Extended disjoint principal-components regression analysis of SAW vapor sensor-array responses

Edward T. Zellers, Tin-Su Pan*, Samuel J. Patrash, Mingwei Han and Stuart A. Batterman

University of Michigan, School of Public Health, Department of Environmental and Industrial Health, Ann Arbor, MI 48109-2029 (USA)

(Received February 5, 1992; in revised form September 1, 1992; accepted September 19, 1992)

Abstract

The application of a disjoint principal-components regression method to the analysis of sensor-array response patterns is demonstrated using published data from ten polymer-coated surface-acoustic-wave (SAW) sensors exposed to each of nine vapors. Use of the method for the identification and quantitation of the components of vapor mixtures is shown by simulating the 36 possible binary mixtures and 84 possible ternary mixtures under the assumption of additive responses. Retaining information on vapor concentrations in the classification models allows vapors to be accurately identified, while facilitating prediction of the concentrations of individual vapors and the vapors comprising the mixtures. The effects on the rates of correct classification of placing constraints on the maximum and minimum vapor concentrations and superimposing error on the sensor responses are investigated.

Introduction

The use of sensor arrays coupled with pattern-recognition analysis constitutes an effective approach to enhancing the selectivity and range of applications of chemical sensors. It has been demonstrated for several gas/sensor technologies that a single sensor array can provide unique response patterns for a number of different individual species or mixtures of species [1–9]. However, few reports have addressed the problem of determining both the identity and concentration of the components of gas/vapor mixtures [7–9].

We are interested in the development of portable instrumentation based on microfabricated chemical sensors for monitoring personnel exposures to toxic organic solvent vapors in the industrial environment. The use of polymer-coated surface-acoustic-wave (SAW) sensor arrays has several potential advantages for this application. First, SAW devices respond with high sensitivity to changes of surface mass and, as a result, can be used for a wide range of potential analytes [10]. Secondly, the amount of vapor sorbed by the sensor coating on the device is typically a linear function of the vapor concentration over the useful concentration range (i.e., less than a few hundred parts per million by volume). Thirdly, efficient operation is possible at ambient temperatures. Indeed, vapor sorption decreases exponentially with increasing temperature, so higher sensitivity is obtained at lower temperatures. Finally, since the response to a given vapor will depend strongly on the functional groups incorporated into the structure of the polymer, judicious selection of polymers can lead to significant differences in the response patterns for different vapors [11].

Various methods have been developed for correlating the pattern of responses from an array of chemical sensors with the identity or class of a given analyte [12–14]. Typically, principal-component analysis (PCA) and cluster analysis (CA) are performed on the concentration-normalized matrix of sensor responses to assess qualitatively the uniqueness of the response pattern for each species. Ideally, responses for different analytes will cluster in different regions of n-dimensional space, where n is the number of sensors used. One of several classification methods can then be used to identify an unknown, provided that its sensor responses are contained in the calibration set (also referred to as the training set). Criteria for classification are established using supervised learning methods such as the K-nearest-neighbor (KNN) technnique or the linear learning machine (LLM). For mixtures, responses will usually trace a locus of points between those of the individual components [2, 5]. For an unknown mixture to be correctly identified it is necessary

*Present address: Department of Nuclear Medicine, University of Massachusetts Medical Center, Worcester, MA 01655, USA.

0925-4005/93/$6.00

to have previously defined the spatial locations associated with the mixture over the range of component concentrations.

Once the identity of an unknown is determined, a second analysis, such as multiple linear regression (MLR), partial least squares (PLS) or principal-component regression (PCR), is performed to determine the concentration(s) of the analyte(s) [7, 14]. For sensors exhibiting non-linear responses with concentration, the data can be transformed in order to linearize the responses [3, 8]. Use of MLR on matrices containing redundant or collinear sensor responses may lead to large quantitation errors, whereas PLS and PCR are less sensitive to collinearity [14].

Disjoint principal-components modelling [15] and its more familiar derivative SIMCA (soft independent modelling of class analogy) [12, 13, 16] incorporate several features of PCR. In these methods, principal-components models are developed for individual groups within a data set. Classification of an unknown is based on the goodness of fit of its response vector to each of the models. This approach differs from standard PCR, where principal components are derived from the data matrix as a whole. Although these methods have been used for classification in a number of analytical chemical applications [17–19], they have not been applied to the analysis of data from sensor arrays.

As will be shown here, the concepts underlying these methods can be extended to permit identification of both individual vapors and the components of vapor mixtures from the sensor response patterns. In this extended disjoint principal-components regression (EDPCR) method advantage is taken of the integration of the qualitative and quantitative aspects of the sensor responses. Since information on the vapor concentrations is retained in the classification models, misclassification can be minimized and estimation of vapor concentrations is facilitated. In addition, by using a single model for the responses to each vapor, the data matrix can be summarized by a series of equations and the computational burden is reduced.

This paper describes the implementation of EDPCR for sensor-array analysis using previously reported data from an array of polymer-coated SAW sensors exposed to each of several vapors. Experimental responses to the individual vapors are combined to simulate all possible binary vapor mixtures and predictions are then made of both the identities and concentrations of vapors. Misclassification rates are examined with and without constraints placed on the vapor concentrations and then with varying amounts of Gaussian error superimposed on the sensor responses. Finally, the accuracy of classification is determined for the case where all possible ternary mixtures are included in the test set.

Data description

The data used here were published by Rose-Pehrsson *et al.* [5] and consist of the responses of ten polymer-coated SAW sensors to each of nine vapors. The vapors examined were dimethylmethylphosphonate (DMMP), dimethylacetamide (DMAC), dichloroethylene (DCE), diethyl sulfide (DES), water (H_2O), isooctane (ISO), toluene (TOL), 1-butanol (1BTL), and 2-butanone (2BTN). The original report contains the structures of the polymer coatings, which are designated here only as P1–P10.

