Eurasian Journal of Analytical Chemistry

Eurasian J Anal Chem 10(1): 68-90, 2015

Diagnosing and correcting three-year drift in metal-oxide chemosensor arrays

Luis Eduardo Muñoz Guerrero

Facultad de Ingenierías, Universidad Tecnológica de Pereira, Pereira, Colombia, E-mail: lemunozg@utp.edu.com, ORCID: 0000-0002-9414-6187

Received: *11/04/2015****; Accepted:*** *24/05/2015*

Abstract

Metal-oxide chemosensor arrays lose analytical validity over time because their sensing layers age, so that a calibration model built today misclassifies samples measured months later. This work characterises that drift and tests what can be done about it without acquiring new reference standards, using a public archive of 13 910 measurements recorded with a sixteen-element array over 36 months for six volatile analytes between 1.0 and 1000 ppmv, each described by eight descriptors per sensor. A variance decomposition shows that the between-session mean square of a typical descriptor exceeds its within-session residual mean square by a median factor of 123, and that the session-to-session shift is shared by the six analytes (median pairwise correlation 0.52, all 15 pairs positive) and confined to a low-dimensional subspace, two components accounting for 79.7% of its variance. Drift susceptibility is not shared equally: the steady-state resistance change and the slowly smoothed decaying-transient descriptors are the most robust, the baseline-normalised steady-state descriptor and the rising-transient descriptors the most fragile, a 3.0-fold difference. Filtering the descriptors on that ranking does not, however, improve prospective accuracy, so the ratio is a diagnostic rather than a variable-selection criterion. Three signal transforms, eight correction strategies and three classifiers were compared under strictly forward-in-time protocols in which a model never sees a measurement session later than those used to calibrate it. Without correction, six-analyte recognition falls from 99.1% within the calibration session to 41.8% over the following three years. Session-wise autoscaling, which needs neither reference standards nor labels nor a drift model, recovers a large part of that loss; component correction performs comparably but needs a drift model, and orthogonal signal correction is worse than no correction at all.

Keywords:

Chemical sensor array; metal oxide semiconductor sensor; sensor drift; chemometrics; multivariate standardisation; machine olfaction

1. Introduction

Arrays of partially selective chemical sensors, read out with multivariate calibration, have become an established instrumental option whenever a volatile mixture has to be recognised quickly, cheaply and without sample preparation [1,2]. In the most widespread implementation the transducers are metal-oxide semiconductor (MOX) resistors whose conductance changes when reducing or oxidising species adsorb on the heated sensing layer and modify the depletion region at the grain boundaries [3,4]. Because a single MOX element responds to almost every reducing vapour, selectivity comes from the array rather than from the transducer: the pattern of responses across several sensing layers is decoded by a chemometric model [5,6], a strategy that underpins applications from indoor air-quality monitoring to food and process control [7,8].

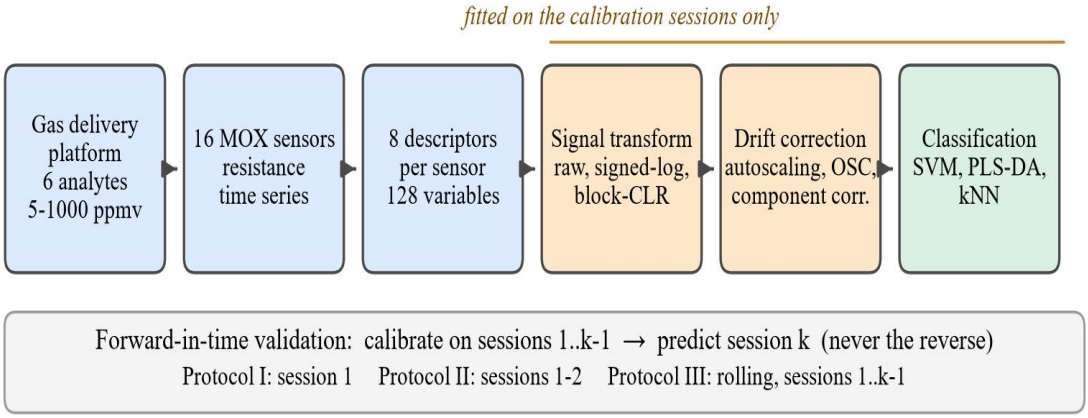
* **Corresponding Author**
E-mail: lemunozg@utp.edu.com
ISSN: 1306-3057
iSER Publication ©2015

The analytical weakness of the approach is that the transducers are not stable. Sintering and coarsening of the polycrystalline grains, slow poisoning by irreversibly adsorbed species and ageing of the heater change the sensing layer continuously, so the response of a given element to a given analyte at a given concentration is not after months of operation what it was on the day of calibration [9,10]. This phenomenon, universally called drift, is not measurement noise but a systematic, slowly varying displacement of the multivariate response, and it is the single largest obstacle to using sensor arrays as routine analytical instruments: a model perfectly validated at the time of construction can be invalid months later, which forces recalibration with fresh standards and defeats much of the cost advantage of the technique.

Two families of counter-measures have been proposed. The first models the drift explicitly and subtracts it. Univariate baseline manipulation and curve fitting of the temporal trend of each sensor were the earliest options [11,12]. Component correction, introduced by Artursson and co-workers, estimates the direction of the drift in the multivariate space from repeated measurements of a reference gas and projects that direction out of the data [13]. The same idea has been implemented with standardisation regressions borrowed from calibration transfer [14], with orthogonal signal correction (OSC), which removes the part of the response matrix orthogonal to the class information [15], and with common principal component analysis, which seeks the drift subspace shared by the analyte-specific covariance matrices [16]. The second family adapts the decision rule instead of the data, by recursive model updating [17] or by ensembles of classifiers trained at different times [18,19]. Related work addresses the mitigation of individual sensor failures [20], arrays in open sampling systems [21] and the design of the calibration experiment [22]; reviews of both families are available [7,23].

Three points remain unsatisfactory. First, most drift studies are carried out on records of weeks to a few months, whereas the practically relevant question is what happens after years. Second, performance is frequently reported with random or cross-validated splits, which let a model borrow information from sessions that would still lie in the future in a real deployment, so the figures of merit are optimistic by construction. Third, and most important for the analyst, the comparison is nearly always made between drift-correction algorithms, while the choice actually under the analyst's control at the method-development stage, which descriptor is extracted from the sensor transient, is rarely examined for drift susceptibility.

The present work addresses the three points on a public archive of 13 910 measurements acquired with a 16-element MOX array over 36 months for six volatile analytes [18,24]. The objectives are (i) to characterise the geometry and magnitude of the drift by decomposing the response variance into between-session and within-session components; (ii) to define a drift-to-noise ratio that ranks each of the 128 descriptors by its susceptibility to ageing; (iii) to compare three signal transforms, eight correction strategies and three classifiers under strictly forward-in-time protocols, so that the contribution of each element of the pipeline can be separated; and (iv) to derive practical recommendations for laboratories that must keep an array in service without frequent recalibration. The literature covered spans 1982 to 2014. The measurements analysed here were recorded at the ChemoSignals Laboratory, BioCircuits Institute, University of California San Diego (USA), over a 36-month period ending in February 2011, and were released through the UCI Machine Learning Repository [24].



descriptors, transformed, corrected for drift and classified. Every element downstream of the descriptor extraction is fitted on the calibration sessions only, and the three validation protocols differ solely in how many past sessions the calibration set is allowed to contain

2. Materials And Methods

The study uses the Gas Sensor Array Drift Dataset at Different Concentrations, distributed through the UCI Machine Learning Repository [24] and described in detail by Vergara and co-workers [18]. The measurements were produced by a fully computer-controlled gas-delivery platform that dosed six pure volatile substances into a test chamber containing a 16-element MOX array: ammonia, acetaldehyde, acetone, ethylene, ethanol and toluene. The array comprised four units each of four commercial Figaro sensor types (TGS2600, TGS2602, TGS2610 and TGS2620); the units are identified below as S1 to S16 in the order in which their descriptors appear in the archive. The complete record contains 13 910 measurements collected over 36 months and is distributed as ten batches, each grouping the measurement months indicated in Table 1. Because every batch corresponds to a distinct period of operation of the same instrument, the batch index is used throughout as the measurement-session index, the two terms being interchangeable.

Table 1 Composition of the measurement archive. Each batch groups the measurement months listed in the second column; the remaining columns give the number of measurements of each analyte

Batch	Months	Ethanol	Ethylene	Ammonia	Acetaldehyde	Acetone	Toluene	Total
1	1, 2	90	98	83	30	70	74	445
2	3, 4, 8, 9, 10	164	334	100	109	532	5	1244
3	11, 12, 13	365	490	216	240	275	0	1586
4	14, 15	64	43	12	30	12	0	161
5	16	28	40	20	46	63	0	197
6	17, 18, 19, 20	514	574	110	29	606	467	2300
7	21	649	662	360	744	630	568	3613
8	22, 23	30	30	40	33	143	18	294
9	24, 30	61	55	100	75	78	101	470
10	36	600	600	600	600	600	600	3600
Total	-	2565	2926	1641	1936	3009	1833	13910

Toluene was not dosed during batches 3, 4 and 5, and only five toluene measurements are

available in batch 2; the imbalance is a property of the archive and is retained here because it reproduces the conditions of a real, opportunistically sampled measurement campaign.

Table 2 Concentration range covered by each analyte across the whole archive

Analyte	*n*	**Minimum (ppmv)**	**Maximum (ppmv)**	**Distinct levels**
Ethanol	2565	2.5	600.0	35
Ethylene	2926	2.5	300.0	38
Ammonia	1641	2.5	1000.0	44
Acetaldehyde	1936	2.5	300.0	37
Acetone	3009	10.0	1000.0	47
Toluene	1833	1.0	230.0	32

The concentration is not supplied to any of the models; it is reported here only to document the analytical span of the archive.

Write the archive as a matrix X whose rows are measurements and whose columns are descriptors, so that xij is descriptor j of measurement i. Each measurement carries an analyte label yi, a nominal concentration ci in ppmv and a session index si running from 1 to 10; every quantity defined below is a function of those four objects and of nothing else.

