Predicting organoleptic scores of sub-ppm flavour notes Part 2.† Computational analysis and results

The Analyst

Timothy C. Pearce[a] and Julian W. Gardner[b]

[a] *Department of Engineering, University of Leicester, University Road, Leicester, UK LE1 7RH*
[b] *Electrical and Electronic Division, School of Engineering, University of Warwick, Coventry, UK CV4 7AL*

Received 28th May 1998, Accepted 22nd July 1998

In Part 1 of this paper (T. C. Pearce and J. W. Gardner, *Analyst*, 1998, **123**, 2047) we describe a novel method for predicting the organoleptic scores of complex odours using an array of non-specific chemosensors. The application of this method to characterising beer flavour is demonstrated here by way of predicting a single organoleptic score as defined under the joint EBC/ASBC/MBAA international flavour wheel for beer. An experimental study was designed to test the accuracy of the odour mapping technique for this prediction of organoleptic scores of added reference compounds within a chemically complex lager beer background. Using the flow injection analyser (FIA) system comprising 24 conducting polymer sensors, also described in Part 1, sampling was conducted on spiked lager beers. A dimethyl sulfide spike was added at the 20–80 ppb v/v level to simulate a range of organoleptic scores (0–5.5 out of 10) for flavour note no. 0730—"cooked vegetable". A certain amount of sensor drift was observed over the 12 d testing period which is shown to account for significant variance in the data-set as a whole. The effect of the sensor drift was reduced by applying a linear drift model, which may be generally applied when the between-class variance due to the difference in odours is small when compared with the within-class variance due to the drift, which increases approximately linearly over time. Careful use of this drift compensation model, coupled with judicious selection of pre-processing and pattern recognition techniques, maximised the between-class variance and so improved the overall classification performance of the system. After applying detailed exploratory data analysis, statistical, and neural classifier techniques, the organoleptic score was predicted with an accuracy of ±1.4 (out of 10) and 95% confidence. Our results show that it is possible to generate subjectively defined organoleptic flavour information using multi-sensor arrays and associated data-processing that is comparable in accuracy to sensory and GC-based techniques.

Background

Since a unified theory of odour analysis does not yet exist, which would be able to explain and predict the sensory impact of simple and complex odours, we are left to describe odours in prototypical terms—by comparison of similar odours. This fact alone has had a tremendous impact on the evolution of flavour analysis which has resorted to two very disparate approaches—instrumental and sensory-based. Unfortunately, the use of these techniques, in isolation, has been found wanting in terms of providing an overall picture of flavour or odour quality. A recent trend towards combined instrumental and sensory analysis has shown some promise in being able to relate a chemical description of flavour, on the one hand, with a perceptual (sensory) description, on the other. As part of this trend we have described (in Part 1 of this paper, ref. 1) an odour mapping technique for correlating the response of chemically sensitive multi-sensor arrays with organoleptic flavour scores as a method for generating a more detailed characterisation of odour quality. This approach leverages upon concepts used in sensory analysis, in particular flavour terminology. The use of flavour terminology systems as a product specific model of odour representation is optimal in view of the lack of any universal standards for odour description (sometimes referred to as a set of primary odours).

In order to assess the efficacy of this technique in generating useful organoleptic data we have designed a flow injection analyser (FIA), based around an array of conducting polymer sensors, as part of an experimental study described in Part 1.[1] In this study lager beer samples have been spiked with a range of concentrations of reference compounds derived from the EBC/ASBC/MBAA international flavour terminology system for beer, to simulate samples with three varying organoleptic flavour notes.[2] Using data collected from the FIA rig sampling on the spiked beer samples with flavour note scores estimated from known sensory-panel sensitivities, the task is then to be able to predict the intensity of the organoleptic flavour note for unseen data using the odour mapping technique described previously. In Part 2 of this paper, we describe the computational analysis for this study and detail the results. Finally, we compare this technique with existing flavour analysis technologies.

Many previous studies reporting upon the application of sensor arrays for odour analysis have provided minimal detail of the reasoning used in the selection of statistical and pattern recognition techniques. Furthermore, there has been little justification for the order in which these are applied during the analysis. In this paper, we outline the typical sequence of operations to be applied and the reasoning behind the use of each technique. In order to approach this in a systematic manner, the data analysis for the flavour note characterisation study was split into three successive stages. Firstly, exploratory data analysis was conducted in order to gain understanding of the data-set under investigation, and so guide the subsequent analysis. Second, a classification of the control lager against the same lager spiked with the three flavour notes was attempted. Finally, the prediction of the organoleptic scores for each spiked

† For Part 1, see ref. 1.

beer sample was attempted in order to illustrate the use of an electronic nose for generating humanly defined odour information.

Exploratory data analysis

Sensor pre-processing

Firstly, in order to investigate odour class discrimination, only the highest concentration of each flavour note was considered (as defined in Table 1 in Part 1, ref. 1). Sensor and array pre-processing of the data was carried out. So as not to introduce bias into the analysis a wide variety of pre-processing metrics were considered at each stage of exploratory data analysis

$$\upsilon'_{kj} = (\upsilon^{\text{in}}_{kj} - \upsilon^{\text{in}}_{kj}) \tag{1}$$

$$\upsilon'_{kj} = \frac{\upsilon^{\text{fi}}_{kj}}{\upsilon^{\text{in}}_{kj}} \tag{2}$$

$$\upsilon'_{kj} = \frac{\upsilon^{\text{fi}}_{kj} - \upsilon^{\text{in}}_{kj}}{\upsilon^{\text{in}}_{kj}} \tag{3}$$

where $\upsilon^{\text{in}}_{kj}$ and $\upsilon^{\text{fi}}_{kj}$ are the initial and final values (resistance or conductance) attained for each sensor, k, in response to an odour sample, j. In particular, difference, relative, and fractional metrics were calculated for both sensor conductance and resistance using eqns. (1), (2) and (3), respectively. Further steady-state sensor parameters for consideration were generated by array normalisation across all sensors

$$\upsilon''_{kj} = \frac{\upsilon'_{kj}}{\sqrt{\sum_{k=1}^{k=R}\left(\upsilon'_{kj}\right)^2}} \tag{4}$$

where R is the total number of sensors in the array. This reduces the concentration dependence of the signal, as well as sensor autoranging using

$$\upsilon'_{kj} = \frac{\upsilon'_{kj} - \min_j\left(\upsilon'_{kj}\right)}{\max_j\left(\upsilon'_{kj}\right) - \min_j\left(\upsilon'_{kj}\right)} \tag{5}$$

which is a popular pre-processing metric for use before input to a statistical classifier in order to maximise the spread of the data in the input space.

The means and standard deviations (s) for the above metrics were then plotted (in addition to the values $\upsilon^{\text{in}}_{kj}$ and $\upsilon^{\text{fi}}_{kj}$) on a per class basis in order to screen the data for outliers as well as illustrate any obvious class differences. For example, the polar plots of the fractional change in conductance, calculated using eqn. (3), are shown in Figs. 1 and 2, for the control lager and control lager with each of the added flavour spikes under consideration. The values range between 0–9% of baseline, although the largest response from sensors 9 and 10 (PAN/$NaHSO_4$/H_2O gave highly variable values between 150–230%) were too large to be displayed. A single dashed line either side of the mean indicates a confidence interval of 1 s indicating the repeatability of the sensor response on a test-to-test basis. As before, this confidence interval is seen to vary depending upon sensor type. Although the mean response values vary slightly between different odour classes, (seen by comparing the values in Fig. 1a with Fig. 1b and Fig. 2a,b) it is clear that these excursions do not extend beyond the confidence intervals shown and so a very high degree of overlap in response is present within the data-set. This is also true of all other pre-processing metrics treated in a similar fashion.

