Assessing the Performance of a Sensory Panel - Panelist monitoring and tracking

Martin Kermit and Valérie Lengard

CAMO Process AS
Nedre Vollgate 8, 0158 Oslo, Norway
www.camo.com

July 3, 2006

Abstract

Sensory science uses the human senses as instruments of measures. This study presents univariate and multivariate data analysis methods to assess individual and group performances in a sensory panel. Green peas were evaluated by a trained panel of 10 assessors for six attributes over two replicates. A consonance analysis with Principal Component Analysis (PCA) is run to get an overview of the panel agreement and detect major individual errors. The origin of the panelist errors is identified by a series of tests based on ANOVA: sensitivity, reproducibility, crossover and panel agreement, complemented with an eggshell-correlation test. One assessor is identified with further need for training in attributes pea flavour, sweetness, fruity and off-flavour, showing errors in sensitivity, reproducibility and crossover. Another assessor shows poor performance for attribute mealiness and to some extent also fruity flavour. Only one panelist performs well to very well in all attributes. The specificity and complementarity of the series of univariate tests are explored and verified with the use of a PCA model.

Keywords:
Sensory panel performance; ANOVA; Agreement error; Sensitivity; Reproducibility; Crossover; Eggshell plot

1 Introduction

The performance level of a descriptive panel of assessors and the quality of the data they provide is of paramount importance for making proper research and business decisions. A good sensory panel should provide results that are accurate, discriminating and precise. Thus, in a successful analysis, it is key to have a set of robust tools for monitoring individual assessor performances as well as the panel as a whole.

Due to its versatility, analysis of variance (ANOVA) has been one of the most often employed statistical tools to study differences between products [1] [2]. This standard univariate method is also used to separate the total variation of sensory data into sources that affect sensory responses [3]. For multivariate analysis across different attributes, principal component analysis (PCA) is the natural choice for consonance analysis [4] and when averaged over assessors [5]. Both these complementary methods are necessary to achieve a representative picture of the performance of a sensory study.

This contribution presents a set of tests for evaluating the performance of individual assessors as well as the total panel for individual attributes. The univariate tool collection is based on sequential ANOVA tests to perform tests on sensitivity, assessor reproducibility, panel agreement and crossover effects. Rank correlations using eggshell plots are also considered. These tests are illustrated using a data set consisting of sensory evaluations of green peas described in [6]. For multivariate tests to identify specificities and possible correlation between the univariate test results, PCA is used. Multivariate methods like three-way regression or Generalized Procrustes Analysis (GPA) are not addressed in this paper.

The next section provides some insight on the kind of errors that may be experienced when analyzing sensory panel data. The sources of these errors and how they affect assessor and panel performance are also described. Section 3 presents the univariate test collection and gives details on the mathematical foundation for the tests. Explanation of the data set and the software used follows in the section thereafter. Section 5 brings the result on the mentioned data set, and section 6 concludes this paper.

2 Performance errors in descriptive sensory evaluations

Within descriptive sensory analysis, it is a well known fact that assessors give uneven results stemming from differences in motivation, sensitivity and psychological response behaviors [7]. Despite the training sessions that each panel undergo, the reliability of the collected data may suffer both from individual assessor errors and from panel agreement errors. A first step in the analysis of sensory data is to identify individual assessors performing abnormally or inconsistently, and have their data for the actual attribute(s) reevaluated in the further analysis. The information about the strengths and weaknesses of each assessor is also key to organize adapted follow-up training sessions and improve the performance of the panel.

2.1 Errors at individual assessor level

Three important errors that can affect individual assessor performance are listed below:

- **Location error**: The assessor uses a different location of the scale than the rest of the panel.
- **Sensitivity error**: The assessor is not able to discriminate between two or more products.
- **Reproducibility error**: The assessor is not able to consistently replicate a judgement for one or more products.

Sensitivity errors are important to identify, so that the assessor can be notified or excluded from further testing for that particular type of product or attribute. Reproducibility errors are also crucial to detect, because extreme ratings in both directions of the scale might lead artificially to a mean rating comparable with the rest of the panel. Such large rating variations should lead to the conclusion that the assessor cannot be trusted and result in the exclusion of the assessor's data from further analysis for the faulty attribute.

2.2 Agreement errors within a sensory panel

A poor assessor performance will eventually lead to disagreement errors in product ratings with respect to the rest of the sensory panel. Four typical patterns of disagreement errors among assessors can be experienced:

- **Magnitude error**: The assessor uses a broader or smaller range of the scale than the rest of the panel.
- **Crossover error**: The assessor rates a product or set of products in the opposite direction from the rest of the panel.
- **Non-discriminator error**: The assessor rates all the products in a set as similar when the rest of the panel rated them as different.
- **Non-perceiver error**: The assessor does not perceive an attribute and scores all the products at "0" when the rest of the panel rated them as different.

