SENSORY ANALYSIS OF TAINTED FISH: ASSESSING THE ASSESSORS USING MIXED MODELS

A Craig, A F Zuur, R J Fryer and I M Davies

FRS Marine Laboratory
375 Victoria Road
Aberdeen AB11 9DB UK

1. INTRODUCTION

FRS Marine Laboratory, Aberdeen, is the designated facility for sensory assessment of fish and shellfish within the UK national contingency plan for response to offshore incidents such as hydrocarbon spills or other chemical releases. The facility has been used in the management of incidents such as the Sea Empress and Braer accidents. Sensory assessment for hydrocarbon taints in fish is accredited at FRS by the United Kingdom Accreditation Service, under international standard ISO 17025.

Assessment of fish and shellfish for petrogenic taint is performed by a panel of assessors, chosen and trained for their ability to recognise hydrocarbon taint in fish. Volunteers are initially screened for their general suitability as panel members (i.e. ability to detect and distinguish between fundamental tastes such as sweetness, bitterness, sourness and saltiness) and are then specifically trained in the recognition and scoring of hydrocarbon-related taint. The panel members participate in a continuous training programme, in which an assessor-specific reporting system is used at bi-monthly intervals to monitor panel performance. Individual performances are assessed and training records kept to support and validate the panel members’ expertise.

This report concerns the results of training sessions held between March 2002 and November 2003. The questions behind the study are:

1. Is each assessor consistent across sessions?
2. Do the assessors give consistent scores within sessions? Is any assessor an ‘outlier’?
3. How well can the unknown PAH concentration of a field sample (as opposed to a training sample) be predicted from the panel scores?

Sensory analysis data are typically analysed using multivariate or ANOVA related techniques. In this report, mixed modelling, an extension of ANOVA, is used.

2. LABORATORY PROCEDURES

2.1 Training of members of the sensory assessment panel

Assessors are trained using artificially tainted rainbow trout. Live rainbow trout, each weighing about 400 g, are obtained from a local fish farm and transferred to a temperature-controlled flow-through aquarium system in the Fish Behaviour Unit at FRS Marine Laboratory. The fish are placed in a holding tank at ambient water temperature (exact temperature dependent on seasonal conditions) and acclimatised over 3 days. During this time, the temperature of the water in the holding tank is gradually increased to 15°C (±1°C).

The number of fish held in the tanks depends on the number of replicate fish and the number of exposure times required by the design of the training programme.
100 litres (±10) of water are transferred from the holding tank into two or more aquaria insulated with expanded polystyrene. The aquaria are supplied with air and at most 8 fish are transferred to each tank. After 30 minutes to settle the fish, 2.5 ml (±0.1) of Forties crude oil are added to the surface of each tank and timing is begun. The required numbers of replicate fish are removed from the tanks at set exposure times and immediately killed, gutted and filleted. Fillets are stored at -20°C until required for sensory assessment.

Preparation of fish samples, and the subsequent assessment, take place in a purpose built laboratory at FRS Marine Laboratory where distractions are minimised and conditions controlled with respect to noise, lighting, and temperature (BSI, 1986; ISO, 1988). The preparation area is separate from the testing area.

75 g (±25) of skinned fillet taken from between the pectoral and ventral fins of each fish, or the whole fillet where fish are smaller, are placed in a casserole dish identified by a 3-digit random number. The fish, with no added condiments, is cooked in a microwave oven to a core temperature of a minimum 65°C. Any skin and bones remaining are removed and the fish tissue is flaked and mixed with a fork to obtain a homogeneous sample. The coded (numbered) casseroles containing the cooked samples are set out on hotplates.

The assessors are provided with fresh drinking water, plain water biscuits, plastic cutlery and plastic cups for expectoration. Each assessor is also given a score sheet (intensity scale record sheet) to record individual assessments of samples. Assessors initially clean their palates by rinsing with fresh drinking water. The first sample is smelled by taking an initial deep inhalation, followed by one or two shallower sniffs. The sample is tasted and expectorated. The intensity of any taint present is recorded on the score sheet provided and the assessor proceeds to the next sample. If taint persists in the mouth after rinsing, the palate is cleansed with a plain water biscuit and rinsed again before proceeding to the next sample.

Sensory assessment of the samples is recorded by assessors using a 6-point structured intensity scale (0-5), with scores of half points accepted within this intensity scale (Table 1).