Each measurement is a 16-channel resistance transient. Following the original feature definition [18,25], eight scalar descriptors are extracted per channel, giving a 128-dimensional response vector per measurement. The first two are steady-state descriptors: ΔR, the difference between the maximum resistance reached in the presence of the analyte and the baseline, and ΔR, its normalised version, obtained as the ratio of the maximum resistance to the baseline. The remaining six summarise the dynamics of the transient through the exponential moving average operator emaα, computed with three smoothing constants (α=0.001, 0.01 and 0.1) on the rising portion of the response (emaαi) and on the decaying portion (emaαd). Each descriptor is identified below by its sensor unit, S1 to S16, and its name.

Three transforms of the response vector were compared. Raw leaves the descriptors unchanged. Signed-log stabilises the variance of descriptors whose values span several decades while preserving the sign of the decaying-transient descriptors,

$$\tilde{x}_{ij} = \operatorname{sgn}\left(x_{ij}\right) \cdot \log_{10}\left(1 + \left|x_{ij}\right|\right) \tag{1}$$

sgn being the sign function. Block-CLR treats the sixteen values of a given descriptor as a composition and applies the centred log-ratio transform of Aitchison [26,27],

$$\operatorname{clr}(x)_k = \ln|x_k| - \frac{1}{16}\sum_{m=1}^{16} \ln|x_m|, \quad k = 1,\ldots,16 \tag{2}$$

the sixteen values being those of one descriptor across the sixteen sensor units. Any common-mode change of the gain of the array, that is, any factor multiplying the whole block of a measurement, cancels exactly under this transform, so Block-CLR is a purely sample-wise, model-free drift compensation requiring no calibration data at all.

Three diagnostics were used. The session effect was obtained by computing, for every analyte g and session s with at least three measurements, the centroid of the autoscaled responses, and subtracting from each analyte the mean of its own centroids over sessions.

Writing xgs for the centroid of analyte g in session s and Sg for the sessions in which that analyte was dosed, the session effect is

$$\mathbf{d}_{gs} = \bar{x}_{gs} - \frac{1}{|S_g|}\sum_{t \in S_a} \bar{x}_{gt} \tag{3}$$

which is free of the analyte main effect by construction and describes how the instrument

itself moved between sessions. The agreement between analytes was quantified as the Pearson correlation between the session-effect vectors of every pair of analytes over the sessions they share, and the dimensionality of the pooled session effect by singular value decomposition [28].

The drift-to-noise ratio (DNR) of a descriptor is the ratio of its between-session mean square to its within-session mean square, both computed within analytes so that the analyte identity

$$\mathrm{MS}_{\mathrm{between}} = \frac{\sum_g \sum_s n_{gs} \left(\bar{x}_{gs} - \bar{x}_g\right)^2}{\mathrm{df}_{\mathrm{between}}} \tag{4}$$

$$\mathrm{MS}_{\mathrm{within}} = \frac{\sum_g \sum_s \sum_{i \in (g,s)} \left(x_i - \bar{x}_{gs}\right)^2}{\mathrm{df}_{\mathrm{within}}} \tag{5}$$

$$\mathrm{DNR} = \frac{\mathrm{MS}_{\mathrm{between}}}{\mathrm{MS}_{\mathrm{within}}} \tag{6}$$

Here xgs is the mean of the descriptor for analyte g in session s, ngs the number of measurements in that cell and xg the analyte mean over the sessions retained; the degrees of freedom are 51 and 13 853. A large DNR means that the descriptor is drift dominated. One caveat must be stated at once: the measurements pooled inside a cell are not replicates, since the concentration varies across them by up to two decades (Table 2), so the denominator is a within-cell residual carrying the response to concentration as well as measurement noise. DNR therefore compares between-session movement with within-session spread of any origin, which is the quantity a practitioner cares about, but it is not a signal-to-noise ratio in the strict sense and is not called one below. Sessions contributing fewer than three measurements of an analyte were excluded from the sums, and because DNR needs at least two calibration sessions it can only be estimated when the calibration set spans more than one.

A third diagnostic places the drift on the same scale as the chemical information. The total sum of squares of the autoscaled archive was split into four sequential terms, the analyte identity, the measurement session within analyte, a linear term in log10 of the concentration fitted inside each analyte-and-session cell, and the within-cell residual,

$$\mathrm{SS}_{\mathrm{total}} = \mathrm{SS}_{\mathrm{analyte}} + \mathrm{SS}_{\mathrm{session}} + \mathrm{SS}_{\mathrm{concentration}} + \mathrm{SS}_{\mathrm{residual}} \tag{7}$$

so that each term can be read as a share of the response variance. The split is sequential, but the terms are hierarchically nested and the concentration term is centred inside each cell, so they are mutually orthogonal here: entering them in any of the six possible orders was verified to reproduce the same four shares to the first decimal.The degrees of freedom are 5 for the analyte term, 51 for the session term nested within analyte, 55 for the concentration slopes and 13 853 for the residual. One caution applies to the session term and is quantified with the results: because the concentration term only removes the slope inside each cell, any difference between sessions in the concentrations that were dosed is charged to the session. The same within-session logic gives the analyte-discriminating power of a descriptor, an F ratio computed inside each session and pooled over sessions,

$$F = \frac{\sum_s \sum_g n_{gs} \left(\bar{x}_{gs} - \bar{x}_s\right)^2 / \mathrm{df}_b}{\sum_s \sum_g \sum_{i \in (g,s)} \left(x_i - \bar{x}_{gs}\right)^2 / \mathrm{df}_w} \tag{8}$$

where xs is the overall mean of session s. Because every comparison happens inside one session, the drift cannot inflate this quantity: it measures how well a descriptor separates the analytes on a single day.

Eight strategies were compared, all applied after the signal transform and before the classifier; Table 3 summarises what each has to be given. Fixed autoscaling centres and scales every descriptor with the mean and standard deviation of the calibration set and applies those constants unchanged to all future sessions; it is the no-correction reference. Session-wise autoscaling instead centres and scales each session with its own mean and standard deviation,

so that for a measurement i in session s the two rules are

$$z_{ij}^{\text{fixed}} = \frac{x_{ij} - \mu_j^{\text{cal}}}{\sigma_j^{\text{cal}}}, \qquad z_{ij}^{\text{session}} = \frac{x_{ij} - \mu_j^{(s)}}{\sigma_j^{(s)}} \tag{9}$$

μcal and σcal being estimated once on the calibration set and μs and σs afresh on each session as it arrives. The second rule uses only the descriptor values of the session being predicted and never its labels, so it needs no reference standards and no drift model; it is the multivariate analogue of re-zeroing an instrument before each working session. Component correction follows Artursson and co-workers [13] in projecting a drift subspace out of the data. In the original method that subspace comes from repeated measurements of a reference gas, whose only source of variation over time is the drift itself. The archive contains no such gas, so the equivalent quantity was built from the calibration sessions: from the centroid of every (analyte, session) group the mean of that analyte over its own sessions was subtracted, leaving a matrix free of the analyte identity and of the concentration distribution that describes only how the instrument moved between sessions. Its leading d principal directions were projected out of every measurement. Collecting them as the columns of a matrix D, the correction is a projection onto the orthogonal complement of the drift subspace,

$$\tilde{\mathbf{Z}} = \mathbf{Z} - \mathbf{Z}\mathbf{D}\mathbf{D}^T \tag{10}$$

in which Z is the autoscaled response matrix. Orthogonal signal correction extracts from the calibration set the d leading score vectors orthogonal to the dummy class matrix, obtains the corresponding weights and loadings by regression on the calibration responses, and subtracts the resulting rank-d component from every measurement, following the direct-orthogonalisation formulation [29,30]. With T those scores, W those weights and P those loadings, the deflation is

$$\tilde{\mathbf{Z}} = \mathbf{Z} - \mathbf{Z}\mathbf{W}\mathbf{P}^T, \quad \mathbf{W} = \mathbf{Z}^+\mathbf{T}, \quad \mathbf{P} = \mathbf{Z}^T\mathbf{T}(\mathbf{T}^T\mathbf{T})^{-1} \tag{11}$$

Deflating with the scores rather than orthogonally projecting out the leading loadings of the Y-orthogonal part is what guarantees that the removed component carries no class information. Component correction and OSC were run with d=2 and d=5, and component correction was additionally combined with session-wise autoscaling. Finally, DNR-based descriptor selection retains only the k descriptors with the lowest drift-to-noise ratio, the ratio being estimated on the calibration sessions alone.

Table 3 The eight drift-correction strategies. The last three columns state what each one has to be given: reference standards or labelled samples, an explicit model of the drift, and whether it looks at the unlabelled measurements of the session being predicted

Strategy	What it estimates	Needs labels	Needs a drift model	Uses the test session
Fixed autoscaling	calibration mean and standard deviation	no	no	no
Session-wise autoscaling	mean and standard deviation of each session	no	no	yes, descriptors only
OSC, d = 2 or 5	scores orthogonal to the class matrix	yes	no	no
Component correction, d = 2 or 5	leading directions of the session effect	yes	yes	no
Component correction + session-wise	both of the above	yes	yes	yes, descriptors only

DNR descriptor selection	drift-to-noise ratio of each descriptor	yes	no	no

Needing labels refers to the calibration set only; no strategy ever sees a label from a session it is asked to predict. “Uses the test session” means that the descriptor values, but never the labels, of the session being predicted enter the transformation.

Three classifiers spanning the usual chemometric range were used. PLS-DA regresses a dummy response matrix Y on the corrected descriptors by partial least squares and assigns each sample to the class with the largest predicted response [31,32,33],

$$\hat{y}_i = \arg\max_g \left(\tilde{\mathbf{Z}}_i \mathbf{B}\right)_g \tag{12}$$

where B is the regression coefficient matrix and the columns of Y are the six analyte indicators. k-nearest neighbours assigns the majority class of the k closest calibration samples in Euclidean distance. Support vector machines with a radial basis function kernel [34,35,36] were used in the one-against-one multiclass formulation. Hyper-parameters were selected exclusively inside the calibration set by stratified cross-validation: the number of latent variables from 2 to 24 for PLS-DA, k from 1 to 9 for nearest neighbours, and the penalty C and the kernel width γ over the grids 2−5 to 210 and 2−10 to 25 for the support vector machine. Ten folds were used, reduced to three with a coarser grid when the calibration set exceeded 3000 measurements, to keep the search tractable.