Sensor responses are presented in Table 1 as the equilibrium frequency shifts caused by exposure to the vapors, Δf_v (in Hz), divided by the frequency shifts caused by initial deposition of the polymer coating on each sensor, Δf_c (in kHz). Presenting the data in this manner accounts for the differences in response arising from variations in the amount of polymer deposited on each sensor [5].

The relationship between the airborne vapor concentration, C_v, and the sensor response is given by

$$C_v = \Delta f_v \rho_c / (\Delta f_c k_m K) \qquad (1)$$

where ρ_c is the polymer density, K is the polymer/air partition coefficient of the vapor and k_m is a constant reflecting the potential role of changes in the polymer modulus on the response of the sensor [20].

The data matrix consists of sensor responses to each of four or eight concentrations of each vapor. Each row can be represented as a response vector composed of the ten sensor responses to a given concentration of a given vapor. For the purpose of analysis, we define a group as the set of response vectors (rows) for a given vapor over all concentrations.

In the original study a number of binary mixtures was also examined, but the mixture responses were not published and therefore could not be incorporated into our analyses. However, the authors state that the responses to vapor mixtures were additive (i.e., linear combinations of the pure vapor responses) or approximately additive in all but a few cases. Thus, the simulation of mixtures performed here under the assumption of additivity has a reasonable foundation.

Since the measured responses consisted of difference frequencies between the coated sensors and uncoated reference sensors, positive values reflect a reduction in the coated-sensor frequency. For most of the vapors the difference-frequency responses were positive, as expected for the increase in mass and/or softening of the polymer coating accompanying vapor sorption. For a few vapor/polymer pairs negative frequency shifts were observed, due either to significant vapor adsorption on the uncoated surface of the reference sensor or to an alternative response mechanism operating on the coated

TABLE 1. Sensor responses in Hz/kHz [5] and linear regression correlation coefficients

Group	Vapor	Conc. (μg/l)	Polymer coatings									
			P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
1	DMAC	81	67.10	9.10	3.97	−0.68	2.60	3.32	158.80	−4.58	2.57	0.47
		40	57.00	7.00	2.97	−0.39	1.26	1.91	124.99	−3.81	1.34	0.29
		20	33.94	3.27	1.54	−0.24	0.55	0.93	92.21	−2.74	0.62	0.22
		11	18.83	1.69	0.81	−0.17	0.28	0.54	69.54	−1.96	0.34	0.17
		(r^2)	(0.846)	(0.903)	(0.921)	(0.999)	(0.999)	(0.993)	(0.948)	(0.900)	(0.998)	(0.998)
2	DMMP	2230	257.62	90.66	38.94	−2.64	45.25	44.84	939.92	−8.46	39.53	4.31
		1100	207.48	56.64	25.05	−2.25	24.92	24.07	414.63	−8.24	21.42	2.72
		559	160.80	31.99	14.44	−1.88	14.11	11.48	346.16	−7.97	10.32	1.09
		372	142.23	24.28	10.84	−2.01	11.48	8.01	321.69	−8.09	6.73	0.74
		137	128.81	20.50	8.31	−1.45	7.02	3.90	228.80	−5.95	2.95	1.13
		84	102.79	13.44	5.53	−1.15	4.97	2.32	183.50	−4.90	1.63	1.70
		52	83.89	8.51	4.02	−0.84	3.62	1.53	157.92	−4.27	0.93	0.44
		29	58.70	4.54	2.33	−0.62	2.44	0.80	124.49	−3.18	0.44	0.40
		(r^2)	(0.873)	(0.979)	(0.979)	(0.715)	(0.996)	(0.998)	(0.964)	(0.496)	(0.998)	(0.870)
3	DCE	176000	35.59	56.62	113.18	16.99	47.49	115.79	69.05	52.82	122.03	22.27
		88600	19.35	30.51	72.07	8.05	23.44	55.64	29.07	21.96	54.86	15.50
		45700	8.98	14.34	39.72	3.80	11.03	27.12	11.99	8.86	23.42	7.78
		30500	6.05	9.04	24.75	2.49	7.78	18.45	8.19	5.97	16.54	5.76
		(r^2)	(0.997)	(0.997)	(0.979)	(0.999)	(0.999)	(0.999)	(0.996)	(0.996)	(0.998)	(0.959)
4	DES	137000	50.54	13.02	48.54	5.86	17.02	66.28	133.40	72.88	35.56	5.33
		68700	26.68	6.99	33.85	2.14	8.50	32.46	69.59	32.56	34.51	3.85
		35400	11.83	3.14	18.40	0.85	4.13	15.63	32.91	13.55	14.84	1.85
		23600	8.29	2.31	12.94	0.32	2.80	10.96	22.75	8.39	9.81	1.21
		(r^2)	(0.997)	(0.997)	(0.961)	(0.994)	(0.999)	(0.999)	(0.999)	(0.999)	(0.728)	(0.936)
5	H_2O	7180	9.71	8.23	2.44	35.23	1.99	1.95	20.49	−2.20	1.35	24.44
		3550	4.86	5.41	0.46	12.85	0.32	0.28	7.57	−1.46	0.72	5.79
		1820	2.51	3.63	0.02	4.91	−0.28	−0.01	4.04	−1.20	0.52	−0.40
		1210	1.89	2.93	−0.03	2.86	−0.38	−0.13	1.58	−1.49	0.24	−2.31
		(r^2)	(0.999)	(0.995)	(0.961)	(0.992)	(0.989)	(0.958)	(0.989)	(0.815)	(0.977)	(0.990)
6	ISO	129000	13.43	0.54	7.33	0.82	1.48	21.25	35.17	78.95	25.22	0.54
		64300	5.90	0.06	4.78	0.61	0.74	10.78	14.88	33.49	11.63	1.02
		33100	2.52	0.04	2.67	0.45	0.38	5.10	6.20	13.84	4.90	0.65
		22000	1.77	0.06	1.98	0.13	0.29	3.57	3.76	9.19	3.42	0.56
		(r^2)	(0.998)	(0.867)	(0.982)	(0.818)	(0.999)	(0.999)	(0.998)	(0.997)	(0.999)	(0.008)
7	TOL	57300	22.07	11.73	40.07	3.22	17.94	69.59	56.83	77.31	77.83	4.07
		28400	14.53	7.41	29.82	1.28	8.67	33.95	29.95	37.10	40.68	3.29
		14500	7.68	3.91	17.24	0.56	4.11	16.52	13.32	15.66	17.89	1.95
		9680	5.11	2.68	12.11	0.17	2.91	11.58	7.84	9.36	11.15	1.25
		(r^2)	(0.971)	(0.982)	(0.938)	(0.998)	(0.999)	(0.999)	(0.996)	(0.999)	(0.997)	(0.883)
8	1BTL	8730	31.90	8.06	19.37	3.46	8.92	13.68	104.22	−4.72	13.26	7.90
		4300	15.45	3.94	11.29	1.64	5.25	6.71	62.13	−4.10	7.01	4.78
		2200	7.46	2.04	6.05	0.91	2.98	4.29	52.67	−3.68	5.36	2.61
		1460	5.42	1.53	4.43	0.50	2.37	2.84	29.02	−3.04	2.87	2.33
		(r^2)	(0.999)	(0.999)	(0.994)	(0.998)	(0.997)	(0.997)	(0.952)	(0.888)	(0.979)	(0.993)
9	2BTN	148000	185.01	45.37	55.67	10.36	41.18	67.73	325.67	17.58	66.37	11.15
		74800	110.72	22.77	35.05	5.69	19.57	35.52	199.64	5.81	30.60	9.33
		38700	65.33	11.27	18.91	2.80	9.68	15.83	136.41	1.32	15.22	4.48
		25800	45.78	7.51	12.62	1.49	6.88	9.73	111.75	−0.13	10.30	3.81
		(r^2)	(0.995)	(0.999)	(0.984)	(0.994)	(0.999)	(0.998)	(0.999)	(0.995)	(0.998)	(0.869)