Principal components analysis

Before applying principal component analysis (PCA) and a host of other parametric statistical techniques, the data-set must conform to the requirements of multivariate normality, homogeneity of the covariance matrices, and independence of observations. Multivariate normality was tested by examining the multivariate scatterplots between all 21 dependent variables (*i.e.*, sensor values) and histograms. While, not surprisingly, the sensor readings showed high correlation across the array there were no non-linear relationships violating the assumption of multivariate normality, the histograms of the raw baseline resistance values were highly skewed and non-Gaussian. This is thought to be a result of the time-dependent sensor drift acting as a systematic error on the odour measurement (thus spreading the distribution and causing asymmetry), and is later shown to be removed in the compensated values. Consequently multivariate analyses of variance (MANOVA) carried out on the drift compensated values are likely to be more meaningful than those on the raw values. Homogeneity of the covariance matrix was verified using Bartlett's χ^2 test, that showed no significant differences between the within-group covariance matrices for each class. Independence of observations was achieved in the data-set by using a Latin Squares experimental design during sampling, as well as sufficient recovery time between tests for the sensors to return to their baseline values (see the Experimental section in Part 1, ref. 1).

PCA was first applied to all types of pre-processed data in order to analyse the nature of the sample variance. Firstly, a scree test of the eigenvalues was used in order to examine the measurement efficiency (or parsimony) of the problem. The scree plot shown in Fig. 3a illustrates that most of the total variance is accounted for by the first six eigenvalues and their corresponding principal components. Here we used Kaiser's stopping rule of extracting components until $\lambda < 1$.[3] This is

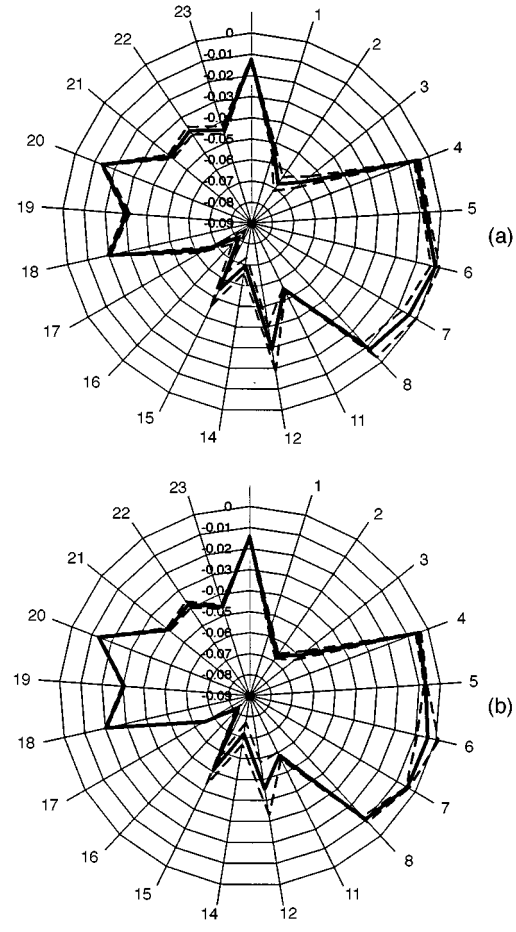

Fig. 1 Polar plots of fractional change in conductance metric eqn. (3) (a) Bass Carling Black Label control lager, (b) control lager spiked with 0.4 ppm 2,3-butanedione. Thick lines represent mean values over six tests and dashed lines show a single s from the mean for each sensor.

equivalent to extracting components until the point where the next component is contributing less variance than that of the average sensor, which is a reasonable assumption in this context since we are using PCA as a dimensionality reduction technique. However, a variety of other stopping rules could have equally been applied (such as Bartlett's sphericity test,[4] cross-validation methods described by Krzanowski,[5] or simply a percentage of variance criterion) and the choice of how many components to extract is always subjective. The cumulative variance explained by these first six principal components is given in Table 1, accounting for a large portion of the total (85.3%).

In order to visualise any class separation present in the data-set, the first three principal components were then plotted, as shown in the example for fractional change in conductance values shown in Fig. 3b, which accounts for the majority of the total variance (66.1%). Class labels have been assigned in order to illustrate class separation (described in the caption of Fig. 3). Little, if any, separation was evident between the samples using PCA. This was also true for a similar treatment of all other varieties of pre-processing metrics. This result indicated that further interpretation of the eigenvectors for the problem was required.

Using an oblique eigenvector rotation method, it was possible to interpret the set of six principal components in terms of the original variables. Rotation methods transform the set of *q*

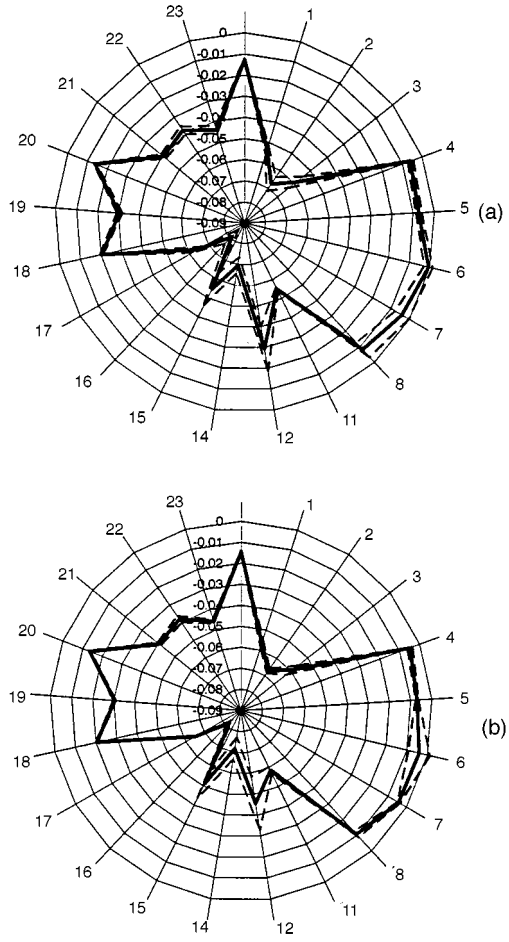

Fig. 2 Polar plots of fractional change in conductance metric eqn. (3) (a) control lager spiked with 80 ppb dimethyl sulfide, and (b) control lager spiked with 400 ppm hop essence. Thick lines represent mean values over six tests and dashed lines show a single *s* for each sensor.

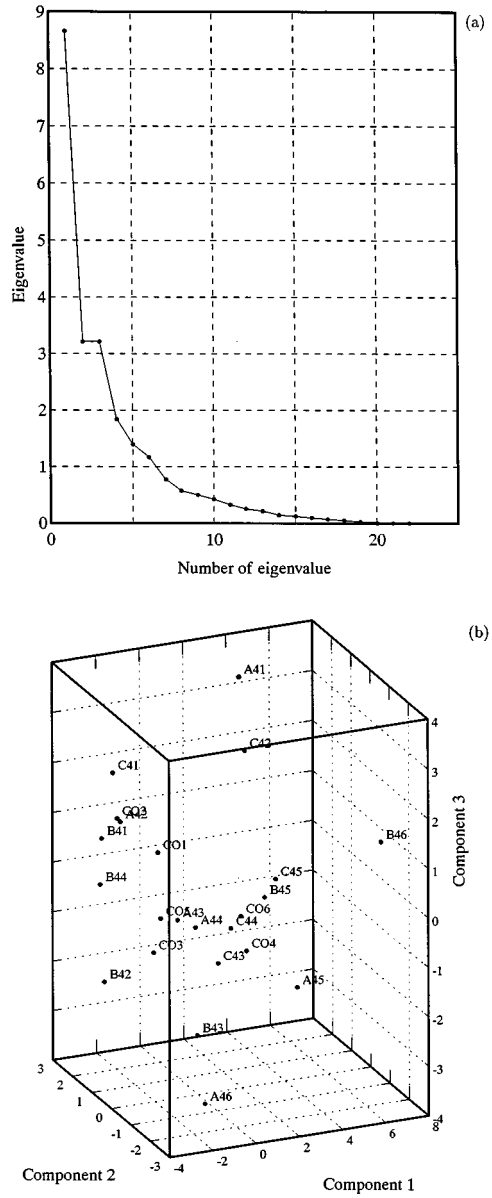

Fig. 3 Results of PCA applied to the fractional change in conductance metric eqn. (3) (a) scree plot of eigenvalues showing their relative importance, and (b) plot of the first three principal components. Class labels assigned as CO, control lager; A, diacetyl spike; B, dimethyl sulfide spike; and C, hop essence spike. First number in label indicates spike intensity (1–4), with 4 corresponding to 0.4 ppm v/v 2,3-butanedione for class A, 80 ppb dimethyl sulfide for class B, and 400 ppm hop essence for class C. Second number indicates its sequence during sampling (1–6).