Ordinary k-fold cross-validation places measurements from the same session in the training and the validation folds at once, so it estimates within-session performance, which the results below show to be a poor guide to a later session. Wherever the classifier itself had to be chosen, and not merely its hyper-parameters, selection was therefore made by leave-one-session-out cross-validation inside the calibration set, the causal analogue of the forward-in-time protocol. Both rules use calibration data only, and both estimates are reported below so that the difference can be seen.

Evaluation is strictly causal: a model never sees a measurement session later than the one it is asked to predict (Fig. 1). Protocol I calibrates on session 1 alone and predicts sessions 2 to 10, reproducing a laboratory that calibrates once and then uses the instrument. Protocol II calibrates on sessions 1 and 2 and predicts sessions 3 to 10; two sessions are the minimum needed to estimate a drift subspace or a drift-to-noise ratio, so all eight strategies are available there. Protocol III predicts session k from a model calibrated on every earlier session, describing a laboratory that re-labels and re-fits whenever reference measurements happen to become available. The figure of merit is the fraction of correctly identified analytes; the macro-averaged F1 score, computed over the analytes actually present in a test session, is reported alongside it because the class balance varies strongly between sessions (Table 1).

Two classifiers evaluated on the same test session were compared with the exact binomial McNemar test on the discordant pairs [37]. Comparisons across the eight test sessions of Protocol II used the Wilcoxon signed-rank test, the recommended non-parametric procedure for comparing two methods over several data sets [38], as did the agreement of the session effect between analytes, applied there to the pairwise correlations. A significance level of 0.05 was adopted throughout; no correction for multiplicity was applied to the per-session McNemar tests, which are reported descriptively.

The complete analysis was implemented in Python with NumPy and SciPy for the numerical work, scikit-learn for the classifiers and the cross-validated hyper-parameter search [39], and matplotlib for the figures [40]. The support vector machines are the scikit-learn interface to LIBSVM [36]. All random number generation was seeded so that the reported figures are exactly reproducible.

3. Results And Discussion

The analysis begins with the structure of the response itself: before attempting any correction it is worth knowing how much of the signal the drift occupies. The variance decomposition of equation (7) gives a first answer (Fig. 2a): the analyte concentration accounts for 35.8% of the

response variance and the analyte identity for 27.2%, while the measurement session accounts for 22.0% and the within-cell residual for 15.0%. That session figure needs one deduction. The archive was not dosed to a fixed design and the sessions differ in the concentrations they contain (Table 1); regressing the (analyte, session) centroids on their own mean log concentration shows that 32.8% of the session term is accounted for by that difference alone. Net of it the measurement session still carries 14.8% of the response variance, comparable to the within-cell residual and about half of what the analyte identity carries. An instrument on which the calendar weighs that heavily is the problem this paper addresses.

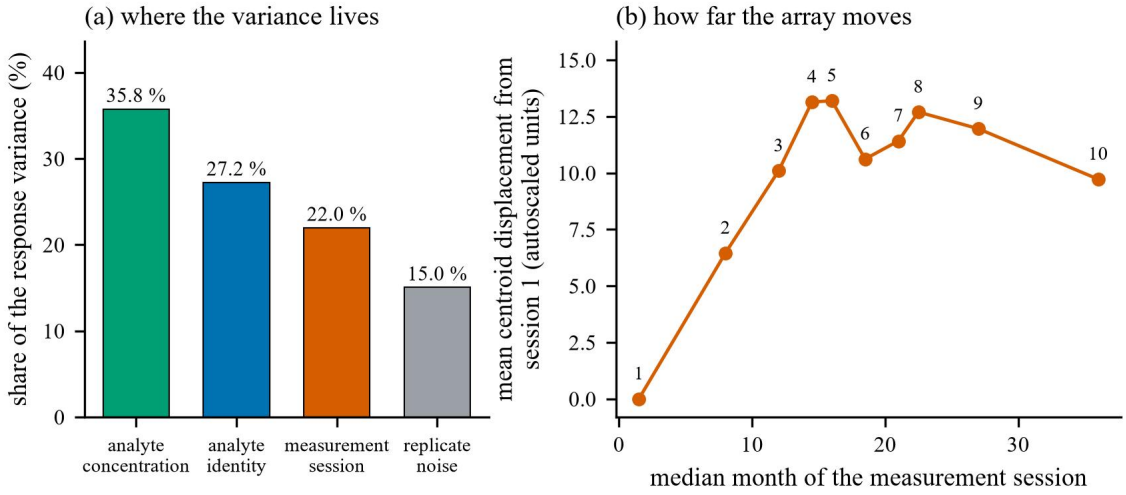


Fig. 2 Structure and magnitude of the drift. (a) Share of the response variance carried by the analyte concentration, the analyte identity, the measurement session and the replicate noise, from the decomposition of equation (7). (b) Mean displacement of the analyte centroids with respect to session 1, in autoscaled units, against the median month of each session; the labels give the session index

The analysis continues with the geometry of that session term. Fig. 3a shows the first two principal components of the autoscaled, signed-log transformed archive. The six analytes occupy overlapping but distinguishable regions, confirming that the array carries the chemical information the recognition task needs. The drift is not visible there because the leading components are dominated by concentration, which varies over more than two decades (Table 2); isolating it requires removing the analyte main effect, as in Fig. 3b, where each point is the centroid of one analyte in one session after the mean of that analyte over sessions has been subtracted, so that only the movement of the instrument remains. The sessions separate cleanly along the two leading components, the early sessions at one end and the late sessions at the other, and the six analytes of a given session move together.

The magnitude of that movement is quantified in Fig. 2b, which plots the mean displacement of the analyte centroids relative to session 1 against the median month of each session. The displacement grows from 6.45 autoscaled units at session 2 to a maximum of 13.19 units at session 5, and is 9.72 units at session 10, month 36. For comparison, the root-mean-square distance of an individual measurement from the centroid of its own analyte and session is 6.43 units. The systematic session-to-session displacement is therefore of the same order as, and at its worst larger than, the entire within-session spread: the instrument moves about as much between sessions as its measurements scatter within one. That is the quantitative statement of the problem the rest of the paper addresses.

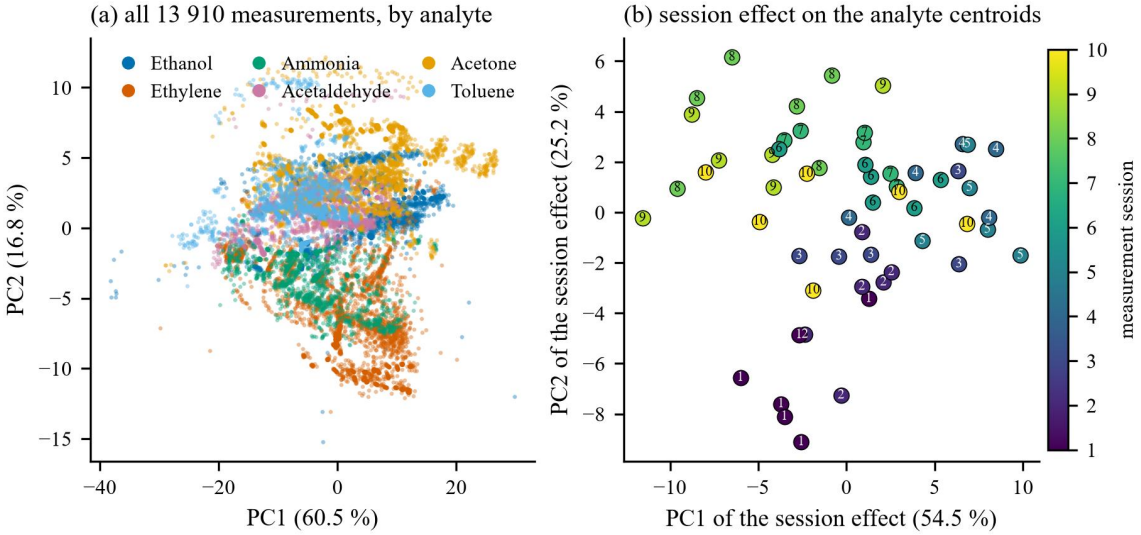


Fig. 3 Principal component analysis of the signed-log transformed, autoscaled archive. (a) All 13 910 measurements, coloured by analyte. (b) Session effect: centroid of each analyte in each session after removal of the analyte main effect, coloured and labelled by session index. Measurements made in the same session move together regardless of which analyte was dosed

If the drift were specific to each analyte, no single correction could remove it for all of them at once and reference-gas schemes would be unusable. The pairwise correlation between the session-effect vectors of the six analytes was therefore computed over the sessions each pair shares. All 15 pairs are positively correlated, with a median coefficient of 0.519 and a range from 0.225 to 0.708 (Wilcoxon signed-rank against zero, $p=3.1\times10-5$). The session-to-session movement is thus a property of the measurement session rather than of the analyte.

What it is not, on this evidence alone, is proof that the sensing layers aged: anything common to a session and external to the chemistry, such as laboratory temperature and humidity, recalibration of the mass-flow controllers or a change of carrier-gas supply, would leave the same signature. The archive records none of those covariates, so the term is named for what identifies it, and ageing of the sensing layer is the most likely but not the only contributor.

The movement is also of low rank. Singular value decomposition of the pooled session effect places 54.5% of its variance in the first component, 25.2% in the second, 79.7% in the first two together and 91.8% in the first five. A correction that removes a handful of directions is therefore geometrically sufficient, which is the premise of component correction [13] and of common principal component analysis [16], and it also explains why a much simpler operation works: if the dominant session effect is a shift and a rescaling the whole array undergoes together, re-centring and re-scaling each session with its own statistics removes most of it without estimating any direction.