sensor (e.g., stiffening of the polymer film). Regardless, all data were used in the analyses presented below.

Method of analysis

The first step in the method is the application of PCA to each group of sensor responses (i.e., the collection of sensor responses to all concentrations of a single vapor). The response vectors for that vapor are then modelled using the most significant principal component(s). Rather than normalizing the response vector by its magnitude (i.e., concentration normalization) or autoscaling, as is commonly performed prior to PCA [4, 5], the only preprocessing that is performed is mean-centering (i.e., subtraction of the mean response vector from the individual response vectors for a group). Therefore, information about the concentrations of the vapors is accessible during classification.

The model used to classify vapor i is given by

$$r = m_i + \sum_{n=1}^{N} \alpha_{i,n} v_{i,n} + \epsilon \qquad (2)$$

where r is the response vector for the vapor at a given concentration, m_i is the mean response vector determined from all of the calibration concentrations measured for that vapor, $\alpha_{i,n}$ is the projection coefficient corresponding to the location of each response vector along the nth principal component represented by the unit vector $v_{i,n}$, ϵ is the residual error vector of the model for the vapor at the measured concentration, and N is the number of principal components.

Where it is unclear how many principal components should be included, a cross-validation method can be used to obtain the optimal number [12]. The accuracy of the model can be assessed merely by inspection of the residual error or by construction of confidence intervals.

Provided that at least some of the sensors in the array respond differently to a given vapor, each vapor will be represented by a unique response model. The response vector from a given concentration of an unknown vapor (r_u) can be tested for its goodness of fit to each of the models established during calibration by replacing r by r_u in eqn. (2) and solving for α_u so that $\|\epsilon\|_2$ is minimized (see Appendix 1). The identity of the unknown is determined from the model for which the smallest $\|\epsilon\|_2$ is obtained. Additional criteria can be imposed by constraining the residual error vector within a certain range, which would be more appropriate when there is no prior knowledge about the nature of the unknown vapor (i.e., where it may not belong to any of the vapors in the calibration set).

Once the identity of the vapor has been established, its concentration can be readily determined by interpolation or by linear regression of α_i versus concentration, because the projection of the response vector along the line (one-principal-component model) is directly related to its concentration when sensor responses are linearly related to concentration. Although this condition generally holds for low concentrations of most organic solvent vapors measured with polymer-coated SAW sensors [5], certain vapors may exhibit non-linear response characteristics. Transformation would then be necessary for accurate quantitation. Figure 1 illustrates the classification and quantitation processes with this approach.

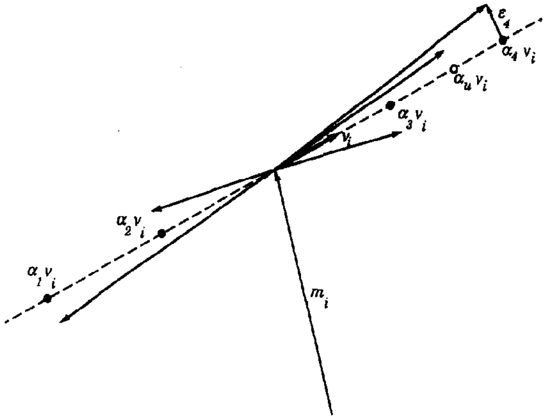

Fig. 1. Graphical representation of a one-principal-component model for an individual vapor showing the four calibration responses (filled symbols), the residual error from projection onto the principal component v (represented by ϵ_4), and the projection of a hypothetical unknown belonging to this class (open symbol). The vapor concentration is proportional to the distance along the principal component.

For mixtures of two components whose responses are additive, the response vector can be projected onto the plane bounded by the pure-vapor response vectors. Each binary mixture can then be thought of as an additional group consisting of two pure vapors, i and j, in some combination of concentrations. A classification model for a binary mixture can be established using the following equation (note: a one-principal-component model is assumed and the subscript $n = N = 1$ has been omitted for simplicity):

$$r = m_i + m_j + \alpha_i v_i + \alpha_j v_j + \epsilon \qquad (3)$$

where r represents the response vector for the mixture of i and j and the remaining variables are defined as above for the pure-vapor case. For each combination of vapor concentrations there will be specific values of α_i and α_j. Here again, the smallest residual error obtained by fitting an unknown response vector to all possible models determines the correct class for the unknown vapor.