Figure 1 illustrates an example of ratings for assessors exhibiting the different types of errors. Magnitude errors are seen for assessors that lack calibration training and scale their ratings on a too broad or too narrow interval, thus incorrectly. Crossover errors are of high importance as they often are the reason for a poor panel consistency. Also in this case, lack of training may be the cause for this type of error. Non-discriminator errors can sometimes be mistakenly interpreted as magnitude errors, since only a small interval on the scale of ratings is used by the assessor.

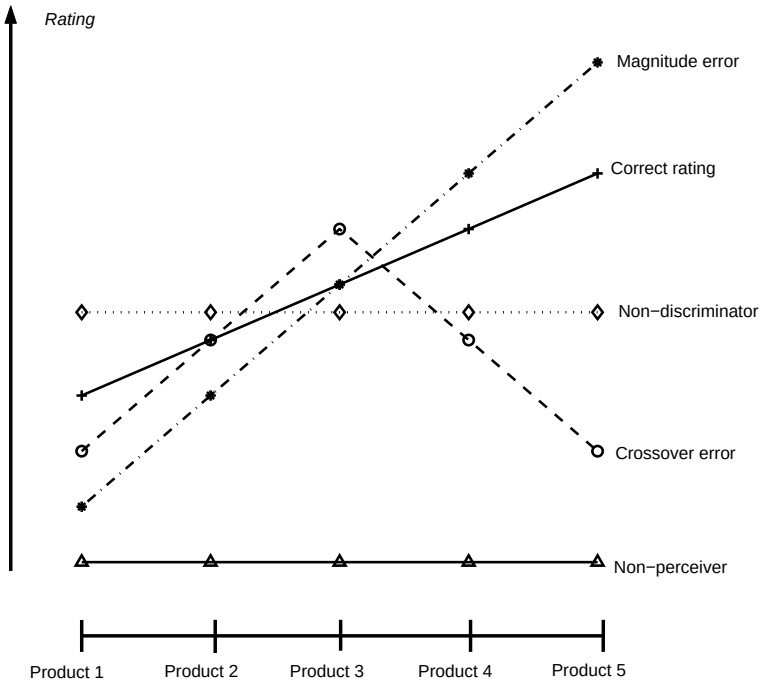

Figure 1: Four types of panel agreement errors.

Most likely, the poor discrimination is due to lack of sensitivity by the assessor. Such errors do not necessarily affect the outcome of the panel result significantly if the rest of panel performs satisfactory, and the number of assessors is not too small. Non-perceiver errors are not crucial to the panel, but still important, since leaving non-perceiver data in the analysis affect the mean score for that attribute. Non-perceiving assessors might be using the lexicon incorrectly or experiencing poor sensitivity.

3 Methods to identify assessor and panel performance

Prior to detailed investigation of errors, it may prove useful to run an exploratory analysis of the panel performance to have a first glance of the panel agreement. This will give a clue to the extent the panel can be trusted and further testing for errors for individual assessors can be continued.

Panel errors at the individual assessor level are easier to identify than for the whole panel, as will be addressed in this section. In the sensitivity test described, only data for the actual assessor is used to identify magnitude, non-perceivers or non-discriminator patterns. For the reproducibility, agreement and crossover tests, the score table for the whole panel must be used, and a more complex data model will be applied.

3.1 Consonance analysis with PCA

The purpose of consonance analysis is to study the level of agreement within the panel. Principal Component Analysis (PCA) is a powerful tool for this purpose and a similar method called consonance analysis has been described in [4]. For each attribute, a PCA is run on the individual

assessors' evaluations (the variable set) for the set of products (the sample set). In this model, the panel agreement is often interpreted along the first principal component, and the variance explained along this component is taken as the percentage of panel agreement for the attribute in consideration [8]. The remaining variance explained by higher principal components can, in our experience, be accounted for by a combination of effects like different use of scale and varying sensitivity.

Further, the plot of loadings shows the replicated evaluations for each assessor, thus allowing a visual detection of assessors outlying from the rest of the panel as well as an identification of individual reproducibility errors. This exploratory multivariate method gives an efficient overview of the panel performance. However, it does not allow to identify the nature of the assessors' errors.