That subspace can be read descriptor by descriptor. Fig. 4 shows the magnitude of the loading of every one of the 128 descriptors on the leading drift direction, which alone carries 54.5% of the session effect. The pattern is not uniform: the steady-state change carries the least weight, a mean absolute loading of 0.058 against 0.097 for the most heavily loaded transient descriptor, and the units differ visibly, S9 and S10 barely participating while S13 and S14 dominate. The drift is therefore neither a property of the array as a whole nor a fault of one element, but is concentrated in the transient part of the response and in a subset of the units, which is what makes both a subspace projection and a session-level rescaling plausible remedies.

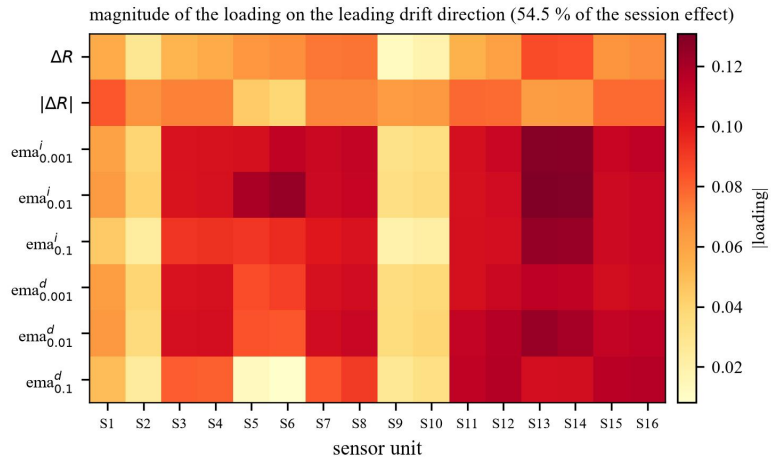


Fig. 4 Magnitude of the loading of each of the 128 descriptors on the leading direction of the session effect, arranged by sensor unit and descriptor type. Darker cells carry more of the drift

The drift-to-noise ratio of the 128 descriptors is summarised in Table 4 and Fig. 5a. Every descriptor is drift dominated: the overall median DNR is 123, two orders of magnitude. The eight types are not equally affected: the steady-state descriptor ΔR is the most robust, with a median DNR of 58, and the fast rising-transient descriptor ema0.1i the most fragile, at 172, a factor of 3.0 higher. The ordering does not follow the steady-state versus transient division the descriptor definitions suggest. It separates instead into a robust group made of ΔR and the two slowly smoothed decaying-transient descriptors (median DNR 58, 84 and 88), and a fragile group made of the normalised steady-state descriptor ΔR and the three rising-transient descriptors, whose medians all lie between 120 and 172. Of the 48 descriptors with the lowest drift-to-noise ratio over the whole archive, 15 of the 16 instances of ΔR, 14 of ema0.01d and 12 of ema0.001d are retained, while not one instance of ΔR or of the two faster rising-transient descriptors survives (Table 4).

Two things follow. First, normalising the steady-state response by the baseline, the usual way of compensating sensor-to-sensor gain differences, gives no protection against ageing: ΔR is among the most drift-prone descriptors in the set. A baseline drifting in proportion to the response would cancel in the ratio and make the normalised descriptor the more robust of the two, so the two must drift differently and the ratio compounds two movements instead of cancelling one. The archive holds no separate record of the baseline, so this cannot be tested here. Second, the rising and decaying halves of the transient behave differently on average but not uniformly: the most heavily smoothed decaying descriptor has a median ratio of 121, indistinguishable from the least drift-prone rising descriptor at 120, so the division is a tendency rather than a rule. Within the rising family susceptibility increases monotonically with the smoothing constant, from 120 at α=0.001 to 138 at α=0.01 and 172 at α=0.1. A mechanistic reading is consistent with the ranking without being established by it: the rise is governed by the surface reaction rate and by transport of the analyte to the grain surface, both sensitive to the state of the sensing layer and to the heater temperature, whereas recovery is governed by desorption, for which incomplete recovery is a classical symptom of ageing. Transient descriptors were introduced to accelerate recognition [25]; a method intended to run for years is nonetheless better served by the steady-state change and the slow decaying-transient descriptors.

It is tempting to ask whether the descriptors that drift most also carry the most analytical information, and the ratios invite an affirmative answer: the drift-to-noise and analyte-discriminating F ratios of equation (8) correlate at +0.54 across the 128 descriptors. That number must be discarded. Both are built on the same denominator, the within-cell mean square of equation (5), so a descriptor with a small within-cell spread is pushed up on both axes and a positive correlation is guaranteed before any data are seen. Comparing the two numerators instead, which share nothing, the association disappears: the rank correlation between the between-session mean square and the between-analyte mean square is −0.15

(Spearman, p=0.08, n=128), and holding the common denominator fixed turns it negative, −0.53 (p=1×10−10). Fig. 5b shows the numerators. There is therefore no evidence in this archive that drift susceptibility and analytical usefulness go together, and any explanation of the selection result below that rests on such a trade-off would be unfounded.

The dispersion across sensor units is smaller than across descriptor types but not negligible: the median DNR per unit ranges from 96 for S15 to 227 for S1, the latter an outlier whose steady-state response falls by more than an order of magnitude over the record. Excluding a single misbehaving unit is a legitimate maintenance action, but it cannot substitute for a correction strategy, since the drift is shared by the whole array.

Table 4 Drift-to-noise ratio of the eight descriptor types, computed over the sixteen sensor units. A larger value means that the descriptor changes more between measurement sessions relative to its replicate variability within a session

Descriptor	Median DNR	Interquartile range	Range	Retained in the 48 lowest
ΔR	58	54 - 61	46 - 172	15
|ΔR|	149	136 - 155	119 - 176	0
$\mathrm{ema}^{\mathrm{i}}_{0.001}$	120	114 - 144	107 - 306	1
$\mathrm{ema}^{\mathrm{i}}_{0.01}$	138	125 - 154	115 - 301	0
$\mathrm{ema}^{\mathrm{i}}_{0.1}$	172	145 - 190	123 - 256	0
$\mathrm{ema}^{\mathrm{d}}_{0.001}$	88	83 - 108	79 - 248	12
$\mathrm{ema}^{\mathrm{d}}_{0.01}$	84	80 - 92	78 - 205	14
$\mathrm{ema}^{\mathrm{d}}_{0.1}$	121	91 - 161	79 - 254	6

The last column gives how many of the sixteen sensor-specific instances of each descriptor type are among the 48 descriptors with the lowest drift-to-noise ratio over the whole archive.

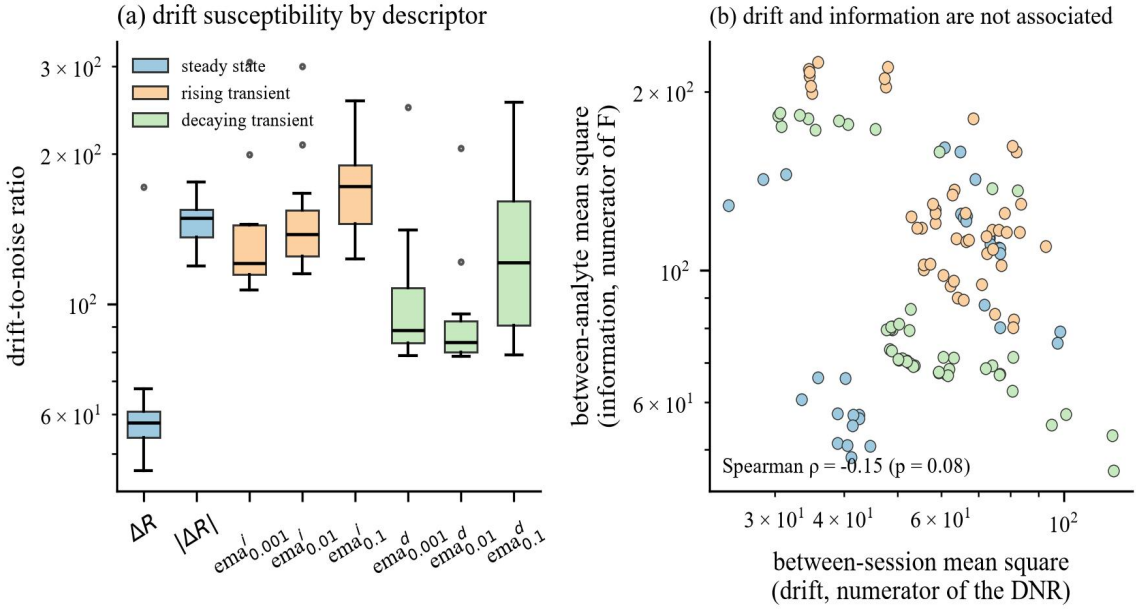


Fig. 5 Drift susceptibility of the descriptors. (a) Distribution over the sixteen sensor units of the drift-to-noise ratio of each descriptor type, on a logarithmic scale, shaded by descriptor family. (b) The two numerators compared directly, the between-analyte mean square of equation (8) against the between-session mean square of equation (4), so that the within-cell mean square they share cannot induce an association; drift susceptibility and analyte-discriminating power are not positively related

Table 5 and Fig. 6 give the outcome of Protocol I, in which the model is calibrated on session 1 alone. Cross-validation inside that session gives a mean accuracy of 99.1%, a figure that would ordinarily be reported as excellent method performance. Applied to the later sessions with fixed autoscaling, the same support vector machine on the raw descriptors reaches only 41.8% over sessions 2 to 10, falling from 49.0% on session 2, two to eight months later, to 38.1% on session 10, three years later. The gap between the cross-validated and the prospective figure measures how misleading a non-causal validation is for this class of instrument, and is why cross-validated accuracies for sensor arrays should always be accompanied by a forward-in-time estimate.

Correction helps even with a single calibration session, but only if it is of the right kind. Session-wise autoscaling raises the mean accuracy to 64.8% with the raw descriptors and 62.7% after the signed-log transform, the best Protocol I configuration overall being the signed-log transform with session-wise autoscaling and the nearest-neighbour classifier, at 65.1%. Orthogonal signal correction, the classical chemometric answer to unwanted structured variance, is actively harmful here: it lowers the mean accuracy in 12 of the 18 combinations of transform, number of components and classifier, and on the raw descriptors with the support vector machine falls to 29.9% for d=2 and 19.5% for d=5 against 41.8% with no correction. The failure is informative, since a single calibration session contains no session-to-session variation, so a drift correction fitted on it is fitted on data containing no drift; this is why Protocol II below requires two sessions.