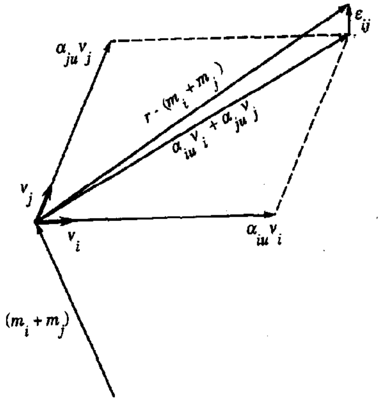

Fig. 2. Graphical representation of a one-principal-component model for a binary mixture and the projections used to determine the concentration of each component (see Appendix 1).

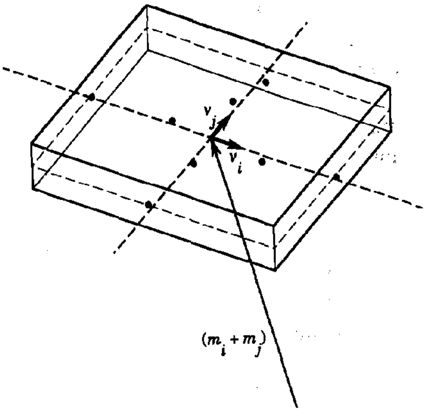

Fig. 3. Binary-mixture classification space with constraints placed on the maximum and minimum concentrations of both vapors in the mixture.

The response vector for the binary mixture can be projected onto the principal component of each vapor and the concentrations can be obtained from the calibration data for the individual vapors. Note that the principal component for one vapor may not be orthogonal to that of the other vapor and this must be taken into account in determining the vapor concentrations (see Fig. 2 and Appendix 1). Models analogous to eqn. (3) can be used for ternary or more complex mixtures.

Use of the concentration predictions in the classification scheme can aid in minimizing classification errors. Misclassification can arise from the intersection of the vector corresponding to one vapor (or vapor mixture) with those of other vapors. Where the response vectors extend to infinite concentration, there is an increased likelihood of intersection. In practical situations there is a limit to the concentrations encountered. In fact, standard quality-control protocols demand that an instrument detector be calibrated over a concentration range that brackets the range to be encountered during normal operation. Measurements obtained outside the calibration range are not strictly valid.

Criteria for establishing the maximum values for hazardous vapors might be based on some multiple of the allowable exposure limit (e.g., the Threshold Limit Value [21] or Permissible Exposure Limit [22]). Constraints on the minimum response would logically be based on the limit of quantitation attainable with the sensor array which, in turn, would be determined by the signal-to-noise ratio. Since the range of possible concentrations is rendered finite, the probability of misclassification is reduced. For a binary mixture, such constraints would result in a 'box' such as that shown in Fig. 3, which accounts for the possibility that residual errors would place the sample vector above or below the ideal mixture plane. An unknown sample initially classified into this group but falling outside these concentration limits would be re-classified to the group for which the next-lowest residual error is obtained and for which the concentration falls within the permitted range.

All computations described here were performed by one of the authors (T.S.P.) either on a DEC-5000 workstation using programs written in PV ~ WAVE® (Precision Visuals, Inc., Boulder, CO) or on a personal computer using programs written in MATLAB® (Mathworks, Inc., Natick, MA).

Results and discussion

Classification of individual vapors

PCA was performed on each group in Table 1. Table 2 shows that the percentage of the total variance explained by the first principal component for each group

TABLE 2. Variance explained by the first principal component of responses to each vapor and the residual errors of the classification models

Vapor	Variance on first principal component (%)	Mean residual error (Hz/kHz)	Mean relative residual error (%)
DMAC	99.4	2.53	2.1
DMMP	99.0	20.6	7.4
DCE	99.4	5.16	7.0
DES	98.9	5.39	7.8
H_2O	99.9	0.59	6.5
ISO	99.9	0.57	2.5
TOL	99.7	2.41	4.8
1BTL	99.1	2.45	5.2
2BTN	99.9	3.44	1.8

was very high. Therefore, a one-principal-component model of the form shown in eqn. (2) with $N = 1$ was adopted for subsequent analyses.

The four response vectors for each vapor (or eight for DMMP) were regressed with the principal components to determine values of α_i and ϵ for each concentration of each vapor i. Table 2 presents the mean residual error of the model for each group in absolute and relative terms, where the relative errors are the averages of the ratios of the residual-error vectors to the response vectors. The relative errors are low ($<8\%$) in all cases, however, the absolute error for DMMP is significantly larger than those for the other vapors. The larger absolute error obtained for DMMP can be attributed to the larger absolute sensor responses and to non-linearities in the responses of several of the sensors as a function of vapor concentration. Linear regression correlation coefficients (r^2) were calculated for each of the sensor responses and are presented in Table 1. Although most sensor responses are linear functions of the vapor concentrations, exceptions are seen for nearly all vapors. Deviations from linearity are most apparent with DMMP, for which seven of the ten r^2 values are below 0.99. Inspection of the individual sensor-response curves (not shown) confirms the exceptional behavior of DMMP. The small r^2 value obtained for ISO with the P10 sensor has a negligible effect on the fit to the model, apparently because of the very low response for this specific sensor.

Plots of the projection coefficients (α_i) versus concentration were linear ($r^2 > 0.992$) for all but three of the vapors. Those for the exceptional vapors are shown in Fig. 4(a)–(c) (note: each point represents the projection of the ten-sensor response vector onto the principal component for that vapor at the indicated concentration). Linear plots for most of the vapors would be expected because the projection coefficients are estimated from combinations of the individual (linear) sensor responses. The predominance of the non-linear sensor responses to DMMP is reflected in Fig. 4(a).