Table 1 Variance explained by the first six principal components of the fractional change in conductance metric eqn. (3). Contributing sensors found using oblique eigenvector rotation

Component No.	Variance	Cum. variance	%	Cumulative %	Sensors contributing
1	8.70367	8.70367	37.84	37.84	3, 4, 9, 10, 16, 17
2	3.24940	1.19531	14.13	51.97	12, 13, 14
3	3.24366	1.51967	14.10	66.07	1, 22, 23
4	1.84440	1.70411	8.02	74.09	5, 6, 18, 19, 20, 21
5	1.40214	1.84433	6.10	80.19	8, 11
6	1.18069	1.96240	5.13	85.32	7

orthogonal eigenvectors (defining a sub-space $\boldsymbol{R}^q$) to be as closely aligned with the co-ordinate system for the original set of R variables (defining space $\boldsymbol{R}^R$) as possible. This results in so-called simple structure, whereby the eigenvectors may be easily interpreted in terms of the original variables. While varimax and quartimax methods preserve eigenvector orthogonality, oblique methods allow some breakdown in the orthogonality of the new sub-space in order to obtain a simpler structure for interpretation. After oblique rotation of the first six eigenvectors, Table 1 includes those sensors that largely contribute to each component. For example, sensors 9 and 10, which were removed from the polar plot (Figs. 1 and 2) due to exceptionally large responses combined with large variance, are by far the largest contributors to the first principal component.

By comparing the contributing sensors in the PCA analysis with the confidence intervals given in Figs. 1 and 2, it is clear that the PCA procedure is largely explaining the within-class variance, as opposed to the desired between-class variance. While PCA is fitting the variance in the data-set very well, any between-class variance is being ameliorated (and thus lost) by the dominant within-class variance (caused by system noise or systematic drift in the data-set), thus leading to poor class separation.

A closer analysis of the original data-set highlighted significant temporal drift over the eleven days of testing, accounting for much of the problematic within-class variance. The drift in the baseline resistance values υ_{kj}^{in} over the period (as measured before the onset of each test) is quantified in Table 2. While the drift is typically modest (0.5–3.9% from baseline over a 14 d period), sensors 7–13 demonstrated poor stability. Indeed, sensors 9, 10 and 13 drifted off scale and were removed before any further analysis. As could be expected, sensor pairs corresponding to each type of polymer show broadly similar stability properties. These differences in device stability between different varieties of polymer film could result from the variety of growth conditions under which the devices were prepared, or from subsequent polymer-specific interaction with the test odours.

Parametric drift compensation

By comparing the baseline resistance values from test-to-test it was possible to characterise the nature of the temporal drift contributing to the problematic within-class variance. Other attempts have been made to model drift in metal oxide gas sensor arrays.[6,7] Fig. 4a shows the fluctuation in the baseline over the course of the test period for each sensor, which show the days (1–12) of the testing period and resistance value. The fluctuation in the final values (over all sample classes) is shown in Fig. 4b. Clearly, these data are well-behaved and linearly dependent upon time, as indicated by the lines of best fit. The day histogram shows the density of tests taken over the 12 d testing period, showing a fairly consistent sampling rate as expected. The sensor histograms give the density of sensor resistance values over their range throughout the study. These are useful for spotting outliers. For example, gross discontinuities in the histogram of Fig. 4b for sensors 7 and 16 show outliers in the final values over time, leading to the poor linear fit for these sensors.

For the case of sensor resistance, it is clear that the drift over time can be modelled here by

$$\upsilon^{fi} = \upsilon_o^{fi} + \delta\upsilon^{fi}(t) = \upsilon_o^{fi} + \alpha^{fi}t \quad (6a)$$

$$\upsilon^{in} = \upsilon_o^{in} + \delta\upsilon^{in}(t) = \upsilon_o^{in} + \alpha^{in}t \quad (6b)$$

where υ_o^{fi} and υ_o^{in} are the intercepts at $t = 0$, and α^{fi} and α^{in} are the gradients of the linear fit shown in Fig. 4. The linear fit for the final values, in eqn. (6a) assumes that the differences in response of the sensors to the different classes of sample are negligible compared with the magnitude of the drift over time.

Similarly, for the difference pre-processing metric defined in eqn. (1), the drift over time is described by

$$\upsilon' = (\upsilon^{fi} - \upsilon^{in}) = \upsilon'_o + \delta\upsilon'(t) \quad (7)$$

where υ'_o is the first difference value observed at $t = 0$ and $\delta\upsilon'(t)$ is the time-dependent drift term. The drift terms are related as

$$\begin{aligned}\delta\upsilon' &= \sqrt{\left(\delta\upsilon^{fi}\right)^2 + \left(\delta\upsilon^{in}\right)^2} \\ &= t\sqrt{\left(\alpha^{fi}\right)^2 + \left(\alpha^{in}\right)^2}\end{aligned} \quad (8)$$

and so the difference values also depend linearly upon time, and may be removed by a simple linear parametric compensation scheme. This relationship can be clearly seen by examining the values of the difference pre-processing metric over time shown in Fig. 5a.

However, for either the relative or fractional pre-processing metrics [defined by eqns. (2) and (3)] the drift terms become

$$\begin{aligned}\left(\frac{\delta\upsilon'}{\upsilon'}\right)^2 &= \left(\frac{\delta\upsilon^{fi}}{\upsilon^{fi}}\right)^2 + \left(\frac{\delta\upsilon^{in}}{\upsilon^{in}}\right)^2 \\ &= \left(\frac{\alpha^{fi}t}{\upsilon_o^{fi} + \alpha^{fi}t}\right)^2 + \left(\frac{\alpha^{in}t}{\upsilon_o^{in} + \alpha^{in}t}\right)^2\end{aligned} \quad (9)$$

and so the drift in these measures varies nonlinearly (second order polynomial) with time. This nonlinear relationship is obvious from the time dependence of the fractional values shown in Fig. 5b.