Table 5 Protocol I. Mean classification accuracy over test sessions 2 to 10 for a model calibrated on session 1 only, for every combination of signal transform, correction strategy and classifier

Correction	Raw / SVM	Raw / PLS	Raw / kNN	S-log / SVM	S-log / PLS	S-log / kNN	CLR / SVM	CLR / PLS	CLR / kNN
None (fixed autoscaling)	0.418	0.273	0.402	0.501	0.501	0.495	0.427	0.444	0.429
Session-wise autoscaling	0.648	0.447	0.638	0.627	0.590	0.651	0.597	0.598	0.590
OSC, d = 2	0.299	0.290	0.343	0.162	0.258	0.327	0.466	0.409	0.458
OSC, d = 5	0.195	0.252	0.321	0.337	0.334	0.354	0.470	0.505	0.451

S-log, signed-log transform; CLR, block-wise centred log-ratio transform; PLS, PLS-DA. Hyper-parameters were tuned by ten-fold cross-validation inside session 1.

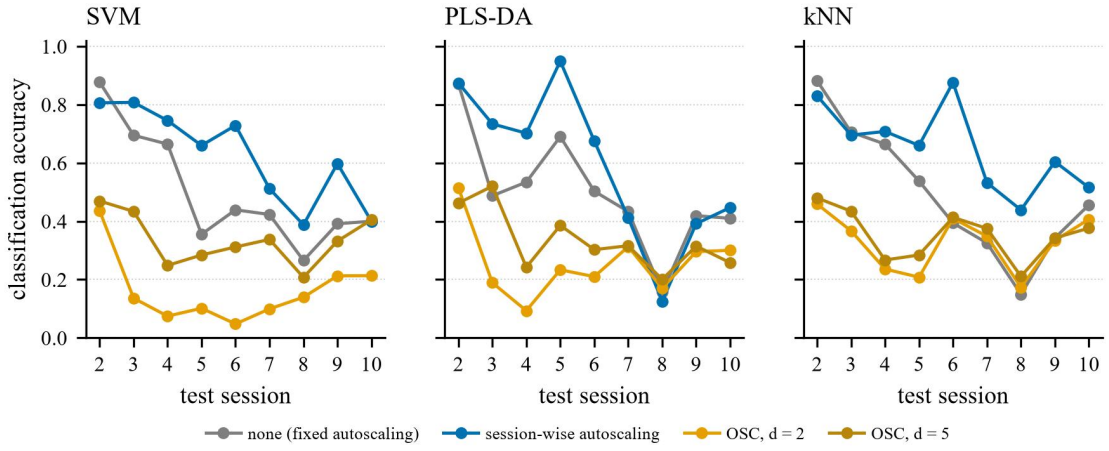


Fig. 6 Protocol I. Accuracy on each test session for a model calibrated on session 1 only, using the signed-log transform, for the three classifiers and four correction strategies

Adding a second calibration session makes the drift identifiable and allows all eight strategies to be compared (Table 6; Fig. 7 traces seven of them session by session). Without correction the mean accuracy over sessions 3 to 10 is 60.4% for the support vector machine on raw descriptors, already higher than under Protocol I because the calibration set now contains some session-to-session variability. Session-wise autoscaling is the most reliable single measure, giving on its own or in combination the best result in 7 of the 9 transform-by-classifier combinations and raising the mean accuracy to 73.0% on the raw descriptors with the support vector machine and to 73.8% with the signed-log transform and PLS-DA. The operation is the sensor-array analogue of the standardisation procedures long used to transfer a multivariate calibration between spectrometers [41,42,43], the difference being that the two instruments to be reconciled here are the same array at two points in its own lifetime.

Component correction is its only serious competitor. With the drift subspace estimated from the two calibration sessions, it improves on the uncorrected reference in all but 1 of the 18 configurations tried, and with five components in all 9 of them; it is the best strategy in the 2 combinations session-wise autoscaling does not win, and with five components and the block-CLR transform it produces the single highest value of the whole grid, 74.0%, marginally above the best session-wise result. The two are not additive: applying component correction first improves on session-wise autoscaling alone in only 8 of the 18 combinations. Both remove the same low-dimensional session effect, one by projecting out its leading directions and the other by re-referencing each session, so once either has been applied little is left for the other. Session-wise autoscaling is the more convenient: it needs no drift model and no subspace dimension, a choice to which component correction is sensitive, five components beating two in 6 of the 9 combinations with no way of knowing the best number before the future sessions are measured.

Orthogonal signal correction behaves quite differently. It is worse than no correction at all in 16 of the 18 configurations tested, and on the raw descriptors with the support vector machine it falls from 60.4% to 50.9% for d=2 and 38.6% for d=5. The component it removes accounts for 55.1% of the calibration variance for d=2 and 60.3% for d=5: what it deletes is the variance orthogonal to the class information, dominated in this archive by the analyte concentration rather than by the drift, so a large part of the analytical signal goes with it while the session displacement, not being orthogonal to the class structure, largely survives. OSC remains an excellent tool where the interfering variance is genuinely orthogonal to the property of interest, as in the near-infrared calibrations for which it was devised [29,30]; long-term chemosensor drift is not such a case.

The two signal transforms behave as expected but modestly. The signed-log transform improves on the raw descriptors for all 3 classifiers when the correction is weak, by preventing the descriptors with the largest numerical range from dominating the Euclidean geometry, and the advantage narrows once a correction is applied. The block-wise centred log-ratio transform, which by construction cancels any factor multiplying the whole array response of a measurement, improves on the raw descriptors in only 1 of the 3 classifiers under fixed autoscaling: the session effect is therefore not predominantly a common-mode multiplicative gain change, and a sample-wise compositional normalisation cannot replace a correction estimated from data. It does combine well with component correction, where the best single value of the grid is found.

Table 6 Protocol II. Mean classification accuracy over test sessions 3 to 10 for a model calibrated on sessions 1 and 2

Correction	**Raw / SVM**	**Raw / PLS**	**Raw / kNN**	**S-log / SVM**	**S-log / PLS**	**S-log / kNN**	**CLR / SVM**	**CLR / PLS**	**CLR / kNN**
None (fixed autoscaling)	0.604	0.627	0.540	0.636	0.657	0.574	0.599	0.645	0.530
Session-wise	0.730	0.716	0.632	0.684	0.738	0.638	0.649	0.723	0.649

autoscaling									
OSC, d = 2	0.509	0.534	0.553	0.538	0.634	0.530	0.406	0.583	0.450
OSC, d = 5	0.386	0.530	0.529	0.608	0.627	0.600	0.464	0.612	0.450
Component correction, d = 2	0.650	0.608	0.576	0.650	0.698	0.587	0.685	0.699	0.587
Component correction, d = 5	0.656	0.629	0.570	0.726	0.660	0.642	0.740	0.696	0.663
Comp. corr. d = 2 + session-wise	0.642	0.689	0.575	0.684	0.687	0.645	0.708	0.708	0.665
Comp. corr. d = 5 + session-wise	0.688	0.700	0.622	0.716	0.711	0.706	0.732	0.695	0.690

Abbreviations as in Table 5. Session-wise autoscaling, alone or combined with component correction, gives the best value in seven of the nine transform-by-classifier combinations; component correction with five components gives the best value in the remaining two.

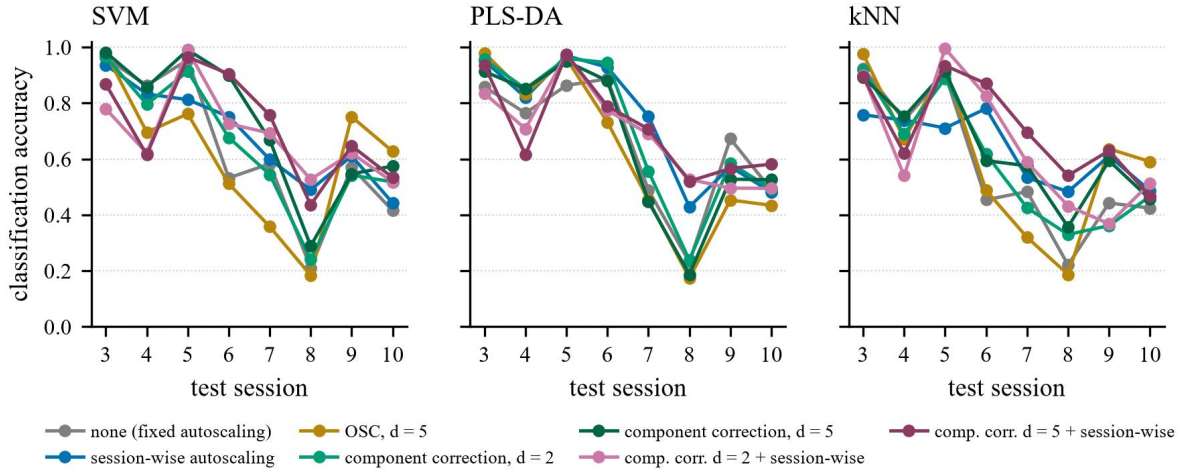


Fig. 7 Protocol II. Accuracy on each test session for a model calibrated on sessions 1 and 2, using the signed-log transform, for the three classifiers and seven correction strategies

Table 7 and Fig. 8 test whether acting on that three-fold spread in drift susceptibility pays off, applying the descriptor filter defined above with each of the three most effective corrections at six values of k. The answer is largely negative, and it is worth stating plainly. When k is fixed in advance, filtering does not improve on keeping all 128 descriptors. Averaged over the six transform-by-classifier combinations, retaining 96 descriptors changes the accuracy by +0.8 percentage points without correction, by −1.6 with session-wise autoscaling and by +1.3 when component correction precedes it; at 64 descriptors the figures are −0.6, −4.5 and +1.2, and every smaller k is worse still.