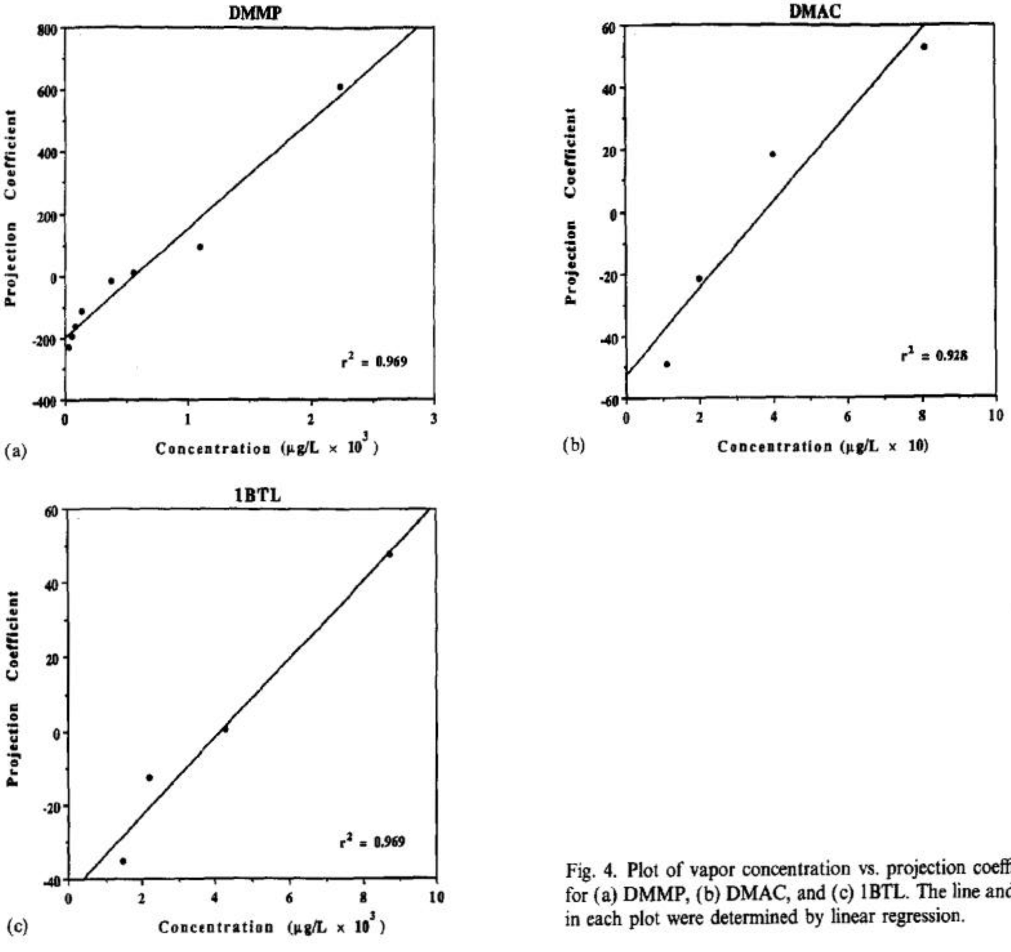

Fig. 4. Plot of vapor concentration vs. projection coefficient, α_i, for (a) DMMP, (b) DMAC, and (c) 1BTL. The line and r^2 value in each plot were determined by linear regression.

Surprisingly, the plots for DMAC (Fig. 4(b)) and 1BTL (Fig. 4(c)) are also clearly non-linear. For DMAC, most of the individual sensors gave linear responses. The shape of the curve in Fig. 4(b) reflects the influence of a minority of sensors (P1–P3 and P7) that gave large non-linear responses. Similarly, for 1BTL the shape of Fig. 4(c) reflects that of the sensor giving the highest response (P7) to this vapor.

That a few sensors can have such a large influence on the overall response pattern arises from the fact that the response vectors were not autoscaled prior to modelling. Thus, sensors providing higher responses have a greater influence on the group response vector. The skewness created by the dominant sensors could be reduced by applying weighting factors to the sensor responses, but this was not performed for the analyses here.

DMMP and DMAC are both high-boiling (> 160 °C) solvents containing polar functional groups. The shapes of Fig. 4(a) and (b) reflect the influence of these factors on their interactions with several of the sensor coatings. The plateaus in the responses at the intermediate-to-high concentrations indicate interactions with, and saturation of, specific sorption sites in the polymers. The subsequent inflection point for DMMP denotes the onset of another mode of sorption at higher concentrations. The DMAC curve suggests adherence to a Langmuir or Freundlich sorption model, while the DMMP curve resembles that of a BET sorption model [23]. For 1BTL, no such physical interpretation is possible and it is suspected that the non-linearity is the result of an error in calibration for the third-highest concentration. Despite the non-linearities in their responses, the residual errors for the DMAC and 1BTL models are still quite small (Table 2).

Table 3 presents the normalized mean residual errors obtained by fitting the responses from one group to each of the other groups. The mean residual error for the correct group was used as the basis for normalization. Each row contains the error obtained on attempting to fit the vapor corresponding to that row to the models for the vapors listed at the top of each column. As expected, the between-group errors are larger than the within-group errors in all cases. Note, however, that for DMMP the errors obtained in attempting to classify it into the other groups are invariably smaller than those for the other vapors. This further confirms that the DMMP data are more scattered about the model values than the data for the other vapors. It also indicates that the response vectors for DMMP fall in fairly close proximity to those for several of the other vapors.

Classification/quantitation of individual vapors and vapor-mixture components

To examine the use of EDPCR for classifying the pure vapors and their binary mixtures, a test set was created using the response vector for the third-highest concentration from each group. Response vectors were then calculated for all possible binary mixtures by combining the responses for the individual vapors. This yielded 45 ‘unknown’ response vectors consisting of the nine individual vapors and the 36 binary mixtures. The limited amount of data available precluded the use of separate data for the test set. Each of these unknown vectors was then tested for goodness of fit to all possible groups using eqns. (2) and (3).