Parametric drift compensation using this simple model was then performed on the initial and final values in order to remove the problematic within-class variance caused by the sensor

Table 2 Percentage change in sensor baseline resistances υ_{kj}^{in} caused by systematic sensor drift over period of testing

No.	Sensor type	Baseline drift (%)	No.	Sensor type	Baseline drift (%)
1	PPy/TEATS/H_2O	2.14	13	PPy/PSA/H_2O	N/A[a]
2	PPy/TEATS/H_2O	2.5	14	PPy/PSA/H_2O	10.7
3	PPy/TEATS/PC	2.6	15	PPy/HpSA/H_2O	2.8
4	PPy/TEATS/PC	2.3	16	PPy/HpSA/H_2O	1.7
5	PPy/pTSA/EtOH	1.2	17	PPy/HxSA/H_2O	3.9
6	PPy/pTSA/EtOH	1.5	18	PPy/HxSA/H_2O	2.5
7	P3MT/TEATFB/CH_3CN	119.6	19	PPy/DSA/H_2O	1.1
8	P3MT/TEATFB/CH_3CN	62.5	20	PPy/DSA/H_2O	1.9
9	PAN/$NaHSO_4$/H_2O	N/A[a]	21	PPy/pTSA/H_2O	0.5
10	PAN/$NaHSO_4$/H_2O	N/A[a]	22	PPy/pTSA/H_2O	1.6
11	PPy/BSA/H_2O	11.8	23	PPy/DSA/H_2O	2.3
12	PPy/BSA/H_2O	14.1	24	PPy/DSA/H_2O	2.1

[a] Sensor removed during study due to drift causing response beyond ADC scale.

instability. The compensated values $\tilde{\upsilon}_{kj}^{fi}$ and $\tilde{\upsilon}_{kj}^{in}$ were calculated by removing the linear drift term defined in eqn. (6)

$$\tilde{\upsilon}_{kj} = (\upsilon_{kj} - \alpha t_i) \quad (10)$$

where t_i is the time at which the ith measurement was taken. Again, this procedure was carried out on all varieties of pre-processing metric. An example of the efficiency of the technique for removing the time-dependent drift is shown by the polar plots in Fig. 6. As before, polar plots show the mean value with one s either side of the mean indicating the variance of samples on a test-to-test basis. While the within-class variance for the raw baseline resistance values is sizeable, the drift compensated values are shown to reduce this to negligible values across the array. After inspection, the poor fit of sensor 4 to linear drift compensation was caused by a single outlier which was later removed. The compensated and the raw values will be compared during the forthcoming classification study.

This method can be generally applied as a simple parametric drift compensation scheme, assuming that the magnitude of the drift variance is far larger than the variance due to the sample differences. Also, the drift must be well-behaved, continuous, and show an approximate linear time dependence.

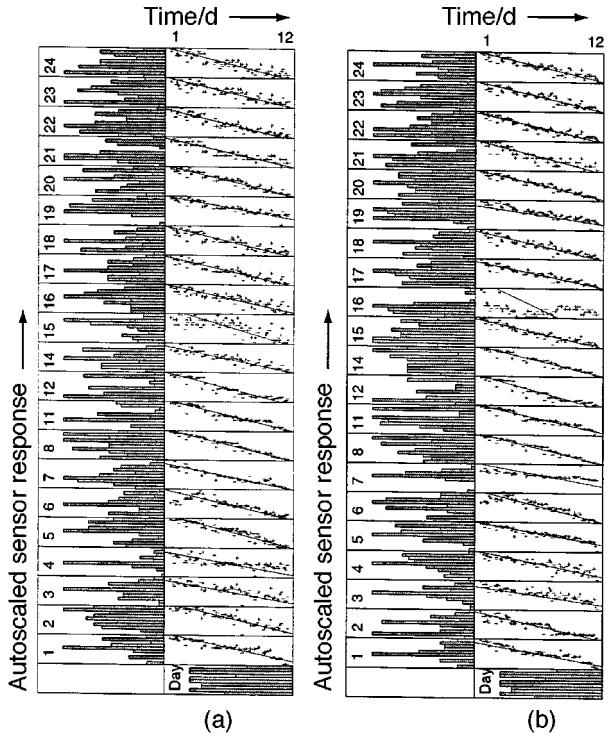

Fig. 4 Graphs of sensor drift with time over the period of the FIA study demonstrating a linear time dependence for (a) initial baseline resistance, υ_{kj}^{in}, and (b) final resistances, υ_{kj}^{fi}. Lines of best fit and histogram over sampling period are shown for each sensor.

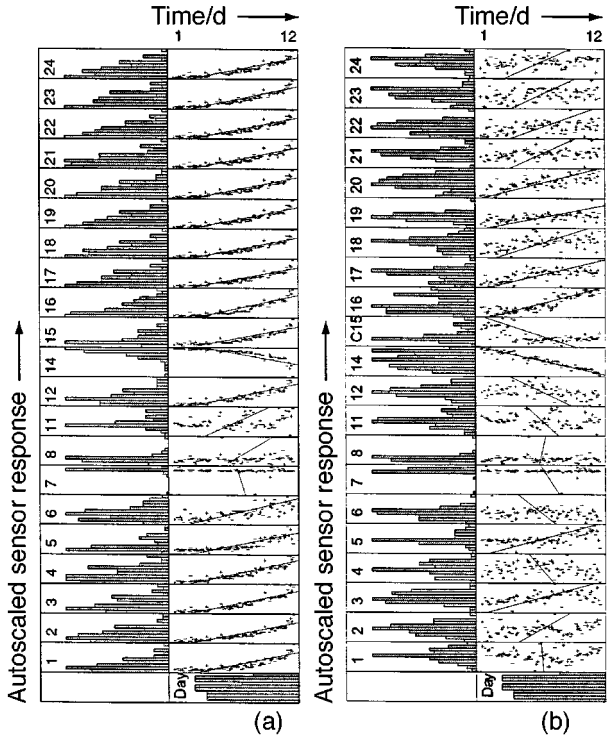

Fig. 5 The effects of systematic sensor drift on pre-processed response values, (a) array normalised difference in resistance showing linear time dependence, and (b) array normalised fractional change in resistance showing non-linear time dependence. Lines of best fit and histogram over sampling period are shown for each sensor.

Multivariate analysis of variance

In order to indicate any differences between the class means within the data-set, MANOVA was carried out. The resulting test statistic, Wilke's Lambda, Λ, represents the amount of variance in the data-set not explained by the independent variables when being used to highlight class separation.

The results of MANOVA as applied to the flavour note data-set are given in Table 3. For all of the varieties of pre-processing metric, values for Λ indicate that a large portion of the variance is explainable in terms of class separation. Overall, the drift

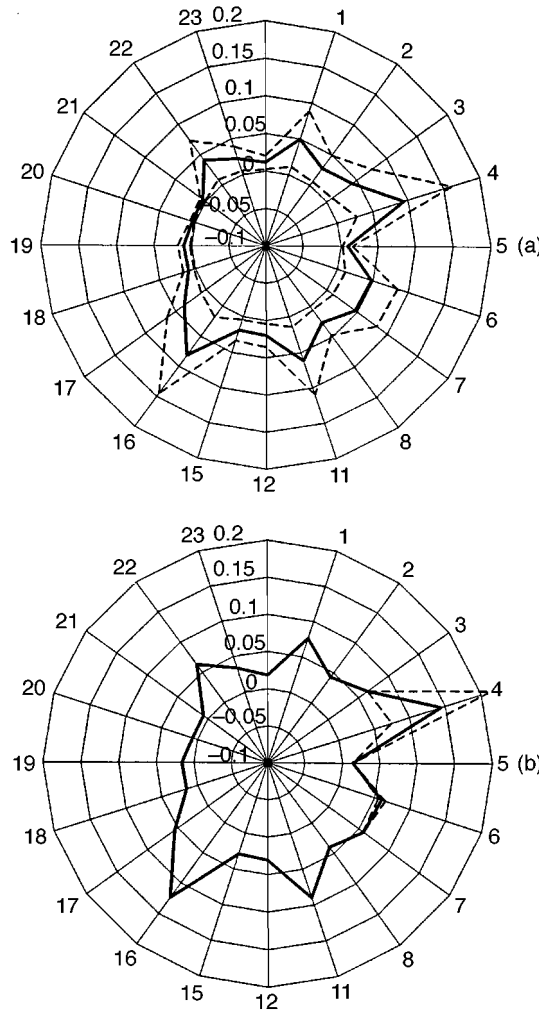

Fig. 6 Polar plots of array normalised baseline resistances prior to sampling control lager, showing the effect of linear parametric drift compensation in reducing problematic within-class variance, (a) raw values, and (b) drift compensated values.

compensated values show far better significance and portions of explained variance as compared to the raw values. However, for only one metric is the significance at a suitable level to demonstrate a clear effect; the array normalised difference in resistance after drift compensation. With a value of $p = 0.078$, the confidence interval of 93% indicates a strong likelihood of the difference between the means for each class being a result of class membership. By considering the process of drift compensation used in eqn. (8) and the fact that the difference metric drifts linearly with changes in initial or final values, it is clear that it provides the best choice of metric for maximising the between-group variance within conducting polymer arrays when subject to linear temporal sensor drift.