### TABLE 1

<table>
<thead>
<tr>
<th>Score</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>absent</td>
</tr>
<tr>
<td>1</td>
<td>slight</td>
</tr>
<tr>
<td>2</td>
<td>moderate</td>
</tr>
<tr>
<td>3</td>
<td>strong</td>
</tr>
<tr>
<td>4</td>
<td>very strong</td>
</tr>
<tr>
<td>5</td>
<td>extremely strong</td>
</tr>
</tbody>
</table>

#### 2.2 Polycyclic aromatic hydrocarbon analysis

Hydrocarbon taint in fish arises primarily from the presence of low molecular weight polycyclic aromatic hydrocarbons (PAH), in particular 2-ring compounds such as naphthalene and substituted naphthalenes. Parallel samples of fish tissue are therefore analysed chemically for PAH and the results are subsequently compared with the taste panel results. Samples of muscle from the same fish used for the sensory assessment are used in the analysis. PAH concentrations are normally measured before the sensory assessment training sessions, to assist in the selection of samples for use in training exercises.
The isolation and estimation of hydrocarbons from biota is based on the method described by Webster et al. (1997). A sample of fish muscle is accurately weighed into a 250 ml round bottom flask. To this is added 100 µl (±1 µl) deuterated aromatic standard which includes D₈-naphthalene, D₁₀-biphenyl, D₈-dibenzothiophene, D₁₀-anthracene, D₁₀-pyrene and D₁₂-benzo[a]pyrene. 40 ml (±4 ml) of sodium hydroxide (10%) are added together with a few anti-bumping granules, and a reflux condenser fitted to the flask. Flasks are lowered onto a sandbath and the mixture is refluxed for 3h 45 ± 5 min before the addition of water (10 ± 1 ml) via the top of the condenser and the refluxing continued for a further 15 ± 2 min.

On completion of this saponification, the hot solution is transferred to a 250 ml separating funnel containing iso-hexane (80 ± 5 ml). Methanol/water (4:1 (v/v), 40 ± 4 ml) is used to rinse the flask before being added to the separating funnel. The mixture is thoroughly shaken and then allowed to settle. The lower aqueous layer is transferred to a second separating funnel containing iso-hexane (80 ± 5 ml) and the solutions thoroughly mixed. While the second extraction is settling, the first iso-hexane extract is washed with methanol/water (1:1 (v/v), 40 ± 4 ml). The mixture is shaken. The aqueous layer from the second iso-hexane extraction is run off to waste and the methanol/water from the first extract mixed with the second iso-hexane extract. After thorough shaking, the two layers are allowed to separate and the aqueous layer is then run to waste.

The iso-hexane extracts are combined and washed with water (3 × 40 ± 4 ml). A sodium sulphate column is prepared by adding washed sodium sulphate to a glass column with a sinter. The combined extracts are run through the column and collected in a 250 ml round bottom flask followed by 50 ± 5 ml of iso-hexane. The solvent is concentrated by rotary evaporation to approximately 300 µl and transferred to a screw top vial. The flask is washed with iso-hexane and the washings transferred to the vial to give a total volume of 500 µl.

HPLC fractionation is performed on a 150 µl (± 10 µl) aliquot using a Genesis metal-free HPLC column. Elution is with iso-hexane at a flow rate of 2 ± 0.1 ml min⁻¹. The aromatic fraction is collected in a 50 ml round bottom flask and the solution concentrated using a rotary evaporator, followed by further reduction to 50 µl (± 10 µl) under a stream of scrubbed nitrogen.

The concentration and composition of the PAH compounds are determined by GC-MSD using a gas chromatograph interfaced with a mass selective detector.

3. MEASURING PANEL PERFORMANCE

The data from the nine training sessions between March 2002 and November 2003 are shown in Figure 1. There were 11 assessors in total, with at least 5 present in any one session. The scores given by an assessor in any one session are approximately linearly related to the square root transformed PAH concentration. The consistency of the panel can therefore be measured by the consistency on these linear score-PAH relationships. We now consider such consistency, first of individual assessors across sessions, then of all assessors within single sessions, and finally of the whole panel.