3.2 Full ANOVA model and notations

Sensory data is typically available as a 4-way data structure $Y_{ijkm}, i = 1, ..., I, j = 1, ..., J, k = 1, ..., K$ and $m = 1, ..., M$. In this structure, the number of assessors I, products J, attributes K and replicates M are the four data modes. Using the notation borrowed from [9], this type of data can be described by an analysis of variance model including only two main effects and the interaction effect due to panelist by product [10]:

$$Y_{ijkm}^{\text{full}} = \mu_k + \alpha_{ik} + \beta_{jk} + (\alpha\beta)_{ijk} + \varepsilon_{ijkm}^{\text{full}}. \tag{1}$$

Here, μ_k is the grand mean for attribute k and α_{ik} the main effect contributed by assessor i for this attribute. The main effect from product j for the kth attribute is represented by β_{jk}. The interaction effect $(\alpha\beta)_{ijk}$ provides the differences between assessors in measuring differences between products. The error term $\varepsilon_{ijkm}^{\text{full}}$ represents the residual variation due to replicates, and the superscript is included to indicate a full ANOVA model for further use in this analysis. Since only one attribute will be considered at a time, the k index is omitted due to clarity in the following text.

3.3 Assessor sensitivity

For a given attribute, an assessor should ideally give equal rating to product repetitions, and different scores to different products. The sensitivity test measures the ability of a single assessor to identify product differences. This can be formulated as a one-way ANOVA test for a single assessor i. The ANOVA model becomes similar to equation (1), but without the effects due to assessor α_i and interaction $(\alpha\beta)_{ij}$,

$$Y_{jm} = \mu + \beta_j + \varepsilon_{jm}^{\text{sens}} \tag{2}$$

This ANOVA model is focused on modeling product differences, and is thus suitable for further F-testing to see whether the assessor is able to discriminate between them. Calculated p-values from the F-test will be low for assessors with good sensitivity.

3.4 Assessor reproducibility

A reproducibility test monitors the ability of a single assessor to reproduce judgements for M replicates of the same product. For proper assessment of reproducibility error, the reproducibility of each assessor should be compared to the reproducibility performances of the panel as a whole. Therefore, ANOVA statistics that model products for each assessor do not directly identify the panelists contributing significantly to a reproducibility effect of the panel. A solution is to partition the error from an ANOVA model in order to calculate how much each assessor contributes to

that error. By using the full two-way ANOVA model in equation (1), the residual $\varepsilon_{ijm}^{\text{full}}$ can be standardized by

$$\epsilon_{jm}^{\text{full}} = \varepsilon_{jm}^{\text{full}} / \sqrt{\text{MS}_{\text{error}}^{\text{full}}} = (Y_{jm} - \hat{Y}_{jm}) / \sqrt{\text{MS}_{\text{error}}^{\text{full}}} \tag{3}$$

where $\hat{Y}_{jm}$) is the fitted ANOVA model. This standardized residual can again be used to conduct a one-way ANOVA for each assessor

$$\epsilon_{jm}^{\text{full}} = \mu + \beta_j + \varepsilon_{jm}^{\text{repro}}. \tag{4}$$

The sum of squares for error,

$$\text{SS}_{\text{error}}^{\text{repro}} = \sum_{j=1}^{J} \sum_{m=1}^{M} (\varepsilon_{jm}^{\text{repro}})^2 = \sum_{j=1}^{J} \sum_{m=1}^{M} (\epsilon_{jm}^{\text{full}} - \hat{\epsilon}_{jm}^{\text{full}})^2, \tag{5}$$

is then a measure of the contribution from the assessor to the sum of squares for error for the whole panel. The individual assessors partitioned sums of squares for error follow a chi-square distribution in that $\text{SS}_{\text{error}}^{\text{repro}} \sim \chi^2_{J(M-1)}$, under the assumption that an assessor's responses are as reproducible as an average panelist. A χ^2-test is then performed to test whether the individual assessor sums of squares for error are equal to those of an average panelist against a two-tailed alternate hypothesis (better or worse than average performer). Thus, a high p-value from the χ^2 test indicates a good reproducibility of the assessor, whereas a low p-value corresponds to a poor reproducibility.