Discriminant function analysis

As MANOVA and discriminant function analysis (DFA) are also parametric techniques, all of the conditions already justified for PCA must hold. Furthermore, DFA is also highly sensitive to outliers in the data-set and so screening of the data was undertaken. During cluster analysis on the array normalised difference in resistance (compensated) values sample B22 was found to be a large distance (in cluster space) from any other data points as identified by particularly late clustering during the iterations of the algorithm. Although cluster analysis is not typically used as a method for identifying multivariate outliers (specific techniques such as Barnett's[8] M-, R-, P-, or C-ordering could have been used), after a closer examination of the values for B22, the readings for sensors 6, 10, and 11 were further than 3 *s* from the mean and so this point was considered to be an outlier and removed from the data-set. All other points showed reasonably close proximity in Euclidean space.

While the MANOVA results suggested some class separation for at least one of the pre-processing metrics, canonical DFA was used in order to visualise this class separation and isolate those sensors contributing most to class separation. The results are given in Table 4 for the three discriminant functions a_1, a_2, and a_3. Since there are only four odour classes (one control and three flavour spikes) only three functions were extracted, accounting for 100% of the total variance in the data-set. The percentage contribution of each discriminant function was calculated from its corresponding eigenvalue λ as

$$\frac{\lambda}{\Sigma_{k=1}^{3}\lambda_k} \tag{11}$$

where λ_i is the ith eigenvalue. This parameter indicates the relative importance of the variables in determining group separation. An F test also provides a p value (significance level) for each discriminant function, as given in Table 4. Clearly for a_1 then $p < 0.05$ indicates a good level of significance. However, both discriminant functions a_2 and a_3 have $p > 0.05$ and so must be considered with caution. By examining the standardised discriminant coefficients used to generate the canonical variables in terms of the original sensor values, it was also possible to identify which sensors contributed most significantly to class separation, also shown in Table 4.

Due to the low significance of the discriminant functions a_2 and a_3, the sensors contributing to class separation are principally, 2, 12, 17, 18, 19, 20, and 22. Interestingly, none of these sensors made large contributions to the total variance, which caused the earlier PCA analysis to fail, verifying our earlier hypothesis that the dominant principal components are modelling the sensor drift. Finally, the separation of class-labelled samples is visualised in Fig. 7, by plotting the first three discriminant functions. The four odour classes are clearly visible as being separated by linear DFA.

Table 3 Results of MANOVA for the FIA study. Metric type; D, difference calculated using eqn. (1), F, fractional calculated using eqn. (3). Measurement parameter: R, resistance; G, conductance. Normalisation: A, array normalised calculated using eqn. (4); N, none. Compensation: C, parametric drift compensation calculated using enq. (10); N, none. (All other varieties of pre-processing metrics resulted in a singular within-class residuals matrix and so could not be calculated)

Metric	Parameter	Normalisation	Compensation	Wilke's Λ	p
D	R	N	N	.33262	.549
D	R	N	C	.21813	.191
F	R	N	N	.33136	.541
F	R	N	C	.22443	.222
D	G	N	N	.32114	.476
D	G	N	C	.23495	.279
F	G	N	N	.33652	.573
F	G	N	C	.22643	.223
D	R	A	N	.31440	.433
D	R	A	C	.18845	.078
F	R	A	N	NA	N/A
F	R	A	C	.22863	.244
D	G	A	N	NA	N/A
D	G	A	C	.23876	.301
F	G	A	N	.32651	.510
F	G	A	C	.23418	.275

Flavour classification

Having carried out a thorough exploratory analysis of the FIA flavour note data-set, the next task was to attempt a classification of each of the samples into one of the four odour classes. Prescriptive DFA uses the canonical linear discriminant functions generated in the previous analysis in order to form a classification rule that assumes an equal cost of misclassification across each class, which is reasonable for this study. The confusion matrix (shown in Table 5) for such a DFA indicated an overall correct classification rate of 80.3%. However, this estimate of the classification error is optimistic, since all of the available samples were used in generating the classification rule. A better estimate of the true error (*i.e.*, to new data samples) is provided by cross-validation methods, thereby training on a subset of the sample data-set and testing on the remainder.

Although DFA is clearly able to discriminate between the trained samples, the classification accuracy after cross-validation (leave-one-out method) fell to 25% for control, 41% for diacetyl spike, 42% for dimethyl sulfide (DMS) spike, and 21% for hop essence. Cross-validation is used here as a technique to better approximate the true error of the classifier to unseen data. By making multiple partitions of the entire data-set into training and testing sets it is possible to obtain a more accurate measure of the likely performance of the system were it to be used, for example, in an on-line odour monitoring application. The classification rates are far lower than those for the non-validated

Table 4 Results of canonical DFA applied to FIA flavour note study

Discriminant function	Eigenvalue, λ	Contribution (%)	Cumulative (%)	Correlation, r	Significance level, p	Sensors contributing
a_1	1.28139	49.2	49.2	0.7494	0.0302	2, 12, 17, 18, 19, 20, 22
a_2	0.98154	37.7	86.8	0.7038	0.22	3, 4, 23, 24
a_3	0.34391	13.2	100.0	0.5059	0.78	16, 17

DFA results, but are still statistically significant when compared to chance for the diacetyl and DMS classes given the 61 remaining samples.

Multi-layer perceptron

In an attempt to improve the classification rates, a multi-layer perceptron (MLP) artificial neural network was applied to the data-set. The network comprised of 21 input nodes (corresponding to the number of sensors), three hidden processing units in a single layer, and four output processing units (corresponding to each odour class). Initially, the default recommended learning parameters were used [momentum term, μ = 0.4, learning rate, η = 0.3, squashing function $s(i)$ = tanh(i)] and the data were range-scaled to [−1, +1] in order to maximise coverage of the input value space into the network. The network error was recorded during training on the entire data-set and compared for the effect of pre-processing metric on convergence. For this study only the array processed metrics were considered in order to remove the concentration-dependence of the data-set (at the expense of increasing noise), and for the drift compensated values only the resistance metrics were considered as the sensor drift was found to be linearly dependent upon it. As shown in Fig. 8, poor convergence was demonstrated by all of the networks, except for the array normalised conductance difference model.

By using 3-fold cross-validation on this network, its ability at classifying unseen samples was tested. A confusion matrix of the results is given in Table 6, showing classification rates that were marginally above chance (again at 25%) for diacetyl and hop essence spike, but statistically significant for the DMS spike at 54%.

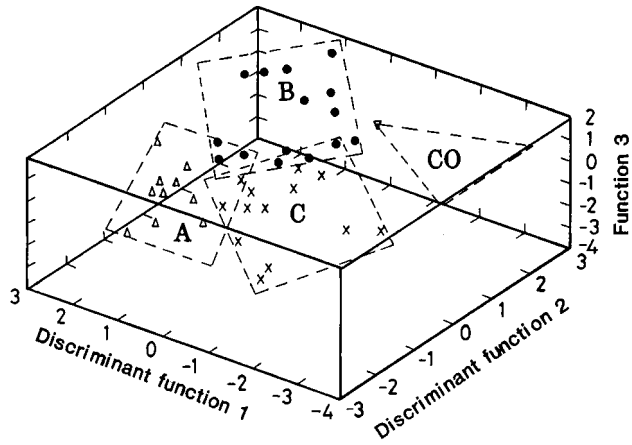

Fig. 7 Plot of the first three canonical discriminant variables after DFA was applied to the array normalised difference in resistance values (after linear drift compensation). The four separate odour classes are shown.