3.1 Consistency of individual assessors across sessions

A good assessor should give consistent scores in all sessions and be able to discriminate between high and low PAH concentrations. Such assessors should give a consistent linear relationship between score and (square root transformed) PAH concentration across sessions with a steep slope for good discrimination. We can thus measure an assessor's
performance by focussing on the vertical direction in Figure 1, which links scores for a single assessor across sessions, and comparing the regression lines. To do so, we fit a mixed model:

Model 1: \[ Y_{ik} = a + b \text{PAH}_{ik} + A_k + B_k \text{PAH}_{ik} + e_{ik} \]

where \( Y_{ik} \) is the assessor’s score for sample \( i \) in session \( k \) and \( \text{PAH}_{ik} \) is the corresponding (square root transformed) PAH concentration. The parameters \( a \) and \( b \) are fixed effects, representing the mean intercept and slope across sessions. They define the mean score-PAH relationship for the assessor. The terms \( A_k, B_k \) are random effects, allowing the intercept and slope of the score-PAH relationship to vary between sessions. They are assumed to be normally distributed with mean 0 and variances \( \sigma^2_A \) and \( \sigma^2_B \) respectively. The term \( e_{ik} \) is the residual variation associated with each score and is assumed to be normally distributed with mean 0 and variance \( \sigma^2_e \). A good assessor would have a steep slope \( b \) for good discrimination, small random effects variances \( \sigma^2_A \) and \( \sigma^2_B \) for a consistent score-PAH relationship across sessions, and a small residual variance \( \sigma^2_e \) so that the score-PAH relationship is well defined within each session.

It is often convenient to formulate mixed models in a less cumbersome notation. Here the mixed model can be written more simply as:

\[
\begin{align*}
\text{fixed} & \sim \text{constant} + \text{PAH} \\
\text{random} & \sim \text{Session} + \text{Session.PAH}
\end{align*}
\]

The fixed component shows that the score is linearly related to PAH, whilst the random component shows that both the intercept (Session) and the slope (Session.PAH) of the linear relationship vary between sessions.

### 3.2 Consistency of assessors within individual sessions

A different way of measuring panel performance is to focus on the horizontal direction in Figure 1 and consider whether all assessors give similar scores in the same session. As before, we can compare the regression lines for all the assessors within a session by fitting a mixed model.

Model 2: \[ Y_{ij} = a + b \text{PAH}_{ij} + C_j + D_j \text{PAH}_{ij} + e_{ij} \]

\( Y_{ij} \) is the score of sample \( i \) by assessor \( j \). The terms \( C_j, D_j \) are random effects allowing the intercept and slope of the score-PAH relationship to vary between assessors and assumed to be normally distributed with mean 0 and variances \( \sigma^2_C \) and \( \sigma^2_D \) respectively. The model can also be written as:

\[
\begin{align*}
\text{fixed} & \sim \text{constant} + \text{PAH} \\
\text{random} & \sim \text{Assessor} + \text{Assessor.PAH}
\end{align*}
\]

where the terms Assessor and Assessor.PAH indicate that the intercept and slope of the score-PAH relationship vary between assessors.

Again, a good panel would have a steep slope \( b \) for discrimination, small variances \( \sigma^2_C \) and \( \sigma^2_D \) for consistency across assessors, and a small residual variance \( \sigma^2_e \) so that the score-PAH relationship is well defined for each assessor. However, we might expect the between-assessor variances \( \sigma^2_C \) and \( \sigma^2_D \) to be larger than the between-session variances \( \sigma^2_A \) and \( \sigma^2_B \) considered earlier, since assessors are likely to have their own range of scores and taste threshold levels.
3.3 Consistency of the panel

To assess the consistency of the panel, we need to assess consistency across both assessors and sessions simultaneously. The two models considered earlier can be combined into a single mixed model of all the panel data:

Model 3: \[ Y_{ijk} = a + b \text{PAH}_{ijk} + A_k + B_k \text{PAH}_{ijk} + C_j + D_j \text{PAH}_{ijk} + E_{jk} + F_{jk} \text{PAH}_{ijk} + e_{ijk} \]

\( Y_{ijk} \) is the score of sample \( i \) of assessor \( j \) in session \( k \). As before, the random effects \( A_k \) and \( B_k \) allow for variation in the score-PAH relationship between assessors and the random effects \( C_j \) and \( D_j \) allow for variation in the score-PAH relationship between sessions. The terms \( E_{jk} \) and \( F_{jk} \) are additional random effects that allow the intercept and slope of the score-PAH relationship to vary between assessors within sessions (i.e. in a different way for each session) and are assumed to be normally distributed with mean 0 and variances \( \sigma^2 \) and \( \sigma^2_{E} \) and \( \sigma^2_{F} \) respectively. The mixed model can also be written as:

\[
\text{fixed} \sim \text{constant} + \text{PAH} \\
\text{random} \sim \text{Session} + \text{Session.PAH} + \text{Assessor} + \text{Assessor.PAH} + \text{Session.Assessor} + \text{Session.Assessor.PAH} \\
\]

where Session.Assessor and Session.Assessor.PAH denote random variation in the score-PAH relationship between assessors within sessions.