3.5 Agreement test

An agreement test should be able to test how much each assessor is contributing to the total agreement error. This can be achieved using the contrast method and the sums of squares reduction method [11]. These methods consist of partitioning the sum of squares interaction effect $(\alpha\beta)_{ij}$ from equation (1) such that the sum of squares from each assessor is isolated. The variance partitioning for agreement is achieved by performing a reduced two-way ANOVA without interaction effect,

$$Y_{ijm} = \mu + \alpha_i + \beta_j + \varepsilon_{ijm}^{\text{red}}. \tag{6}$$

The unstandardized residuals from this reduced ANOVA are then subtracted from the unstandardized residuals from the full ANOVA,

$$\phi_{ijm}^{\text{agree}} = \varepsilon_{ijm}^{\text{full}} - \varepsilon_{ijm}^{\text{red}} \tag{7}$$

These differences ϕ^{agree} are the model effects for assessor by product interaction. A one-way ANOVA is conducted for each assessor i using the model effects from equation (6)

$$\phi_{jm}^{\text{agree}} = \mu + \beta_j + \varepsilon_{jm}^{\text{agree}}. \tag{8}$$

In this ANOVA analysis, the sums of squares for the product effect are a measure of the contribution of each assessor to the sums of squares assessor by product interaction effect. The agreement test is concluded by applying an F-test to determine whether each individual assessor contributes significantly to the sums of squares assessor by product

$$F = \text{MS}_{\text{red}}^{\text{agree}} / \text{MS}_{\text{error}}^{\text{full}} \tag{9}$$

From the F-test, it should be noted that a low p-value corresponds to a high agreement error.

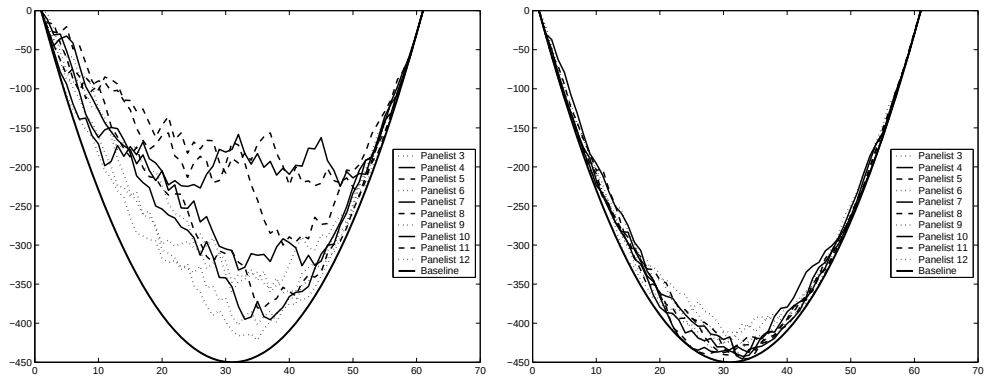

Figure 2: Eggshell plots for 10 panelists rankings against the consensus ranking for two attributes *Left*: Attribute Off-flavor. Noticeable panel agreement. *Right*: Attribute Hardness. High panel agreement.

3.6 Crossover effects

As discussed in section 2.2, one of the most severe contributions of errors in the panel, is a panelist showing crossover effects. Crossover effects occur when a panelist scores products opposite in intensity to the rest of the panel. A way to identify crossover effects is to partition the agreement error to determine if such effects are occurring. We define the difference between the panel mean from the product mean as

$$t_j = \bar{Y}_{.j.} - \bar{Y}_{...} = \frac{1}{IM}\sum_{i=1}^{I}\sum_{m=1}^{M} Y_{ijm} - \frac{1}{IJM}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{m=1}^{M} Y_{ijm}. \tag{10}$$

The sign of the difference values t_j can then be compared to the sign of the subtracted residuals ϕ_{jm}^{agree} to calculate the sums of squares crossover effect for panelist i. This can be formalized as

$$\text{SS}_{\text{crossover}} = \sum_{m=1}^{M}\sum_{j_-:t_{(j)}<0} r_{j_-}\phi_{j_-m}^2 + \sum_{m=1}^{M}\sum_{j_+:t_{(j)}>0} r_{j_+}\phi_{j_+m}^2 \tag{11}$$

where r has the function as a sign comparator with the following values

$$\begin{aligned} r_{j_-} &= 1 \quad \text{if} \quad \phi_{j_-m} > 0 \\ r_{j_-} &= 0 \quad \text{if} \quad \phi_{j_-m} < 0 \\ r_{j_+} &= 1 \quad \text{if} \quad \phi_{j_+m} > 0 \\ r_{j_+} &= 0 \quad \text{if} \quad \phi_{j_+m} < 0. \end{aligned} \tag{12}$$

The sums of squares crossover error $\text{SS}_{\text{crossover}}$ is then compared to the sums of squares agreement error $\text{SS}_{\text{agreement}}$ found from equation (8) to give the portion of the agreement error that is due to crossover. If the agreement error for a panelist is very small, even a large portion of the error stemming from crossover can be neglected. For large agreement error, it is adequate to investigate the amount due to crossover effects.