Further improvements in the classification rates were obtained by optimising the network and learning parameters used in the original MLP. Firstly, the effect of the number of hidden units on the unbiased estimate of the true error was considered. An MLP (with default learning parameters) was trained on the array normalised difference in conductance data (that demonstrated good convergence in the previous stage of optimisation), and the error rate observed during variation of the number of hidden units, as shown in Fig. 9a. Clearly, a single hidden unit provides by far the most consistent network error during training. However, the lowest network error was obtained using two hidden units after only 200 training iterations. These values were therefore used in future stages of optimisation.

Further experiments in optimising the response showed that little improvement in network error was obtained by varying the learning rate, η, and so this was maintained at the default value of 0.4. However, a marginal effect for the momentum term was obtained, as shown in Fig. 9b. The best selection of the learning term μ, was seen to be 0.3.

After optimisation of the learning and network parameters used in the MLP, an attempt was made to reclassify the samples. Table 7 shows the confusion matrix for the classification using the optimised network. Clearly, the MLP is making choices between two classes, diacetyl and DMS. While the classification results have improved substantially for both of these categories, with diacetyl being identified 52.4% and DMS 85.7% of the time (against chance at 25%), it was not possible to discriminate control and hop essence samples from any other samples investigated.

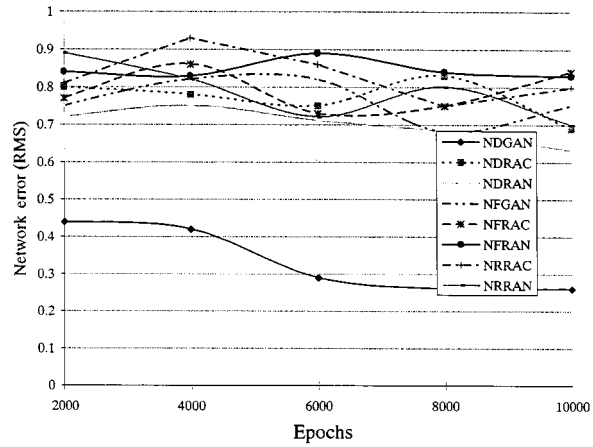

Fig. 8 The effect of type of pre-processing metric on the convergence of an MLP during training. Legend coding: 1st letter; N, no logarithm. 2nd letter: D, difference; F, fractional; R, relative. 3rd letter: R, resistance; G, conductance. 4th letter: A, array normalised. 5th letter: N, raw values; C, parametric linear drift compensation.

Table 5 Confusion matrix of classification results of control lager from control lager with three added flavour spikes using a classification rule based upon linear canonical DFA variables (no cross-validation used)

	Predicted class			
Actual class	Control	Diacetyl	DMS	Hop essence
Control	4	0	0	0
	100.0%	0.0%	0.0%	0.0%
Diacetyl	1	15	1	3
	5.0%	75.0%	5.0%	15.0%
DMS	2	1	17	1
	9.5%	4.8%	81.0%	4.8%
Hop essence	0	2	1	13
	0.0%	12.5%	6.3%	81.3%

Table 6 Confusion matrix of classification results of control lager from control lager with three added flavour spikes using an MLP with the array normalised difference in conductance metric (3-fold cross-validation used)

	Predicted class			
Actual class	Control	Diacetyl	DMS	Hop essence
Control	0	1	1	1
	0.0%	33.3%	33.3%	33.3%
Diacetyl	1	6	5	9
	0.5	27.9%	23.8%	42.8%
DMS	2	2	12	6
	0.9%	0.9%	54.0%	27.2%
Hop essence	0	5	5	4
	0.0%	35.7%	35.7%	28.6%

Flavour note characterisation

The problem of flavour note intensity prediction requires a mapping function f[A] between sensor space bounded by $\boldsymbol{V}$ and odour space bounded by $\boldsymbol{Y}$, such that $f: A \subset \boldsymbol{R}^R \rightarrow \boldsymbol{R}^P$. A full discussion of the theoretical background to the odour mapping technique is given in Part 1, ref. 1. In this study organoleptic data were available for the range of concentrations tested for all three flavour notes defining $\boldsymbol{Y}$, shown in Table 1 of Part 1 of this paper (ref. 1), although poor classification was obtained for diacetyl and, particularly, hop essence. Consequently organoleptic mapping was considered for only one of the flavour notes, namely DMS, that performed best during the odour classification study detailed above.

The theoretical organoleptic scores for each flavour note, as might be expected to result from an FPA trail conducted by a trained sensory panel, have been given in Table 1, in Part 1, ref. 1. That is, for the DMS concentration ranges of 20–80 ppb, the expected scores for the cooked vegetable note (ASBC no. 0720 on the flavour terminology wheel shown in Fig. 1 of Part 1, ref. 1) would be 0, 1, 3, and 5–6, respectively. In order to generate the required mapping function, f[A], between the array data and the organoleptic scores, the MLP was chosen to be trained upon the DMS samples only.

Initially an MLP with 21 input units (corresponding to each sensor), 2 hidden units in a single layer, and 1 output unit was trained using default learning parameters [μ = 0.4, η = 0.3, tanh activation function] using the target function, t

$$t = 1.8\left(\frac{I}{I_{\max}}\right) - 0.9 \qquad (12)$$

where I represents the perceived intensity of the flavour note or organoleptic score to varying concentrations of DMS and $I_{\max}$ represents the flavour score corresponding to the highest concentration of flavour spike investigated (5–6 for DMS). This function was used to condition the target values to lie between [−0.9 and +0.9], while still retaining the non-linear relationship between concentration and perceived intensity that has been modelled by Steven's power law of psychophysics[9]

$$R = c\,(C - C_0)^n \qquad (13)$$

where R represents the flavour intensity as reported by the perceiver and c is a constant of proportionality that is specific to the individual, C is the concentration and C_0 the discrimination threshold for the analyte, and n is an index that varies between 0.2–0.8 depending upon the odorant in question.

Convergence of the network to a wide variety of pre-processing parameters is shown in Fig. 10a. Obviously, no array normalised values were considered for flavour intensity prediction as these remove the concentration-dependence of the data. All of the pre-processing parameters led to network convergence, except the relative measures that caused paralysis in learning after only 300 training iterations, probably resulting from a local minima in the error space. The relative measures were therefore not considered further.

An unbiased estimate of the true network error (to unseen samples) was provided by using 6-fold cross-validation on the remaining six pre-processing metrics under consideration.

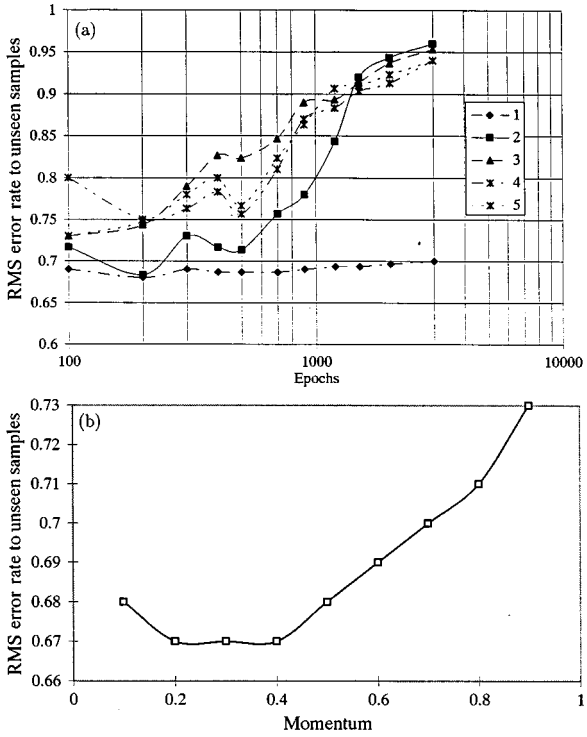

Fig. 9 The effect on the MLP network error after 3-fold cross-validation training on the array normalised difference in conductance data to (a) variation in the number of hidden units, and (b) variation in the momentum term.