4. RESULTS

To reduce correlation between the parameter estimates, the square root transformed PAH concentrations were centred on zero before fitting the mixed models. The intercepts in the mixed models therefore correspond to the score when the PAH concentration is about 900 ng g\(^{-1}\).

4.1 Consistency of individual assessors across sessions

Model 1 was fitted to the data for each assessor in turn, apart from assessor 1 who took part in only one session. The results are summarised in Table 2 and Figure 2. The estimated mean intercepts and slopes are similar for most of the assessors, although assessors 3, 6, and 7 tend to give lower scores and have lower slopes than the others, so might have difficulty discriminating between low and high PAH concentrations (Figure 2). Assessor 10 is the least consistent, having large between-session variation in both slope and intercept (Figure 2). This is partly due to poor performances in sessions 2 and 4 (Figure 1).
TABLE 2

Estimates of the mean intercept (a) and slope (b), with standard errors in brackets, and of the Session ($\sigma_A^2$), Session.PAH ($\sigma_B^2$) and residual ($\sigma_e^2$) variance components for each assessor.

<table>
<thead>
<tr>
<th>Assessor</th>
<th>a</th>
<th>b</th>
<th>$\sigma_A^2$</th>
<th>$\sigma_B^2$</th>
<th>$\sigma_e^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>2.00</td>
<td>0.071</td>
<td>0.45</td>
<td>0.012</td>
<td>0.86</td>
</tr>
<tr>
<td>3</td>
<td>1.12</td>
<td>0.038</td>
<td>0.00</td>
<td>0.000</td>
<td>0.97</td>
</tr>
<tr>
<td>4</td>
<td>1.79</td>
<td>0.031</td>
<td>0.41</td>
<td>0.016</td>
<td>0.82</td>
</tr>
<tr>
<td>5</td>
<td>2.10</td>
<td>0.086</td>
<td>0.54</td>
<td>0.014</td>
<td>0.91</td>
</tr>
<tr>
<td>6</td>
<td>1.31</td>
<td>0.051</td>
<td>0.00</td>
<td>0.000</td>
<td>1.12</td>
</tr>
<tr>
<td>7</td>
<td>0.68</td>
<td>0.027</td>
<td>0.51</td>
<td>0.018</td>
<td>0.52</td>
</tr>
<tr>
<td>8</td>
<td>2.11</td>
<td>0.077</td>
<td>0.19</td>
<td>0.015</td>
<td>1.10</td>
</tr>
<tr>
<td>9</td>
<td>1.83</td>
<td>0.064</td>
<td>0.29</td>
<td>0.011</td>
<td>1.34</td>
</tr>
<tr>
<td>10</td>
<td>2.94</td>
<td>0.070</td>
<td>0.89</td>
<td>0.032</td>
<td>1.30</td>
</tr>
<tr>
<td>11</td>
<td>2.35</td>
<td>0.096</td>
<td>0.64</td>
<td>0.024</td>
<td>1.01</td>
</tr>
</tbody>
</table>

4.2 Consistency of assessors within individual sessions

The results of fitting model 2 to each session in turn are summarised in Table 3 and Figure 3. The estimated mean intercepts and slopes are similar for all sessions (Figure 3), indicating that the panel, as a whole, perceived PAH concentrations in a reasonably consistent way. However, within sessions, there was considerable variation in both intercept and slope between assessors (Figure 3). For example, session 4 had large between-assessor variation in intercept, which can in part be attributed to the results for assessor 10 (Figure 1). Similarly, session 5 had large between-assessor variation in slope, which can in part be attributed to assessor 7 (Figure 1).
TABLE 3

Estimates of the mean intercept (a) and slope (b), with standard errors in brackets, and of the Assessor (σ_C^2), Assessor.PAH (σ_D^2) and residual (σ_e^2) variance components for each session.