With no constraints placed on the concentrations of the vapors, 36 of the 45 unknowns (80%) were correctly classified. The classification test was repeated with the added constraint that the predicted concentrations had to be greater than zero. In this case the correct classification rate was 38/45, or 84%. Most misclassifications (six out of seven) involved DMMP: pure DMMP was misclassified as a mixture of DMMP with DMAC; and in four of the five remaining misclassifications involving DMMP the error resulted from assignment of the unknown as a mixture of DMAC with another vapor rather than DMMP with another vapor. The misclassifications are consistent with the data in Table 3, which show that the residual between-group:within group error ratio for classification of DMMP as DMAC is quite small (2.2 Hz/kHz).

The residual errors for these six misclassifications were relatively large (≈ 7–16 Hz/kHz) compared to

TABLE 3. Ratios of within-group to between-group residual errors

Vapor	DMAC	DMMP	DCE	DES	H_2O	ISO	TOL	1BTL	2BTN
DMAC	1	17	44	24	10	17	28	6.6	21
DMMP	2.2	1	6.4	4.3	3.2	3.8	4.8	3.2	3.0
DCE	23	67	1	11	5.4	8.2	6.9	12	37
DES	14	42	13	1	4.4	5.7	6.1	6.8	21
H_2O	180	564	182	141	1	70	120	96	337
ISO	194	595	163	101	35	1	69	106	345
TOL	45	138	21	17	9.3	11	1	23	76
1BTL	12	38	41	20	8.7	17	26	1	17
2BTN	5.4	30	34	18	12	17	24	5.7	1

those for the 38 correctly classified samples (generally <5 Hz/kHz). This suggests that a limit could be placed on the maximum allowable residual error as a threshold for classification. The same test set was analyzed again with the further constraint that the minimum and maximum vapor concentrations could not exceed those in the calibration set. In this case, the correct classification rate increased to 43/45, or 96%. The two misclassifications involved mixtures of DMMP with other vapors. The increase in correct classifications as increasing constraints are placed on the concentrations of the vapors illustrates the advantage of retaining and utilizing concentration information in the classification analysis.

For the above analyses, the test set used to examine the classification models was extracted from the calibration set. To represent practical conditions more accurately, where sensors might exhibit random fluctuations about their true values, tests of classification were repeated with the addition of varying amounts of Gaussian error to the sensor responses. For each level of superimposed error, 100 simulations were performed. The rates of correct classification are presented in Table 4. Standard deviations were less than 6.3%, indicating that the estimation procedure was reasonably stable. The correct classification rate is seen to decline fairly rapidly with added error. To put these results into perspective, a second classification method based on the K-nearest-neighbor technique ($K = 1$) was applied to the data (after normalizing the data matrix). Table 4 also presents the classification rates using the KNN technique. Note that, by definition, the classification rate with the KNN method was 100% for the case of no superimposed error, because the samples used in the test set were identical to those in the calibration set. Comparing the results for the two methods shows that as the error increases, the correct classification rates are virtually the same for both methods. The slightly lower rates obtained with the EDPCR method reflect modelling error; however, this is offset by the reduction in the number of computations involved in classification relative to the KNN method.

Concentrations were then predicted for the 43 samples correctly classified with the EDPCR method under the last test condition (i.e., maximum and minimum bounds). 77 predictions were made, corresponding to 68 binary-mixture components (i.e., 2 × 34 correctly classified binary-mixture components) and 9 individual vapors. Projection coefficients were determined as described in Appendix 1 and concentrations were then predicted by linear regression using the relationship between the projection coefficients (α_i) and the vapor concentrations discussed above. The results showed that 66 of the 77 (86%) values were within 20% of the actual concentration values and most were within 10%. Of the 11 concentrations that were not accurately predicted, 8 of these involved 1BTL. In fact, none of the concentration predictions for 1BTL was accurate. As can be seen in Fig. 4(c), this is an artifact of having chosen the third-highest concentration for 1BTL, which is well above the regression line. Since the prediction for pure 1BTL was in error, the predictions of 1BTL concentrations in the mixtures were also in error. The other three concentrations that were not predicted accurately involved mixtures of DMMP and DMAC with each other or with other solvents.

In practice, the approach taken above for concentration predictions would likely be modified. That is, for those vapors giving linear plots of concentration versus α_i, the concentration values obtained from linear regression would be used to create a calibration curve that would account for the slight deviations of the individual points from the overall linear trend. This approach would reduce the errors obtained with 1BTL and would also improve the concentration predictions for most of the other vapors. For DMAC and DMMP, whose projection coefficients vary non-linearly with concentration, a log

TABLE 4. Correct classification rates (based on 100 trials) as a function of superimposed Gaussian error applied to the sensor responses for EDPCR and KNN methods

Superimposed error	EDPCR		KNN	
	Correct classification	Standard deviation	Correct classification	Standard deviation
0	95.5	0	100	0
5	88.6	3.6	91.9	3.5
10	78.7	4.5	80.6	4.7
15	68.9	5.2	70.8	5.2
20	60.2	6.3	61.7	5.9
25	52.4	6.2	53.9	5.8
30	46.6	5.6	47.8	6.2