Only the combination with component correction is ever positive, and then by little more than a percentage point. The highest single value anywhere in the Protocol II experiments, 75.6%, is reached with 48 descriptors, the block-CLR transform, component correction followed by session-wise autoscaling and the PLS-DA model, slightly above the best cell of the main grid. It should not be taken at face value: k was chosen there by looking at the test sessions, so the figure is optimistically biased in exactly the way the head-to-head comparison below warns against, and the same configuration with all 128 descriptors gives 70.8%.

The reason is instructive, and it is not the one that suggests itself. Drift susceptibility and analyte-discriminating power are not positively associated, as shown above, so the filter is not trading information for stability. What it does instead is remove whole descriptor families at

once: the ranking is dominated by descriptor type, so cutting at any k removes all three rising-transient descriptors of every sensor before it removes any steady-state descriptor, and a classifier uses descriptor combinations rather than descriptors one at a time, a univariate ranking being blind to the complementarity between a fragile descriptor and a robust one on the same sensor. The drift-to-noise ratio is therefore best used as a diagnostic: it tells the analyst which parts of the sensor response are ageing, and is exactly the quantity to monitor when deciding that an array needs servicing or that a descriptor has become unusable. It is not a variable-selection criterion, and this paper reports it as such.

Table 7 Protocol II. Mean accuracy over test sessions 3 to 10 when only the k descriptors with the lowest drift-to-noise ratio are retained, followed by session-wise autoscaling

Retained descriptors	S-log / SVM	S-log / PLS	S-log / kNN	CLR / SVM	CLR / PLS	CLR / kNN
16	0.611	0.565	0.598	0.543	0.556	0.499
32	0.614	0.650	0.601	0.579	0.675	0.530
48	0.632	0.684	0.564	0.590	0.714	0.582
64	0.600	0.625	0.567	0.663	0.734	0.623
96	0.641	0.682	0.618	0.650	0.732	0.659
128	0.684	0.738	0.638	0.649	0.723	0.649

The drift-to-noise ratio is estimated on calibration sessions 1 and 2 only. k=128 corresponds to keeping all descriptors. Fig. 8 gives the same comparison for the other two corrections.

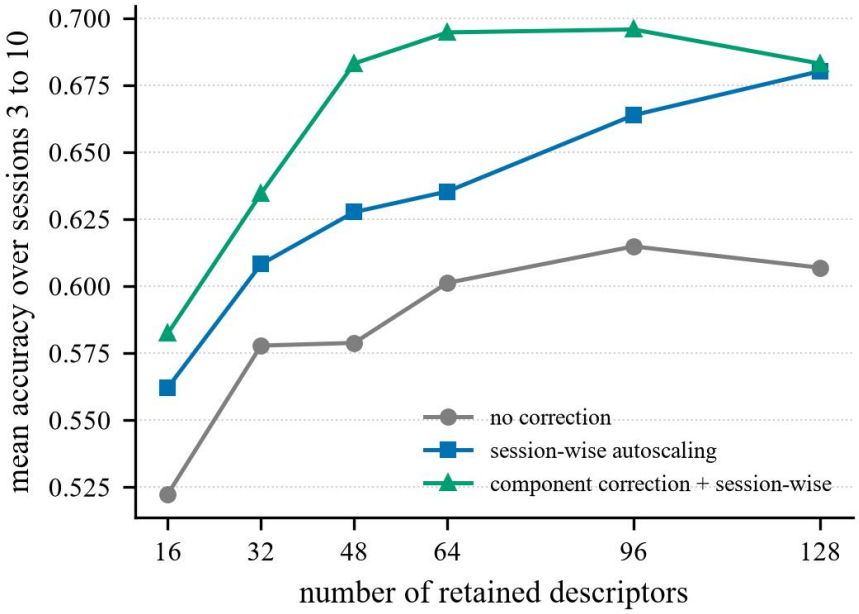


Fig. 8 Protocol II. Mean accuracy over test sessions 3 to 10 against the number of descriptors retained by the drift-to-noise criterion, averaged over the six combinations of signal transform and classifier, for three correction strategies

Protocol III describes the realistic middle ground in which the laboratory occasionally acquires new labelled measurements and refits (Table 8, Fig. 9).
Simply accumulating labelled history is very effective: averaged over sessions 2 to 10 the uncorrected model on raw descriptors reaches 79.3%, against 60.4% under Protocol II, and the mean over the last three sessions, 79.2%, exceeds that over the first three, 74.1%, even though the last sessions lie furthest in time from the start of the record.
The second result of this protocol is a negative one, and it sharpens the message of the earlier sections. Session-wise autoscaling helps enormously when the calibration set is a single

session, which is what session 2 tests, adding +23.8 percentage points there. From session 3 onwards, when the calibration set spans two or more sessions, the mean effect turns negative, −4.1 percentage points, improving only 4 of those 8 sessions and giving 78.3% over the nine sessions together against 79.3% without correction; the largest single effect is a loss of 37.8 percentage points on session 8. A calibration set covering a range of instrument states has taught the classifier to tolerate the session effect, so re-referencing each test session no longer removes a bias and only adds the variance of statistics estimated from a small batch. Session 8 holds 294 measurements, almost half of them one analyte, so its location and spread describe what was dosed as much as what the array was doing; across the nine sessions, however, the relation between compositional imbalance and the benefit of the correction is not statistically significant, so this describes the sessions where the correction fails rather than a general rule.

Component correction behaves differently and becomes the strongest strategy under this protocol, reaching 89.0% with the signed-log transform, component correction, d=5 and the support vector machine. Unlike session-wise autoscaling it re-estimates nothing on the test session, applying a fixed projection learned from the calibration history, so it adds no variance and continues to remove whatever part of the session effect the classifier has not absorbed. Taken with Protocols I and II, this gives a practical rule: corrections matter most when calibration data are scarce, precisely the situation in which a laboratory cannot afford to recalibrate. Accumulating labelled history is an effective substitute, but it buys accuracy with exactly the reference standards and analyst time that the correction was meant to save.

Table 8 Protocol III. Mean classification accuracy over test sessions 2 to 10 for a model calibrated on every earlier session

Correction	**Raw / SVM**	**Raw / PLS**	**Raw / kNN**	**S-log / SVM**	**S-log / PLS**	**S-log / kNN**	**CLR / SVM**	**CLR / PLS**	**CLR / kNN**
None (fixed autoscaling)	0.793	0.767	0.796	0.869	0.854	0.797	0.798	0.847	0.715
Session-wise autoscaling	0.783	0.827	0.734	0.798	0.871	0.744	0.755	0.823	0.755
OSC, d = 2	0.838	0.780	0.851	0.784	0.794	0.759	0.763	0.839	0.709
OSC, d = 5	0.799	0.798	0.848	0.812	0.811	0.790	0.774	0.866	0.740
Component correction, d = 2	0.809	0.821	0.788	0.855	0.870	0.820	0.847	0.867	0.748
Component correction, d = 5	0.809	0.827	0.811	0.890	0.860	0.793	0.864	0.854	0.754
Comp. corr. d = 2 + session-wise	0.768	0.808	0.743	0.776	0.835	0.756	0.834	0.841	0.790
Comp. corr. d = 5 + session-wise	0.774	0.798	0.762	0.815	0.858	0.747	0.824	0.863	0.753

Abbreviations as in the Protocol I table. Session 2 is predicted from session 1 alone, so the two component-correction rows and their combinations start at session 3.

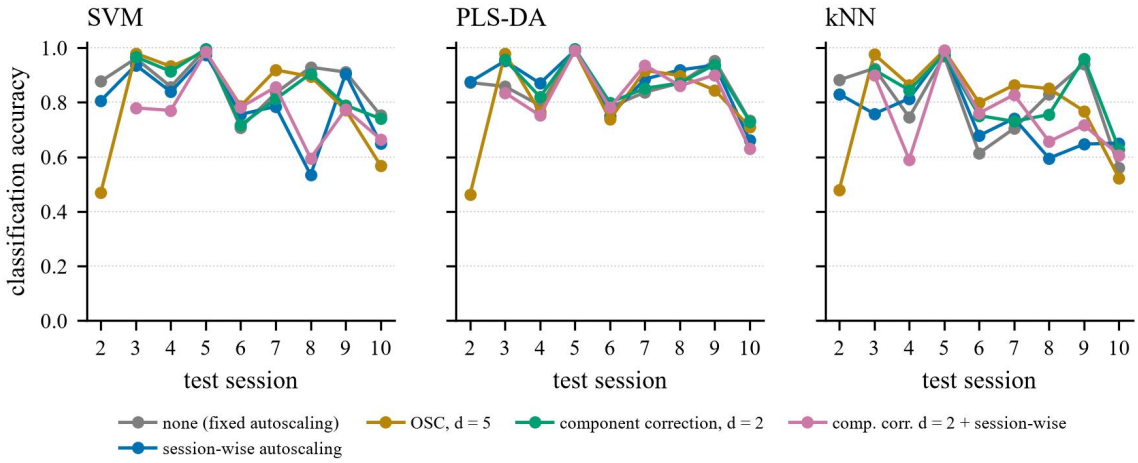


Fig. 9 Protocol III. Accuracy on session k for a model calibrated on all sessions up to k−1, using the signed-log transform, for the three classifiers and the main correction strategies

The comparisons above scan a grid, so the best cell is chosen with knowledge of the test sessions and its performance is optimistically biased. The comparison that follows therefore uses two pipelines a laboratory could have specified at calibration time: the uncorrected reference, raw descriptors with fixed autoscaling, and the pipeline motivated by the variance decomposition and the geometry of the session effect above, the signed-log transform followed by session-wise autoscaling. For both, the classifier and its hyper-parameters were selected by leave-one-session-out cross-validation inside sessions 1 and 2, which picked the support vector machine for the reference and the PLS-DA model for the recommended pipeline. The test sessions played no part in any choice.

The choice of selection rule matters, and it matters in the direction this paper predicts (Table 9). For the recommended pipeline the ordinary k-fold estimate inside the calibration set is 99.3% for the support vector machine, 97.9% for PLS-DA and 98.9% for nearest neighbours, and it would select the support vector machine; the leave-one-session-out estimate of the same three models is 73.3%, 75.0% and 73.2%, and it selects PLS-DA, which is in fact the better of the two on the future sessions. The two estimates differ by roughly 25 percentage points, a second measurement of the same phenomenon quantified by Protocol I, and only the second is an honest forecast. This is also why the accuracies in Table 10 differ from the corresponding entries of Table 6, where the classifier is fixed in advance and only C, γ or the number of latent variables is tuned, by k-fold.