3.7 Eggshell plot

The eggshell plot is a graphical technique for illustration of assessor differences based on cumulative ranks instead of using assessor scores directly. The idea is to compute the consensus ranking and

then plot each of the assessors' ranks against the consensus [12]. Details on the implementation are described by Hirst & Næs [13]. Here, only a short overview is given as follows:

1. Identify the product ranking for each assessor and define a consensus ranking for the panel. A good approach to achieve a consensus ranking is to perform a principal component analysis on the matrix of all assessor ranks using assessors as variables. The scores of the first principal component are then ranked to give the consensus.

2. For every assessor, calculate the cumulative ranks in the consensus order and subtract the cumulative rank for an hypothetic assessor who ranks all products the same.

3. The cumulative rank difference can then be plotted to give one curve for every assessor against the consensus.

An interesting property of the eggshell plot, apart from its aesthetic appeal [14], is the available rank correlation, also called Spearman's Rho [15]. The rank correlation, defined by one minus the area between the assessor's curve and the consensus baseline, is a measure of the assessor's correlation with the consensus ranking. Figure 2 shows eggshell plots for a set of assessors for two different attributes with different level of rank correlation.

4 Material and methods for the application example

The sensory data used in the application was collected on 60 samples of wrinkle-seeded green peas (*Pisum sativum L.*) from 27 varieties. The peas were submitted to blanching and quick-freezing treatments, then packed and finally stored at -20°C. The sensory analysis was performed by 10 assessors over two replicates after steaming the peas over boiling water. Six attributes were scored on a scale from 1 to 9: pea flavour, sweetness, fruity flavour, off-flavour, mealiness and hardness. More details on the peas preparation and on the sensory data collection are given in [6].

The panel data is first submitted to a consonance analysis with PCA, then to the five assessor and panel tests described above. A PCA model is used on the test results to describe the properties of the set of tests. The multivariate analysis software *The Unscrambler*® [16] was used for PCA modeling; the *Panelist Monitoring and Tracking* software [17] was used for the set of ANOVA and eggshell tests.

5 Results and discussion

The five panel tests were run on 10 assessors from the described data set with 2 replication of 60 green peas. Two assessors (assessor 1 and 2) were left out of the analysis due to missing data. 6 different attributes were tested. The results from the five tests has then been submitted to a Principal Component Analysis for two purposes: firstly, to check the specificity and complementarity of the tests, and secondly, to validate some of the tests results multivariately.

5.1 Consonance analysis with PCA

The full data of sensory evaluations is submitted to a consonance analysis based on PCA models. The rows in the table correspond to the 60 pea samples; the columns contain the evaluations per assessor per replicate per attribute, that is to say 10 assessors x 2 evaluations x 6 attributes giving a total of 120 columns. Six successive PCA models are run, each of them focusing on one specific attribute. The correlation loadings for each of these models are given in Figure 6. On the plots, label i.j indicates the loading of assessor i, replicate j. The explained variance on the first principal

	Agreement		Sensitivity		Reproducibility		Crossover		Eggshell	
Attribute	Expert	Train	Expert	Train	Expert	Train	Expert	Train	Expert	Train
Pea flavour	9	-	3, 6, 7 9, 11, 12	8	7	4	3, 12	5, 8, 10	-	5, 8, 10
Sweetness	3, 7, 12	5, 8, 9	3, 4, 5, 6 7, 11, 12	8	3, 7	5, 9	4, 5	8, 9	4, 5	8
Fruity flavour	6, 9	3	3, 5, 6 9, 11, 12	7, 8	7, 10	8, 11	3, 6	7, 8	3	-
Off flavour	6, 9	3, 12	3, 6, 9 12	8	6, 9	4, 11	3	5, 8, 10	-	5, 8, 10
Mealiness	5, 6, 10 12	3, 4, 7	all but 7	7	6, 11	4, 8, 9	3, 4	7	3, 4, 9 10, 12	7
Hardness	3, 5, 8 10	4	all	-	4	9	4, 5, 9 10, 11	-	all	-

Table 1: Strengths (expert) and weaknesses (train) per attribute for the 10 assessors (numbered 3 through 12)

component can be interpreted as a percentage of panel agreement. The best panel agreement is observed for attribute hardness (81%); the lowest panel agreement is seen for attribute off-flavour (58%). The latter attribute may need to be further defined for the assessors. Assessor 8 is outlying for sweetness and needs further calibration training for this attribute; assessor 7 is outlying for mealy and requires further explanation and training for this attribute. Both assessors 7 and 8 are outlying for attribute fruity. Assessors 4, 5, 8 and 10 are slightly outlying for attribute pea flavour, possibly because of reproducibility or crossover errors. The nature of their errors will be identified in the detailed panel tests.