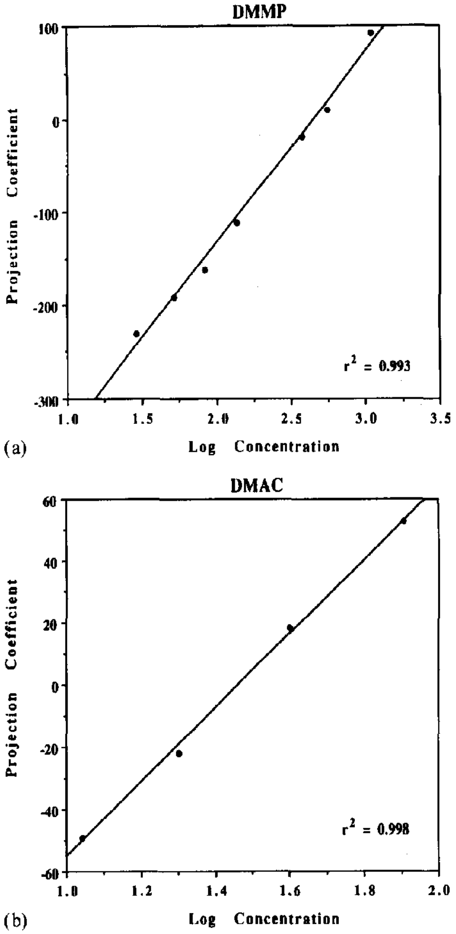

Fig. 5. Plots of log concentration vs. projection coefficient for (a) DMMP (highest concentration omitted) and (b) DMAC. The line and r^2 value in each plot were determined by linear regression.

transformation could be performed prior to quantitation. Figure 5(a) and (b) shows that plots of log concentration versus α_i for these vapors (omitting the highest DMMP concentration) describe the data better than the regression lines in Fig. 4(a) and (b). Commensurate improvements in the accuracy of predicted concentrations would be expected.

A final analysis was performed by including all possible ternary mixtures in the test set. Correct classification rates using the third sample from each of the individual response vectors from a group were 67, 71 and 85% for the unconstrained, partially constrained (i.e., positive values only) and fully constrained (maximum and minimum bounds) conditions, respectively. For the last condition, 18 of the 19 misclassifications involved DMMP and 15 of these involved DMMP in ternary mixtures with other vapors. Correct classifications were obtained for the remaining 110 vapors and vapor mixtures.

Summary and conclusions

The EDPCR method described here is an alternative to other methods that have been used for analyzing sensor-array responses. It is particularly well suited for arrays of polymer-coated SAW sensors where the responses to individual vapors are linear and responses to mixtures of vapors are additive. Deviations from linearity can be tolerated without adversely affecting the classification outcomes, but in extreme cases linearizing transformations may be required as a pre-processing step. Accommodation for non-additive mixture responses should also be possible with this method, provided that some proportionality in responses is maintained between the mixture components over the relevant concentration ranges.

The EDPCR method takes advantage of the inherent structure of the response data in both the classification and quantitation procedures, which improves the computational efficiency. Additional samples are easily incorporated into the data set, since the principal-components models are based on the individual vapor responses.

Due to limitations on the data available, the test set examined here was derived from the calibration set and mixture responses were simulated rather than generated experimentally. Work is currently underway to generate a data set for a wide range of organic vapors using several different polymer coatings that will yield separate calibration and test samples. Mixture responses are also being collected to test the validity of the additivity assumption further. EDPCR analyses of these data will be the topic of a subsequent report.

Acknowledgements

Funding for this work was provided by Grant No. K01OH00077 from the National Institute for Occupational Safety and Health of the Centers for Disease Control and by Contract No. 02112409 from the US Department of Energy.

References

1 H. Sundgren, I. Lundström, F. Winquist, I. Lukkari, R. Carlsson and S. Wold, Evaluation of a multiple gas mixture with a simple MOSFET gas sensor array and pattern recognition, *Sensors and Actuators B*, *2* (1990) 115–123.

2 R. Muller, High electronic selectivity obtainable with nonselective chemosensors, *Sensors and Actuators B*, *4* (1991) 35–39.

3 J. W. Gardner, Detection of vapours and odours from a multisensor array using pattern recognition. Part 1. Principal component and cluster analysis, *Sensors and Actuators B*, *4* (1991) 109–115.

4 J. Stetter, J. P. Jurs and S. L. Rose, Detection of hazardous gases and vapors: pattern recognition analysis of data from an electrochemical sensor array, *Anal. Chem.*, *58* (1986) 860–869.

5 S. L. Rose-Pehrsson, J. W. Grate, D. S. Ballantine, Jr. and P. C. Jurs, Detection of hazardous vapors including mixtures using pattern recognition analysis of responses from surface acoustic wave devices, *Anal. Chem.*, *60* (1988) 2801–2811.

6 D. S. Ballantine, Jr., S. L. Rose, J. W. Grate and H. Wohltjen, Correlation of surface acoustic wave device coating responses with solubility properties and chemical structure using pattern recognition, *Anal. Chem.*, *58* (1986) 3058–3066.

7 W. P. Carey, K. R. Beebe and B. R. Kowalski, Multicomponent analysis using an array of piezoelectric crystal sensors, *Anal. Chem.*, *59* (1987) 1529–1534.

8 G. Horner and Chr. Hierold, Gas analysis by partial model building, *Sensors and Actuators B*, *2* (1990) 173–184.

9 Chr. Hierold and G. Horner, Quantitative analysis of gas mixtures with non-selective gas sensors, *Sensors and Actuators*, *17* (1989) 587–592.

10 M. S. Nieuwenhuisen and A. Venema, Surface acoustic wave chemical sensors, *Sensors Mater.*, *1* (1989) 261–300.

11 J. W. Grate and M. H. Abraham, Solubility interactions and the design of chemically selective sorbent coatings for chemical sensors and arrays, *Sensors and Actuators B*, *3* (1991) 85–112.

12 M. A. Sharaf, D. L. Illman and B. R. Kowalski, *Chemometrics*, Wiley–Interscience, New York, 1986.

13 D. L. Massart, B. G. M. VanDeginste, S. N. Demming, Y. Michotte and L. Kaufman, *Chemometrics: A Textbook*, Elsevier, Amsterdam, 1988, pp. 403–407.

14 W. P. Carey, K. R. Beebe, E. Sanchez, P. Geladi and B. R. Kowalski, Chemometric analysis of multisensor arrays, *Sensors and Actuators*, *9* (1986) 223–234.

15 S. Wold, Pattern recognition by means of disjoint principal components models, *Pattern Recognition*, *8* (1978) 127–139.