Table 7 Confusion matrix of classification results of control lager from control lager with three added flavour spikes using an optimised MLP and the array normalised difference in conductance metric (3-fold cross-validation used)

	Predicted class			
Actual class	Control	Diacetyl	DMS	Hop essence
Control	0	0	3	0
	0.0%	0.0%	100.0%	0.0%
Diacetyl	0	11	10	0
	0.0	52.4%	47.6%	0.0%
DMS	0	3	18	0
	0.0%	14.3%	85.7%	0.0%
Hop essence	0	11	3	0
	0.0%	78.6%	21.4%	0.0%

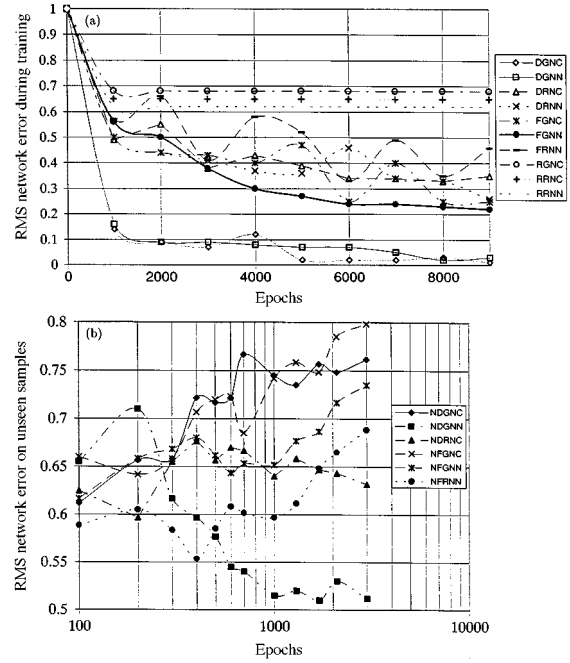

Fig. 10 (a) The effect of pre-processing metric on the network error during training. Legend coding: 1st letter; D, difference; F, fractional; R, relative. 2nd letter: R, resistance; G, conductance. 3rd letter: N, no normalisation. 4th letter: N, raw values; C, parametric linear drift compensation; (b) the effect of pre-processing metric on the unbiased estimate of the true error of an MLP during training on flavour intensity data. Legend coding: 1st letter; N, no logarithm. 2nd letter: D, difference; F, fractional; R, relative. 3rd letter: R, resistance; G, conductance. 4th letter: N, no normalisation. 5th letter: N, raw values; C, parametric linear drift compensation.

During training, the networks were tested on unseen data to provide an average RMS error to unseen samples, as shown in Fig. 10b. Similar to the case for odour classification, the results demonstrate that the difference in conductance pre-processing metric (without normalisation) is the most predictive of flavour note intensity.

Before proceeding to optimise the learning and network parameters for the network, the data-set was screened for outliers and extreme values. Box plots of the difference in conductance pre-processing metric for all sensors are shown in Fig. 11b. It is clear that a small subset of sensors were leading to unreliable results, in particular, sensors 7, 8, 9 and 22. The extreme values (> 5 *s* from the mean) were removed from the data-set, while the outliers were retained for further consideration.

As before, the network and learning parameters were optimised in order to further improve the prediction performance. Firstly, the number of hidden units were varied and again the true error rate estimated by using 6-fold cross-validation. The results of varying the number of hidden units for an MLP with default learning parameters and training, on the difference in conductance metric are shown in Fig. 11a. The effect of the number of hidden units on the root-mean-squared error is small, with two hidden units showing marginally better performance in predicting the unseen samples.

The momentum and learning terms were varied in order to optimise the prediction performance, as shown in Fig. 12. A clear effect for both parameters is shown, with the optimum values obtained when $\mu = \eta = 0.3$. After optimisation, then, the final MLP network selected for the flavour note prediction study was therefore a $21 \times 2 \times 1$ MLP with the learning parameters, $\mu = \eta = 0.3$, and tanh activation function, training on the difference in conductance values.

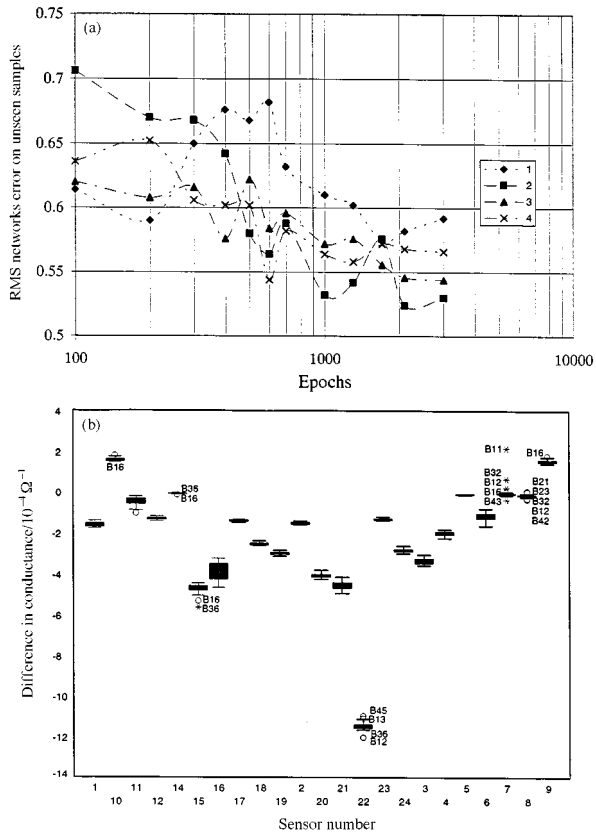

Fig. 11 (a) The effect of the number of hidden units on the unbiased estimate of the true MLP network error, and (b) box plots of the difference in conductance pre-processing metric, showing outliers and extreme values. Labels shown identify each sample: B, indicates DMS spiked samples; 1st number indicates concentration level (*i.e.*, 1 = 20 ppb, 3 = 40 ppb, 3 = 60 ppb, 4 = 80 ppb); 2nd number indicates sample number in series. Outliers: B13, B23, B21, B36, B45, B43. Extreme values: B11, B12, B16, B32.

Using the optimised MLP network, the organoleptic scores for each of the samples were then predicted using 6-fold cross-validation. After removal of the outliers identified using the box plots earlier, the performance of the MLP network for predicting the corresponding flavour note scores of the DMS spike is shown in Fig. 13. Clearly the MLP network is able to predict the organoleptic scores, although this becomes less accurate at higher concentrations of flavour spike. This inaccuracy could result from either the response of the sensor array to DMS saturating above 60 ppb, or that the neural network is having difficulty mapping onto the highly logarithmic flavour scores. In the latter case this may be remedied by the use of a logarithmic post-processing function.