Table 9 What the two cross-validation rules estimate inside the calibration sessions, in per cent, for each candidate classifier and each of the two pipelines compared below

Classifier	Reference, k-fold	Reference, session-out	Corrected, k-fold	Corrected, session-out
Support vector machine	99.3	75.4	99.3	73.3
Pls-da model	96.4	72.1	97.9	75.0
Nearest-neighbour classifier	99.2	70.8	98.9	73.2

The selected model is the row with the highest session-out estimate in each pipeline. For the corrected pipeline the k-fold rule would have selected a different classifier, and the worse of the two on the future sessions.

On this fully causal comparison (Table 10) the recommended pipeline is better on 8 of the 8 test sessions, the mean accuracy over sessions rising from 48.6% to 70.7% and the pooled accuracy over all 12 221 test measurements from 45.1% to 69.6%; the difference over sessions is significant by the Wilcoxon signed-rank test (p=0.0078, the smallest value

attainable with eight paired observations), and the exact McNemar test is significant at the 0.05 level on 8 of the 8 sessions individually.
For reference, the best cell of the whole Protocol II grid reaches 74.0% (block-CLR transform, component correction, d=5, SVM); the causally specified pipeline is 3.3 percentage points below that hindsight optimum, which bounds what the grid search could have added.

Table 10 Protocol II. Session-by-session comparison of the uncorrected reference pipeline with the recommended corrected pipeline

Session	*n*	**Accuracy, reference**	**Accuracy, corrected**	**Macro-F1, reference**	**Macro-F1, corrected**	**Discordant pairs**	*p*
3	1586	0.810	0.931	0.793	0.915	91 / 284	< 0.001
4	161	0.671	0.783	0.612	0.800	3 / 21	< 0.001
5	197	0.614	0.959	0.626	0.963	3 / 71	< 0.001
6	2300	0.440	0.817	0.406	0.736	17 / 885	< 0.001
7	3613	0.385	0.710	0.366	0.709	95 / 1270	< 0.001
8	294	0.221	0.483	0.146	0.485	11 / 88	< 0.001
9	470	0.372	0.440	0.261	0.404	64 / 96	0.014
10	3600	0.377	0.534	0.335	0.521	312 / 880	< 0.001
Mean	12221	0.486	0.707	0.443	0.692	-	0.0078 (W)

Discordant pairs are given as the number of measurements correct only for the reference pipeline and correct only for the corrected pipeline. p values are from the exact binomial McNemar test; the value marked (W) in the last row is the Wilcoxon signed-rank test over the eight sessions. Selected models: reference, support vector machine with C=29 and a kernel width of 2−9; corrected, PLS-DA with 16 latent variables.

Table 11 Protocol II. Analyte-wise figures of merit pooled over test sessions 3 to 10, for the uncorrected reference pipeline and the recommended corrected pipeline

Analyte	*n*	**Recall, ref.**	**Recall, corr.**	**Precision, ref.**	**Precision, corr.**	**F1, ref.**	**F1, corr.**
Ethanol	2311	0.692	0.782	0.250	0.824	0.367	0.803
Ethylene	2494	0.737	0.929	0.856	0.907	0.792	0.918
Ammonia	1458	0.840	0.858	0.850	0.982	0.845	0.916
Acetaldehyde	1797	0.269	0.385	0.259	0.407	0.264	0.395
Acetone	2407	0.151	0.867	0.997	0.520	0.262	0.650
Toluene	1754	0.000	0.203	n.d.	0.728	n.d.	0.317

ref., uncorrected reference pipeline; corr., recommended corrected pipeline. Toluene is absent from test sessions 3, 4 and 5. n.d., not defined: the reference pipeline never predicts toluene, so its precision and F1 for that analyte do not exist.
The analyte-wise breakdown (Table 11, Fig. 10) shows that the gain is not uniform. Ethylene remains the best recognised analyte, with a recall of 0.929 after correction, while toluene is the hardest, at 0.203. The uncorrected reference does not merely recognise toluene poorly: it never predicts that analyte at all, so its recall is exactly zero and its precision undefined, and the correction recovers at least part of the class.
The explanation is coverage rather than selectivity: the calibration sessions contain only 79 toluene measurements out of 1 689, and a class represented that thinly is unlikely to be predicted whatever correction is applied. The residual confusions are between the analytes

whose response patterns are most similar across the array, reduced but not eliminated by the correction, as expected: session-wise standardisation removes the instrumental displacement but adds no information the calibration set never contained.

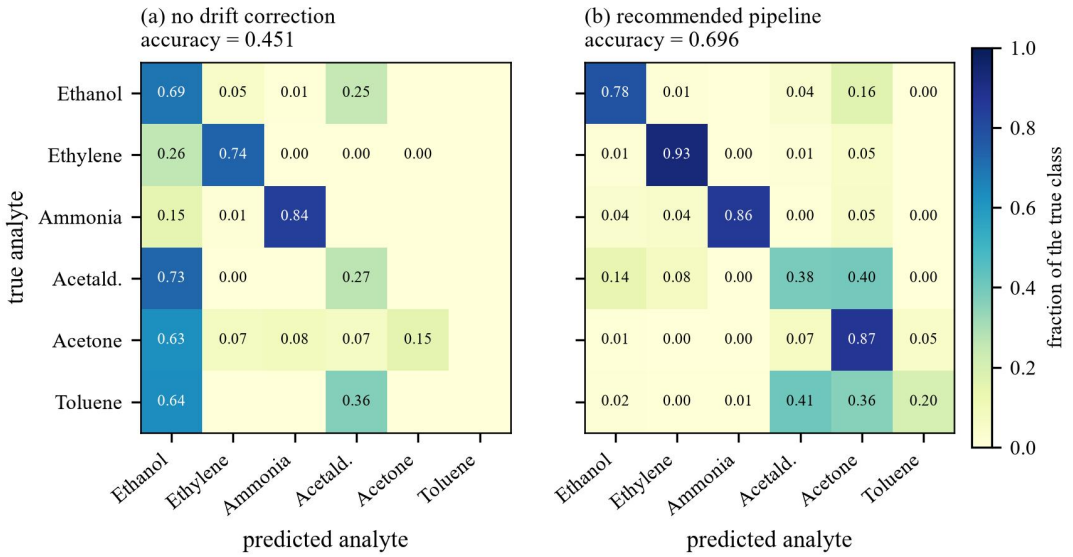


Fig. 10 Confusion matrices pooled over test sessions 3 to 10 under Protocol II, normalised by the number of measurements of each true analyte, for (a) the uncorrected reference pipeline and (b) the recommended corrected pipeline

Three practical conclusions follow. First, a figure of merit for a chemical sensor array should be obtained with a forward-in-time protocol, and where only a few calibration sessions exist, holding out one whole session at a time is a far better proxy than random k-fold; the same rule should govern model selection, since only the session-out rule chose the model that performed better later. Second, the cheapest effective counter-measure is not an algorithm that models the drift but a standardisation that re-references each session to its own location and spread, which needs no reference standards, no labels and no assumption about the shape of the drift, and asks only that a session contain a reasonable variety of analytes, as routine operation normally provides.

Component correction reaches comparable accuracy and remains sound where a reference gas exists, but it needs a drift model and a subspace dimension the calibration data alone cannot fix; orthogonal signal correction should not be used for this purpose at all.

Third, the drift-to-noise ratio belongs in method development as a diagnostic and not as a selection rule: it tells the analyst which part of the response is degrading and how fast, so that an array can be serviced or a descriptor retired on evidence rather than on a fixed schedule, but discarding the drift-prone descriptors does not improve prospective accuracy.

The limitations should be stated plainly. The archive comes from a single array of one design, so the numerical values of the drift-to-noise ratios are specific to these Figaro sensor types and to this operating protocol, even though the ordering of the descriptor families has a physical rationale that should generalise. Session-wise standardisation requires a batch of measurements rather than a single sample, so it applies to campaign-style operation rather than to a truly one-sample-at-a-time instrument, and it implicitly assumes that the composition of the batch is not grossly unbalanced; a session containing a single analyte would be mis-standardised.

Finally, the class imbalance of the archive, and in particular the absence of toluene from three test sessions, makes the accuracies of individual sessions less comparable than their means, which is why the macro-averaged F1 score is reported alongside.

4. Conclusion

Three findings are firm. First, on this array the measurement session is not a nuisance term of

second order: after deducting the part attributable to differences in what was dosed it still carries 14.8% of the response variance, comparable to the within-cell residual, and it moves the analyte centroids by as much as those centroids scatter within a single session. Second, that movement is shared by all six analytes and lies in a subspace of two or three dimensions, which is why it can be removed without a model of its physical cause. Third, and least expected, the correction that works is the one that assumes least: re-referencing each session to its own mean and standard deviation, an operation needing neither reference standards nor labels nor a drift model, recovered accuracy from 48.6% to 70.7% across eight test sessions spanning three years and was better on every one of them, while orthogonal signal correction, the textbook remedy for unwanted structured variance, was worse than doing nothing at all.

The added value lies less in any one of those numbers than in what the three protocols together say about how such a method should be reported. An accuracy obtained by cross-validation inside one measurement session overstates what the same instrument will do a year later by tens of percentage points, and the same optimism infects model selection: the rule that chose the better classifier here was the one that held out a whole calibration session, not the one that shuffled measurements at random. A corollary is negative: ranking descriptors by drift susceptibility diagnoses which part of the response is degrading but is not a variable-selection criterion. For a laboratory that must keep an array in service for years the practical route is therefore neither a better classifier nor a more elaborate drift model: it is forward-in-time validation, model selection by leave-one-session-out cross-validation, and a standardisation applied afresh to every batch of measurements as it is acquired.

Funding

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Authors' Contributions

Luis Eduardo Muñoz Guerrero designed the study, implemented the chemometric analysis, interpreted the results and wrote the manuscript.