5.2 Univariate test results

From the set of univariate tests previously described, results are summarized in table 1. Since most of the panel in this case performed very well with only a few outliers, only assessors performing poor or extremely well are indicated to illustrate the key features of the test set. If the agreement test is evaluated first, the assessors not performing well are indicated. These assessors can be tracked further having a problem with either sensitivity, reproducibility or crossover, or eventually a combination. It should also be noted the large degree of correlation between the crossover error and the eggshell test, since crossover errors eventually will lead to an incorrect ranking. From the eggshell test, it can be seen that assessors performing well not necessarily will show good agreement. This is often due to poor reproducibility, which is not visible in the eggshell plot.

Specifically, assessor 8 clearly needs further training in attributes pea flavour, sweetness, fruity and off-flavour, with considerable errors in sensitivity, reproducibility and crossover. Assessor 7 shows poor performance for attribute mealiness and to some extent also fruity flavour. Only assessor 6 performs well in all situations, in that this assessor is not found to need any further training for all tests and attributes.

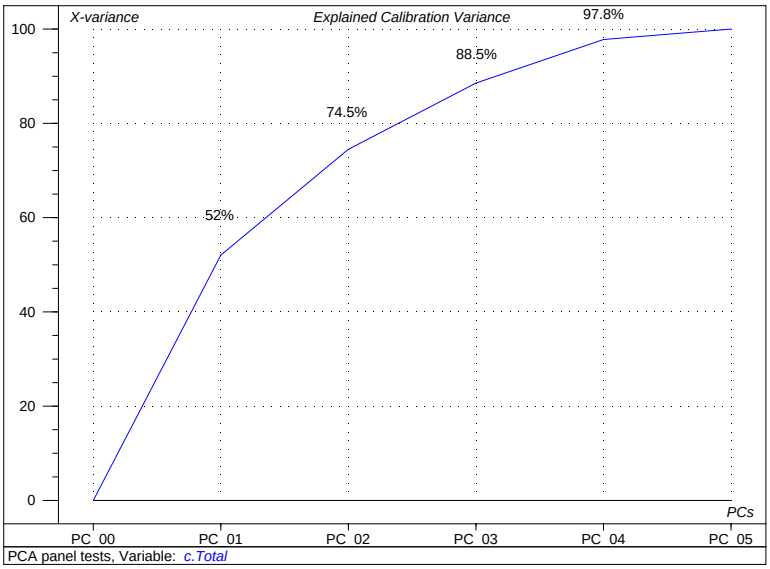

Figure 3: PCA model of the five panel tests, cumulative explained variance along five principal components. Four principal components are needed to describe the variation in the five tests, indicating a poor correlation and therefore a high specificity and complementarity of the tests.

5.3 Specificity and complementarity of the panel tests

A Principal Component Analysis (PCA) is conducted on the results from the five assessor and panel tests: assessor sensitivity, assessor reproducibility, crossover, panel agreement and eggshell test. The rows in the table are the six attributes over 10 assessors; the columns are the results from each of the five tests (p-values for the ANOVA tests and rank correlation for the eggshell test). The data was standardized and the model was cross-validated over sub-segments of 6 samples each, leaving out results from one assessor at the time.

Figure 3 shows the cumulative explained variance expressed in percentages along each model component. The first two principal components describe respectively 52 and 22% of the variance. Four principal components are required to describe most of the structured variance in the five tests. This shows that despite some clear correlations between the tests, each of the five tests checks for a specific type of panel error. The tests are to a certain extent complementary of one another.

The correlation loadings plot in Figure (4) maps the loadings for the five tests. On the plot, the outer circle represents 100% explained variance while the inner circle indicates the limit for 50% explained variance. General conclusions regarding the specificities of the five tests can be drawn, although one should note that these generalities remain to be verified by an external validation on other data sets.

The agreement test and the eggshell-correlation test are correlated to each other as they both give a measurement for panel agreement. The crossover test is negatively correlated to them, indicating that a high crossover effect leads to poor panel agreement. The reproducibility and sensitivity tests are not strongly correlated to any other tests, indicating that each of them checks for a specific type of error. The reproducibility test is described along PC2 and is logically not correlated to the eggshell test, which builds on assessor averages over replicates.