Comparing flavour analysis technologies

Finally, a comparison between the accuracy obtained by FIA using the electronic nose system reported here with both FPA as

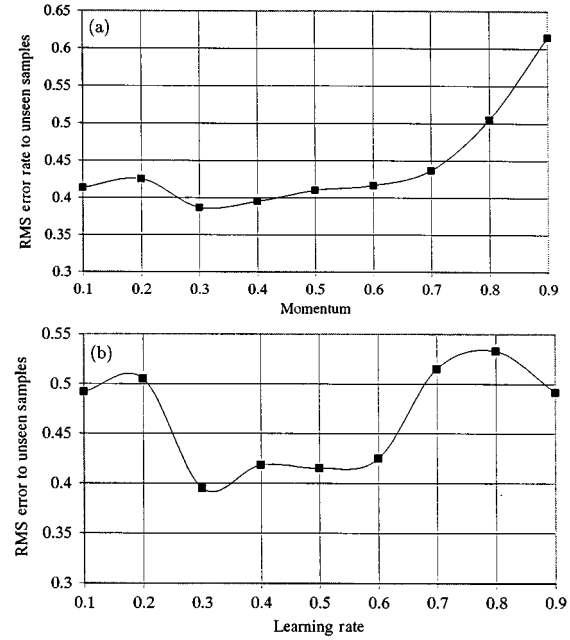

Fig. 12 The effect of varying, (a) μ, the momentum term (for $\eta = 0.3$), and (b) η, the learning term (for $\mu = 0.4$) on the prediction error for a $21 \times 2 \times 1$ MLP network training on the difference in conductance values.

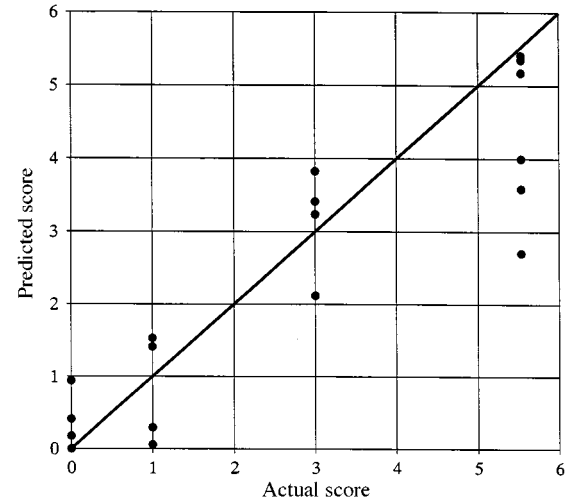

Fig. 13 Predicted organoleptic scores under the 'cooked veg' flavour note using a sensor array and MLP for lager beer spiked with four different levels of DMS (20, 40, 60, and 80 ppb), simulating a beer with scores of 0, 1, 3, 5–6.

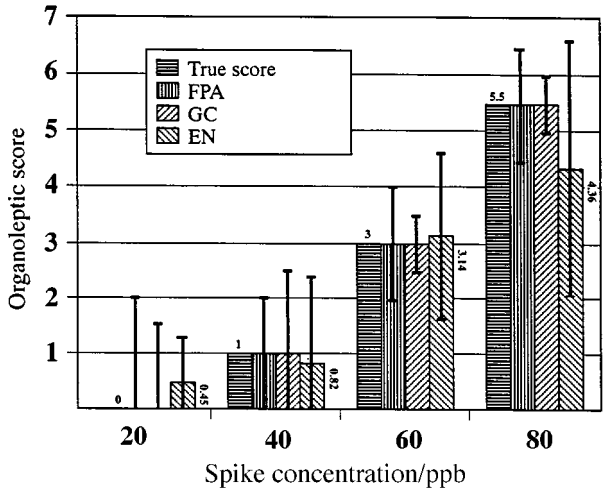

Fig. 14 A comparison of sensory-based FPA, instrumental techniques (GC-based methods) with the electronic nose for the prediction of DMS spike (at 20–80 ppb) in a background of lager beer. The mean predicted organoleptic scores generated by an electronic nose (EN) compared favourably with the theoretical score (arrived at from the known concentrations of DMS added to the samples), and the GC and FPA methods. Error bars indicate the 95% confidence intervals for each method. (GC and FPA data courtesy of Bass Brewers).

conducted by a trained tasting panel and GC-based instrumental methods is indicated by plotting the 95% confidence intervals for the data against typical known accuracies for the other analysis methods in Fig. 14. For higher concentrations (> 40 ppb) the accuracy of the electronic nose is shown to be comparable, but lower than both the GC-based and FPA methods. However, for lower concentrations, expected accuracies across all three methods are broadly similar, with results indicating that organoleptic flavour note prediction using the electronic nose may be more accurate than GC-MS analysis. This is an encouraging result, when regarding the cost of the competing flavour analysis methods and the potential for the low-cost commercial electronic nose exploiting mass produced chemical sensors. However, the future economies of scale for electronic noses will very much depend upon the future maturation and uptake of the technology in routine odour and flavour monitoring applications.

Conclusions

A 21-element conducting polymer electronic nose system has been developed to analyse the headspace of beers, described in detail in Part 1, ref. 1. In terms of the task of discriminating between the control lager beer and the same control with three different flavour spikes added, a certain amount of sensor drift was observed over a 12 d testing period. This drift has been shown to account for significant within-class variance in the data-set, the effect of which was reduced by applying a linear drift model. Careful use of this drift compensation model, coupled with judicious selection of pre-processing and pattern recognition techniques, maximised the between-class variance and so improved the overall classification performance of the system. In a related experiment, a novel data-processing method for sensor-arrays was applied to the same data-set to demonstrate its efficacy in predicting organoleptic flavour note scores. Consequently, we have shown that it is possible to predict an organoleptic score for a sub-ppb beer note present within a complex odour background of lager beer comprising more than 600 flavour active compounds. It may be possible to improve the accuracy of this prediction through the use of a fuzzy-based mapping algorithm, rather than the crisp ones employed here. A fuzzy approach may more accurately describe the nature of subjective organoleptic scores taken as an aggregate from a panel of human tasters.

Acknowledgements

The authors thank various colleagues on the LINK programme: Ms S. Friel for assistance in chemical hardware development, the late Mr C. Bidmead for technical assistance in the development of the FIA rig, Mr R. Cope (Bass Breweries Ltd., UK) for kindly providing the lager beer samples, Dr P. Heggerty (Bass Breweries Ltd., UK) for releasing information on the sensory-panel scores for the same samples, Professor P. N. Bartlett and in particular Dr N. Blair at the University of Southampton for their work on sensor fabrication. Finally, we are grateful to our industrial collaborators (Bass Breweries Ltd. and Neotronics Ltd., UK) and Government establishments (Department of Trade and Industry and Ministry of Agriculture, Fisheries and Food) for their financial support of this project.

We are also indebted to the helpful comments and suggestions of the referees and the following individuals in reading and reviewing this manuscript; Dr J. E. Dixon (University of Warwick, UK), Dr J. L. White, and Dr J. S. Kauer (Tufts University Medical School, USA).

References

1 T. C. Pearce and J. W. Gardner, *Analyst*, 1998, **123**, 2047.
2 ASBC Technical Committee and Editorial Committee, *Methods of Analysis of the American Society of Brewing Chemists*, ASBC, St. Paul, USA, 1976.
3 H. F. Kaiser, *EPM*, 1960, **20**, 141.
4 M. S. Bartlett, *British Journal of Psychology, Statistical Section*, 1950, **3**, 77.
5 W. J. Krzanowski, *Biometrics*, 1987, **43**, 575.
6 C. DiNatale, F. A. M. Davide and A. D'Amico, *Sens. Actuators B*, 1995, **27**, 237.
7 M. Holmberg, F. Winquist, I. Lundström, F. Davide, C. DiNatale and A. D'Amico, *Sens. Actuators*, 1996, **36**, 528.
8 V. Barnett, *Journal of the Royal Statistical Society, Series A*, 1976, **139**, 318.
9 S. S. Stevens, *Psychophysics*, McGraw-Hill, New York, 1975.

Paper 8/04019B