<table>
<thead>
<tr>
<th>Session</th>
<th>a</th>
<th>b</th>
<th>σ_C^2</th>
<th>σ_D^2</th>
<th>σ_e^2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.25</td>
<td>0.06</td>
<td>0.53^2</td>
<td>0.020^2</td>
<td>1.09^2</td>
</tr>
<tr>
<td>2</td>
<td>2.06</td>
<td>0.06</td>
<td>0.88^2</td>
<td>0.020^2</td>
<td>0.99^2</td>
</tr>
<tr>
<td>3</td>
<td>1.42</td>
<td>0.06</td>
<td>0.31^2</td>
<td>0.012^2</td>
<td>0.82^2</td>
</tr>
<tr>
<td>4</td>
<td>1.98</td>
<td>0.04</td>
<td>0.97^2</td>
<td>0.013^2</td>
<td>1.19^2</td>
</tr>
<tr>
<td>5</td>
<td>1.57</td>
<td>0.08</td>
<td>0.56^2</td>
<td>0.029^2</td>
<td>1.14^2</td>
</tr>
<tr>
<td>6</td>
<td>1.89</td>
<td>0.10</td>
<td>0.78^2</td>
<td>0.032^2</td>
<td>1.14^2</td>
</tr>
<tr>
<td>7</td>
<td>1.72</td>
<td>0.06</td>
<td>0.90^2</td>
<td>0.031^2</td>
<td>0.76^2</td>
</tr>
<tr>
<td>8</td>
<td>1.45</td>
<td>0.06</td>
<td>0.71^2</td>
<td>0.031^2</td>
<td>1.02^2</td>
</tr>
<tr>
<td>9</td>
<td>1.45</td>
<td>0.05</td>
<td>0.57^2</td>
<td>0.030^2</td>
<td>0.86^2</td>
</tr>
</tbody>
</table>

4.3 Consistency of the panel

Fitting model 3 to the whole data set showed that the Session.Assessor.PAH variance component was non-significant. The model was therefore refitted having omitted this term and the results are given in Table 4. Variation between assessors is greater than between sessions. However, the largest variance is the residual term, which measures variation about the lines shown in Figure 1. The effect of these variance components on the performance of the panel is considered next.

TABLE 4

Estimates of the mean intercept and slope (with standard errors in brackets) and of the variance components obtained by fitting model 3 to the whole data set. The correlations between the Session and Session.PAH and between the Assessor and Assessor.PAH random effects were estimated to be 0.50 and 0.86 respectively.

<table>
<thead>
<tr>
<th>term</th>
<th>notation</th>
<th>estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean intercept</td>
<td>a</td>
<td>1.81 (0.47)</td>
</tr>
<tr>
<td>Session</td>
<td>σ_A^2</td>
<td>0.24^2</td>
</tr>
<tr>
<td>Assessor</td>
<td>σ_C^2</td>
<td>0.63^2</td>
</tr>
<tr>
<td>Session.Assessor</td>
<td>σ_E^2</td>
<td>0.31^2</td>
</tr>
<tr>
<td>mean slope</td>
<td>b</td>
<td>0.064 (0.027)</td>
</tr>
<tr>
<td>Session.PAH</td>
<td>σ_B^2</td>
<td>0.009^2</td>
</tr>
<tr>
<td>Assessor.PAH</td>
<td>σ_D^2</td>
<td>0.019^2</td>
</tr>
<tr>
<td>residual</td>
<td>σ_e^2</td>
<td>1.06^2</td>
</tr>
</tbody>
</table>

5. CALIBRATION OF PANEL SCORES

This section investigates how well the unknown PAH concentration of a field sample can be estimated from the mean score of J assessors. To do so, it is first necessary to estimate the variance of the mean score as a function of the unknown concentration. For ease of notation, denote the session by k, and the square root concentration by PAH. Then, if Y_jk is the score of assessor j, we have from model 3:

\[ Y_{jk} = a + b \text{PAH} + A_k + B_k \text{PAH} + C_j + D_j \text{PAH} + E_{jk} + e_{jk} \]
SENSORY ANALYSIS OF TAINTED FISH