Figure 4 shows the scores of the PCA model for the five panel tests. The circle indicates a confidence interval of 95%, as this is given by Hotelling T^2 statistic. Three objects are detected

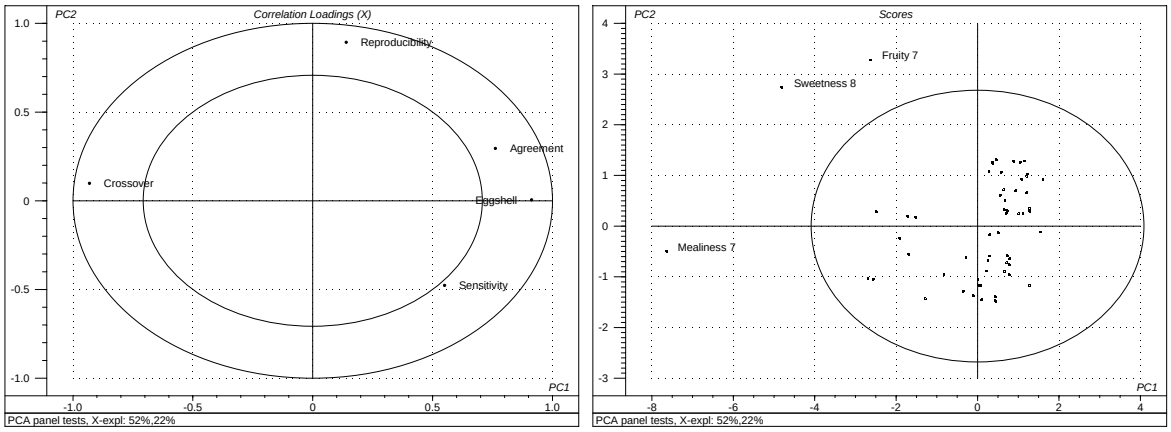

Figure 4: PCA model of the five panel tests. *Left*: Correlation loadings plot for the first two principal components. High crossover effects (left along PC1) lead to a poor ranking agreement in the eggshell test (right along PC1). The eggshell test does not detect not correlated to reproducibility. *Right*: Plot of scores for the first two principal components with 95% confidence ellipse. Three evaluations from assessors 7 and 8 are outlying.

as outliers as they lie outside the circle. These outlying evaluations are for mealiness by assessor 7, sweetness by assessor 8 and fruity flavour by assessor 7. This matches the results presented in table 1 which indicated sensitivity and crossover errors for assessors 7 and 8 for these attributes.

5.4 Stability assessment of the panel tests model

The PCA model describing the panel tests results was re-validated twice, first across assessors, then across attributes. In the model validated across assessors, 10 cross-validation segments are built each including 9 only of the 10 assessors. By studying the 10 resulting sub-models, we can observe the influence of each of the assessors on the global model for panel test description. Similarly, in the model validated across attributes the stability of the 6 sub-models in the cross-validation reflects the influence of each of the attributes on the global model for panel test description. The method is based on Jack-knifing and is described in details in [18].

Figure 5 shows the stability plots of loadings for the model validated across assessors (left) and across attributes (right). Each variable in the global model is surrounded by a swarm of its loadings from each of the sub-models. The middle of each swarm is the loading for the variable in the total model.

The validation across assessors shows that the crossover test is stable over assessors. Three of the tests show a large deviation for one of the sub-models in particular: the sub-model without assessor 8. This indicates that assessor 8 is the one with most variation from the group in eggshell, crossover and sensitivity tests. Assessor 3 is influential in the agreement and reproducibility tests. The validation across attributes indicates a very high stability of all tests, indicating that the results of the panel tests vary more across assessors than across attributes in this data set.