16 C. Albano, W. Dunn, III, U. Endlund, E. Johansson, B. Norden, M. Sjostrom and S. Wold, Four levels of pattern recognition, *Anal. Chim. Acta*, *103* (1978) 429–443.

17 W. J. Dunn III, S. L. Emery, W. G. Glen and D. R. Scott, Preprocessing, variable selection, and classification rules in the application of SIMCA pattern recognition to mass spectral data, *Environ. Sci. Technol.*, *23* (1989) 1499–1505.

18 N. B. Vogt, F. Brakstad, K. Thrane, S. Nordenson, J. Krane, E. Aamot, K. Kolset, K. Esbensen and E. Steinnes, Polycyclic aromatic hydrocarbons in soil and air: statistical analysis and classification by the SIMCA method, *Environ. Sci. Technol.*, *21* (1987) 35–44.

19 D. R. Scott, Determination of chemical classes from mass spectra of toxic organic compounds by SIMCA pattern recognition and information theory, *Anal. Chem.*, *58* (1986) 881–890.

20 J. W. Grate, M. Klusty, R. A. McGill, M. H. Abraham, G. Whiting and J. Andonian-Haftvan, The predominant role of sorbent phase swelling or modulus changes in determining the responses of polymer-coated surface acoustic wave vapor sensors, *Anal. Chem.*, *64* (1992) 610–624.

21 American Conference of Governmental Industrial Hygienists, *Threshold Limit Values for Chemical Substances and Physical Agents and Biological Exposure Indices*, ACGIH, Cincinnati, OH (1991–1992).

22 OSHA Final Rule Air Contaminants Permissible Exposure Limits, *Code of Federal Regulations, 29CFR1910.1000*, US Department of Labor, Occupational Safety and Health Administration, Jan. 19, 1989.

23 C. E. Rogers, in J. Comyn (ed.), *Polymer Permeability*, Elsevier Applied Science, London, 1985, Ch. 2, pp. 11–74.

24 K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press, Orlando, FL, 1972, p. 30.

Appendix 1

The procedure for determining the projection coefficients used to fit an unknown response vector r_u to models based on pure vapor i or a mixture of vapors i and j using a one-principal-component model is described. From eqn. (2), classification of r_u as a specific vapor i is indicated if

$$r_u = m_i + \alpha_{ui} v_i + \epsilon_u \tag{A1}$$

and the Euclidean distance $\|\epsilon_u\|$ is small. From the orthogonality principle [24], $\|\epsilon_u\|$ is minimized when ϵ_u is orthogonal to v_i, which means that the inner product of ϵ_u and v_i is zero, i.e., $(\epsilon_u, v_i) = 0$. Therefore, taking the inner product of eqn. (A1) with v_i and solving for α_{ui}, we have

$$\alpha_{ui} = (r_u - m_i, v_i) \tag{A2}$$

Similarly, if r_u consists of a mixture of vapors i and j, r_u can be expressed as

$$r_u = m_i + m_j + \alpha_{ui} v_i + \alpha_{uj} v_j + \epsilon_u \tag{A3}$$

and, again, $\|\epsilon_u\|$ has to be small. In this case, ϵ_u has to be orthogonal to v_i and v_j, i.e., $(\epsilon_u, v_i) = 0$ and $(\epsilon_u, v_j) = 0$. Taking the inner product of eqn. (A3) with v_i and v_j separately,

$$(r_u - m_i - m_j, v_i) = \alpha_{ui} + \alpha_{uj}(v_j, v_i) \tag{A4a}$$

and

$$(r_u - m_i - m_j, v_j) = \alpha_{ui}(v_j, v_i) + \alpha_{uj} \tag{A4b}$$

one can derive, in matrix notation,

$$\begin{bmatrix} \alpha_{ui} \\ \alpha_{uj} \end{bmatrix} = \begin{bmatrix} 1 & (v_i, v_j) \\ (v_i, v_j) & 1 \end{bmatrix}^{-1} \begin{bmatrix} (r_u - m_i - m_j, v_i) \\ (r_u - m_i - m_j, v_j) \end{bmatrix} \tag{A5}$$

Inversion of the matrix is performed to account for the fact that v_i and v_j may not be orthogonal.

Biographies

Edward T. Zellers earned a BA degree in chemistry from Rutgers University in 1978 and MS (1984) and PhD (1987) degrees in environmental health science from the University of California, Berkeley. From 1978 to 1981 he worked at Bell Telephone Laboratories on the synthesis of electrically conducting organic materials. Since 1987 he has been an assistant professor of occupational health at the University of Michigan, School of Public Health. His primary research interests are in the occupational and environmental health applications of chemical sensors.

Tin-Su Pan earned BS and MS degrees in electrical engineering from National Tsing-Hua University, Taiwan, and a PhD degree in electrical engineering in 1991 from the University of Michigan. Since then he has been working as a postdoctoral research scientist in the Department of Nuclear Medicine at the University of Massachusetts Medical Center, Worcester, MA. His research interests include pattern recognition, medical imaging and signal processing.

Samuel J. Patrash earned a BS degree in chemistry from the University of Michigan in 1977. He was employed for ten years as a research chemist in government and industry and is currently a doctoral candidate in environmental health science at the University of Michigan, working on the development of coated SAW sensor arrays for monitoring toxic organic vapors.

Mingwei Han earned a BSE degree in chemical engineering in 1985 from Tsinghua University in Beijing, China and an MS degree in chemical engineering in 1988 from the University of Michigan. He is presently working toward a doctoral degree in industrial health at the same university on the development of acoustic-wave chemical sensors for occupational health applications.

Stuart A. Batterman earned a BS degree in environmental science from Rutgers University in 1979, and MS (1981) and PhD (1986) degrees in civil engineering from the Massachusetts Institute of Technology. He is currently an assistant professor of environmental health science at the University of Michigan, with research interests in the development of models and sampling and analytical methodologies for environmental pollutants.