Conflict Of Interest

No conflict of interest to declare.

Supporting Information

The measurement archive analysed here is publicly available from the UCI Machine Learning Repository as the Gas Sensor Array Drift Dataset at Different Concentrations [24].

References

1. Persaud K, Dodd G (1982) Analysis of discrimination mechanisms in the mammalian olfactory system using a model nose. Nature, 299: 352-355. https://doi.org/10.1038/299352a0
2. Gardner JW, Bartlett PN (1994) A brief history of electronic noses. Sensors and Actuators B: Chemical, 18: 210-211. https://doi.org/10.1016/0925-4005(94)87085-3
3. Barsan N, Weimar U (2001) Conduction model of metal oxide gas sensors. Journal of Electroceramics, 7: 143-167. https://doi.org/10.1023/a:1014405811371
4. Wang C, Yin L, Zhang L, Xiang D, Gao R (2010) Metal oxide gas sensors: sensitivity and influencing factors. Sensors, 10: 2088-2106. https://doi.org/10.3390/s100302088
5. Hierlemann A, Gutierrez-Osuna R (2008) Higher-order chemical sensing. Chemical Reviews, 108: 563-613. https://doi.org/10.1021/cr068116m
6. Gutierrez-Osuna R (2002) Pattern analysis for machine olfaction: a review. IEEE Sensors Journal, 2: 189-202. https://doi.org/10.1109/jsen.2002.800688
7. Marco S, Gutierrez-Galvez A (2012) Signal and data processing for machine olfaction

and chemical sensing: a review. IEEE Sensors Journal, 12: 3189-3214. https://doi.org/10.1109/jsen.2012.2192920

8. Zhang L, Tian F, Nie H, Dang L, Li G, Ye Q, Kadri C (2012) Classification of multiple indoor air contaminants by an electronic nose and a hybrid support vector machine. Sensors and Actuators B: Chemical, 174: 114-125. https://doi.org/10.1016/j.snb.2012.07.021
9. Korotcenkov G, Cho BK (2011) Instability of metal oxide-based conductometric gas sensors and approaches to stability improvement (short survey). Sensors and Actuators B: Chemical, 156: 527-538. https://doi.org/10.1016/j.snb.2011.02.024
10. Romain AC, Nicolas J (2010) Long term stability of metal oxide-based gas sensors for e-nose environmental applications: an overview. Sensors and Actuators B: Chemical, 146: 502-506. https://doi.org/10.1016/j.snb.2009.12.027
11. Holmberg M, Davide FA, Di Natale C, D'Amico A, Winquist F, Lundstrom I (1997) Drift counteraction in odour recognition applications: lifelong calibration method. Sensors and Actuators B: Chemical, 42: 185-194. https://doi.org/10.1016/s0925-4005(97)80335-8
12. Haugen JE, Tomic O, Kvaal K (2000) A calibration method for handling the temporal drift of solid state gas-sensors. Analytica Chimica Acta, 407: 23-39. https://doi.org/10.1016/s0003-2670(99)00784-9
13. Artursson T, Eklov T, Lundstrom I, Martensson P, Sjostrom M, Holmberg M (2000) Drift correction for gas sensors using multivariate methods. Journal of Chemometrics, 14: 711-723. https://doi.org/10.1002/1099-128x(200009/12)14:5/6<711::aid-cem607>3.0.co;2-4
14. Tomic O, Ulmer H, Haugen JE (2002) Standardization methods for handling instrument related signal shift in gas-sensor array measurement data. Analytica Chimica Acta, 472: 99-111. https://doi.org/10.1016/s0003-2670(02)00936-4
15. Padilla M, Perera A, Montoliu I, Chaudry A, Persaud K, Marco S (2010) Drift compensation of gas sensor array data by orthogonal signal correction. Chemometrics and Intelligent Laboratory Systems, 100: 28-35. https://doi.org/10.1016/j.chemolab.2009.10.002
16. Ziyatdinov A, Marco S, Chaudry A, Persaud K, Caminal P, Perera A (2010) Drift compensation of gas sensor array data by common principal component analysis. Sensors and Actuators B: Chemical, 146: 460-465. https://doi.org/10.1016/j.snb.2009.11.034
17. Perera A, Papamichail N, Barsan N, Weimar U, Marco S (2006) On-line novelty detection by recursive dynamic principal component analysis and gas sensor arrays under drift conditions. IEEE Sensors Journal, 6: 770-783. https://doi.org/10.1109/jsen.2006.874015
18. Vergara A, Vembu S, Ayhan T, Ryan MA, Homer ML, Huerta R (2012) Chemical gas sensor drift compensation using classifier ensembles. Sensors and Actuators B: Chemical, 166-167: 320-329. https://doi.org/10.1016/j.snb.2012.01.074
19. Vergara A, Huerta R, Ayhan T, Ryan MA, Vembu S, Homer ML (2011) Gas sensor drift mitigation using classifier ensembles. Proceedings of the Fifth International Workshop on Knowledge Discovery from Sensor Data, ACM, New York, 16-24. https://doi.org/10.1145/2003653.2003655
20. Fonollosa J, Vergara A, Huerta R (2013) Algorithmic mitigation of sensor failure: is sensor replacement really necessary? Sensors and Actuators B: Chemical, 183: 211-221. https://doi.org/10.1016/j.snb.2013.03.034
21. Vergara A, Fonollosa J, Mahiques J, Trincavelli M, Rulkov N, Huerta R (2013) On the

performance of gas sensor arrays in open sampling systems using inhibitory support vector machines. Sensors and Actuators B: Chemical, 185: 462-477. https://doi.org/10.1016/j.snb.2013.05.027

22. Rodriguez-Lujan I, Fonollosa J, Vergara A, Homer M, Huerta R (2014) On the calibration of sensor arrays for pattern recognition using the minimal number of experiments. Chemometrics and Intelligent Laboratory Systems, 130: 123-134. https://doi.org/10.1016/j.chemolab.2013.10.012
23. Di Carlo S, Falasconi M (2012) Drift correction methods for gas chemical sensors in artificial olfaction systems: techniques and challenges. In: Advances in Chemical Sensors, InTech, Rijeka, 305-326. https://doi.org/10.5772/33411
24. Bache K, Lichman M (2013) UCI Machine Learning Repository. University of California, School of Information and Computer Sciences, Irvine, CA. http://archive.ics.uci.edu/ml
25. Muezzinoglu MK, Vergara A, Huerta R, Rulkov N, Rabinovich MI, Selverston A, Abarbanel HDI (2009) Acceleration of chemo-sensory information processing using transient features. Sensors and Actuators B: Chemical, 137: 507-512. https://doi.org/10.1016/j.snb.2008.10.065
26. Aitchison J (1982) The statistical analysis of compositional data. Journal of the Royal Statistical Society, Series B, 44: 139-160. https://doi.org/10.1111/j.2517-6161.1982.tb01195.x
27. Egozcue JJ, Pawlowsky-Glahn V, Mateu-Figueras G, Barcelo-Vidal C (2003) Isometric logratio transformations for compositional data analysis. Mathematical Geology, 35: 279-300. https://doi.org/10.1023/a:1023818214614
28. Wold S, Esbensen K, Geladi P (1987) Principal component analysis. Chemometrics and Intelligent Laboratory Systems, 2: 37-52. https://doi.org/10.1016/0169-7439(87)80084-9
29. Wold S, Antti H, Lindgren F, Ohman J (1998) Orthogonal signal correction of near-infrared spectra. Chemometrics and Intelligent Laboratory Systems, 44: 175-185. https://doi.org/10.1016/s0169-7439(98)00109-9
30. Fearn T (2000) On orthogonal signal correction. Chemometrics and Intelligent Laboratory Systems, 50: 47-52. https://doi.org/10.1016/s0169-7439(99)00045-3
31. Geladi P, Kowalski BR (1986) Partial least-squares regression: a tutorial. Analytica Chimica Acta, 185: 1-17. https://doi.org/10.1016/0003-2670(86)80028-9
32. Barker M, Rayens W (2003) Partial least squares for discrimination. Journal of Chemometrics, 17: 166-173. https://doi.org/10.1002/cem.785
33. Ballabio D, Consonni V (2013) Classification tools in chemistry. Part 1: linear models. PLS-DA. Analytical Methods, 5: 3790-3798. https://doi.org/10.1039/c3ay40582f
34. Cortes C, Vapnik V (1995) Support-vector networks. Machine Learning, 20: 273-297. https://doi.org/10.1023/a:1022627411411
35. Brereton RG, Lloyd GR (2010) Support vector machines for classification and regression. Analyst, 135: 230-267. https://doi.org/10.1039/b918972f
36. Chang CC, Lin CJ (2011) LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2: 27. https://doi.org/10.1145/1961189.1961199
37. Dietterich TG (1998) Approximate statistical tests for comparing supervised classification learning algorithms. Neural Computation, 10: 1895-1923. https://doi.org/10.1162/089976698300017197
38. Demsar J (2006) Statistical comparisons of classifiers over multiple data sets. Journal of

Machine Learning Research, 7: 1-30. http://jmlr.org/papers/v7/demsar06a.html

39. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: machine learning in Python. Journal of Machine Learning Research, 12: 2825-2830. http://jmlr.org/papers/v12/pedregosa11a.html
40. Hunter JD (2007) Matplotlib: a 2D graphics environment. Computing in Science and Engineering, 9: 90-95. https://doi.org/10.1109/mcse.2007.55
41. Wang Y, Veltkamp DJ, Kowalski BR (1991) Multivariate instrument standardization. Analytical Chemistry, 63: 2750-2756. https://doi.org/10.1021/ac00023a016
42. Bouveresse E, Massart DL (1996) Standardisation of near-infrared spectrometric instruments: a review. Vibrational Spectroscopy, 11: 3-15. https://doi.org/10.1016/0924-2031(95)00055-0
43. Feudale RN, Woody NA, Tan H, Myles AJ, Brown SD, Ferre J (2002) Transfer of multivariate calibration models: a review. Chemometrics and Intelligent Laboratory Systems, 64: 181-192. https://doi.org/10.1016/s0169-7439(02)00085-0