(The Session.Assessor.PAH random effect has been dropped for simplicity, since it was non-significant.) The mean score is then:

$$\bar{Y} = a + bPAH + A_i + B_jPAH + \frac{C}{J} + \frac{D}{J}PAH + \frac{E_{ij}}{J} + \frac{e_{ij}}{J}$$

where the dot stands for summation over assessors. If we assume that a and b are known, the variance of the mean score is:

$$\text{var}(\bar{Y}) = \sigma_A^2 + 2PAH\sigma_{AB}^2 + PAH^2\sigma_B^2 + \frac{1}{J}(\sigma_C^2 + 2PAH\sigma_{CD}^2 + PAH^2\sigma_D^2) + \frac{\sigma_e^2}{J} + \frac{\sigma_{ij}^2}{J}$$

where $\sigma_{AB}^2$ and $\sigma_{CD}^2$ are the covariances between the Session and Session.PAH and between the Assessor and Assessor.PAH random effects respectively. The variance of the mean score increases as the unknown PAH concentration increases, but can be reduced by increasing the number of assessors (Figure 4). The variance will always be at least

$$\sigma_A^2 + 2PAH\sigma_{AB}^2 + PAH^2\sigma_B^2$$

unless there is more than one session.

Given the mean score, an approximate 95% confidence interval for the unknown concentration is the set of values that satisfy

$$\bar{Y} \in [a + bPAH \pm 2\sqrt{\text{var}(\bar{Y})}]$$

For example, Figure 5 shows that if there are 15 assessors, and the mean score of the panel is about 2, the unknown PAH concentration is between about 400 and 2000 ng g$^{-1}$ with 95% confidence. Similarly, if the mean score is about 3, the PAH concentration is larger than about 1100 ng g$^{-1}$ with 95% confidence, although no upper limit can be specified. And if the mean score is about 1.5, the PAH concentration is between about 200 and 1200 ng g$^{-1}$ with 95% confidence. Figure 6 shows how the widths of the confidence intervals increase when the number of assessors is reduced to 10.

To understand which variance components cause the wide confidence bands, we calculated the percentage contribution to $\text{var}(\bar{Y})$ of the following four terms:

- **Session + Session.PAH**: $\sigma_A^2 + 2PAH\sigma_{AB}^2 + PAH^2\sigma_B^2$
- **Assessor + Assessor.PAH**: $(\sigma_C^2 + 2PAH\sigma_{CD}^2 + PAH^2\sigma_D^2)/J$
- **Session. Assessor**: $\sigma_e^2 / J$
- **Residual**: $\sigma_{ij}^2 / J$

(Figure 7). For a panel of 10 assessors, the between-session variances are between 25 and 40% of the total variance. The between-assessor variances are negligible at low concentrations, but are up to 35% of the total variance at large concentrations. The Session.Assessor variance is negligible for all concentrations. The residual variance is about 60% of the total variance at small concentrations and 30% at large concentrations. This might indicate that the model gives a poor fit at low PAH concentrations. In particular, the variance structure might not be well modelled here.

Throughout this section we have assumed that the mean intercept and slope are known without error. In practice, the variability in these estimates should also be accounted for when predicting concentrations from mean scores. This will only increase the widths of the confidence intervals.
6. DISCUSSION

In this report, three sets of mixed models have been fitted to data from sensory assessment panel training sessions using fish tainted with hydrocarbons. The first approach considered the consistency of individual assessors across sessions, and identified one assessor who was particularly inconsistent. The second approach revealed a lot of variation in the scores given by different assessors within the same session. The third approach estimated variance components that could be used to assess the performance of the panel as a whole, and in particular, to investigate the precision with which the unknown concentration of a field sample could be estimated from the mean score of the panel. The results suggest that the confidence intervals on the unknown concentration would be very wide.

The between-session, between-assessor and residual variance components all make substantial contributions to the wide confidence intervals. There are several ways that the variation might be reduced. Giving assessors replicate samples would reduce the contribution of the residual component. The between-session variation could be ignored if the session-specific intercept and slope for the specific session were used to construct the calibration line, rather than the mean intercept and slope from multiple sessions. However, there might be no benefit in doing this if the session-specific intercept and slope are poorly estimated. Reducing the assessor effect would be particularly useful. One option might be to remove assessors who do not use the entire range of scores, as these assessors contribute to the large variation in mean score for samples with higher PAH concentrations. However, this option is unattractive because there are few trained assessors. Alternatively, the training regime could be revised, with greater corrective feedback on what constitutes strong and extremely strong taint. Another option would be to change the scoring system to a 0-1 scale denoting absence-presence of taint. This would remove the need to discriminate between samples with differing degrees of taint. Such data could be analysed using generalised mixed models for binomial data, a generalisation of the methods applied in this report. The current data could be converted and used to test such an approach.

7. REFERENCES


List of Figures and Tables

Figure 1. Scores plotted against square root transformed PAH concentrations (ng g\(^{-1}\)), with a linear least squares fit, for each assessor (A1, ..., A11) and session (S1, ..., S9). The concentration axis is on the transformed scale, but the original concentrations are used as labels.

Figure 2. The consistency of individual assessors within sessions. The horizontal axes show the estimated mean intercept \(a\) (top) and mean slope \(b\) (bottom) for each assessor with 95% confidence limits. The plotting symbols denote the assessor. The vertical axes shows the corresponding between-session standard deviations in intercept \(\sigma_A\) and slope \(\sigma_B\).

Figure 3. The consistency of assessors within individual sessions. The horizontal axes show the estimated mean intercept \(a\) (top) and mean slope \(b\) (bottom) for each session with 95% confidence limits. The plotting symbols denote the session. The vertical axes shows the corresponding between-assessor standard deviations in intercept \(\sigma_C\) and slope \(\sigma_D\).

Figure 4. Contour plot showing how the variance of the mean score changes with the number of assessors and the PAH concentration.

Figure 5. Calibration line for 15 assessors, showing the expected relationship between mean score and PAH concentration \(a + b\text{PAH}\), with pointwise 95% prediction bands \(a + b\text{PAH} \pm 2\sqrt{\text{var} \bar{Y}}\). The dashed lines are used to calculate an approximate 95% confidence interval on an unknown PAH concentration when the mean score is about 2 (top), 3 (middle), and 1.5 (bottom).

Figure 6. Calibration line for 10 assessors, showing approximate 95% confidence intervals on the unknown PAH concentrations when the mean score is about 2 (top), 3 (middle), and 1.5 (bottom).

Figure 7. The percentage contribution of different variance components to the total variance of the mean score, as a function of PAH and number of assessors.

Table 1. Six point sensory assessment scale

Table 2. Estimates of the mean intercept (a) and slope (b), with standard errors in brackets, and of the Session (\(\sigma_A^2\)), Session.PAH (\(\sigma_B^2\)) and residual (\(\sigma_e^2\)) variance components for each assessor.

Table 3. Estimates of the mean intercept (a) and slope (b), with standard errors in brackets, and of the Assessor (\(\sigma_C^2\)), Assessor.PAH (\(\sigma_D^2\)) and residual (\(\sigma_e^2\)) variance components for each session.

Table 4. Estimates of the mean intercept and slope (with standard errors in brackets) and of the variance components obtained by fitting model 3 to the whole data set. Correlations between the Session and Session.PAH and between the Assessor and Assessor.PAH random effects were estimated to be 0.50 and 0.86 respectively.
Figure 1: Scores plotted against square root transformed PAH concentrations (ng g$^{-1}$), with a linear least squares fit, for each assessor (A1, ..., A11) and session (S1, ..., S9). The concentration axis is on the transformed scale, but the original concentrations are used as labels.
Figure 2: The consistency of individual assessors within sessions. The horizontal axes show the estimated mean intercept $a$ (top) and mean slope $b$ (bottom) for each assessor with 95% confidence limits. The plotting symbols denote the assessor. The vertical axes show the corresponding between-session standard deviations in intercept $\sigma_A$ and slope $\sigma_B$. 
Figure 3: The consistency of assessors within individual sessions. The horizontal axes show the estimated mean intercept $a$ (top) and mean slope $b$ (bottom) for each session with 95% confidence limits. The plotting symbols denote the session. The vertical axes shows the corresponding between-assessor standard deviations in intercept $\sigma_C$ and slope $\sigma_D$. 
**Figure 4:** Contour plot showing how the variance of the mean score changes with the number of assessors and the PAH concentration.
Figure 5: Calibration line for 15 assessors, showing the expected relationship between mean score and PAH concentration $a + bPAH$, with pointwise 95% prediction bands $a + bPAH \pm 2\sqrt{\text{var} \bar{Y}}$. The dashed lines are used to calculate an approximate 95% confidence interval on an unknown PAH concentration when the mean score is about 2 (top), 3 (middle), and 1.5 (bottom).
Figure 6: Calibration line for 10 assessors, showing approximate 95% confidence intervals on the unknown PAH concentrations when the mean score is about 2 (top), 3 (middle), and 1.5 (bottom).
Figure 7: The percentage contribution of different variance components to the total variance of the mean score, as a function of PAH and number of assessors.