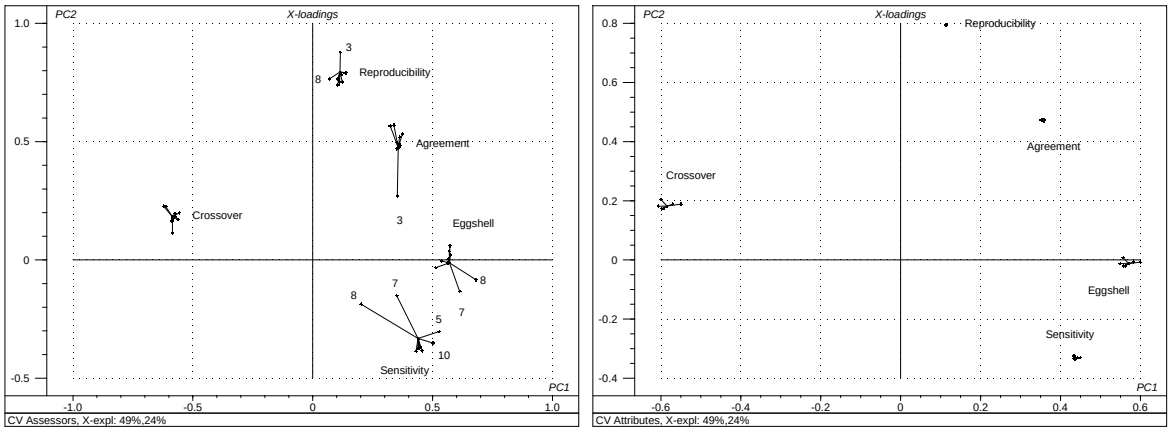

Figure 5: PCA model of the five panel tests. *Left*: Stability plot of loadings over assessors. The swarm around a variable corresponds to 10 sub-models, each including 9 only of the 10 assessors. The middle of the swarm is the loading for the variable in the total model. The sub-model without assessor 8 deviates largely from the other sub-models for four out of five panel tests. All assessors perform equally well in reproducibility. *Right*: Stability plot of loadings over attributes. The swarm around a variable corresponds to 6 sub-models, each including 5 only of the 6 attributes. The middle of the swarm is the loading for the variable in the total model. The sub-models without attributes fruity, mealy and sweetness deviate most from the other sub-models in the agreement, egg-shell and crossover tests. Very little variation is observed from attribute to attribute in reproducibility and sensitivity performances.

6 Conclusions

Five univariate test methods for assessing panel and assessor performance were described. Multivariate methods using PCA was used to conduct further analysis from the univariate tests and to conduct consonance analysis. A data set collected from a sensory panel tasting green peas has been used to illustrate the performance of the method. The 5 different univariate tests were shown to be able to pick up different types of panel and assessor errors in that two assessors clearly need retraining. Only a single assessor was performing well in all situations. A single test is not enough for identifying such errors, as the tests are complementary of one another. Univariate tests are more detailed than multivariate tests, which only pick-up major errors in reproducibility and panel agreement.

References

[1] Per Lea, Tormod Næs, and Marit Rødbotten. *Analysis of Variance for Sensory Data*. John Wiley and Sons Ltd., 1997.

[2] Per Bruun Brockhoff. Statistical testing of individual differences in sensory profiling. *Food Quality and Preference*, 14, 2003.

[3] David S. Lundahl and Mina R. McDaniel. The panelist effect- fixed or random? *Journal of Sensory Studies*, 3:113–121, 1988.

[4] Garmt Dijksterhuis. Assessing panel consonance. *Food Quality and Preference*, 6(1):7–14, 1995.

[5] Magni Martens. Sensory and chemical quality criteria for white cabbage studied by multivariate data analysis. *Lebensmittel-Wissenschaft und -Technologie*, 18:100–104, 1985.

[6] Tormod Næs and Bruce R. Kowalski. Predicting sensory profiles from external instrumental measurements. *Food Quality and Preference*, 4/5(1):135–147, 1989.

[7] David S. Lundahl and Mina R. McDaniel. Influence of panel inconsistency on the outcome of sensory evaluations from descriptive panels. *Journal of Sensory Studies*, 6:145–157, 1991.

[8] Marjorie C. King, John Hall, and Margaret A. Cliff. A comparison of methods for evaluating the performance of a trained sensory panel. *Journal of Sensory Studies*, 16(6):567–582, 2001.

[9] Tormod Næs. Handling individual differences between assessors in sensory profiling. *Food Quality and Preference*, 2:187–199, 1990.

[10] Charles E. McCulloch and Shayle R. Searle. *Generalized, Linear and Mixed Models.* Wiley Series in Probability & Statistics. John Wiley & Sons Inc, 2000.

[11] David S. Lundahl and Mina R. McDaniel. Use of contrasts for the evaluation of panel inconsistency. *Journal of Sensory Studies*, 5:265–277, 1990.

[12] Tormod Næs. Detecting individual differences among assessors and differences among replicates in sensory profiling. *Food Quality and Preference*, 9(3):107–110, 1998.

[13] David Hirst and Tormod Næs. A graphical technique for assessing differences among a set of rankings. *Journal of Chemometrics*, 8:81–93, 1994.

[14] Per Lea, Marit Rødbotten, and Tormod Næs. Measuring validity in sensory analysis. *Food Quality and Preference*, 6:321–326, 1995.

[15] S. Siegel and N.J. Castellan. *Nonparametric Statistics for the Behavioral Sciences.* McGraw-Hill, New York, 2nd edition, 1988.

[16] Camo (USA, Norway, and India). *The Unscrambler®.* www.camo.com, 2004.

[17] Camo (USA, Norway, and India). *Panelist Monitoring and Tracking.* www.camo.com, 2004.

[18] Harald Martens and Magni Martens. Modified jack-knife estimation of parameter uncertainty in bilinear modelling by partial least squares regression (plsr). *Food Quality and Preference*, 11:5–16, 2000.

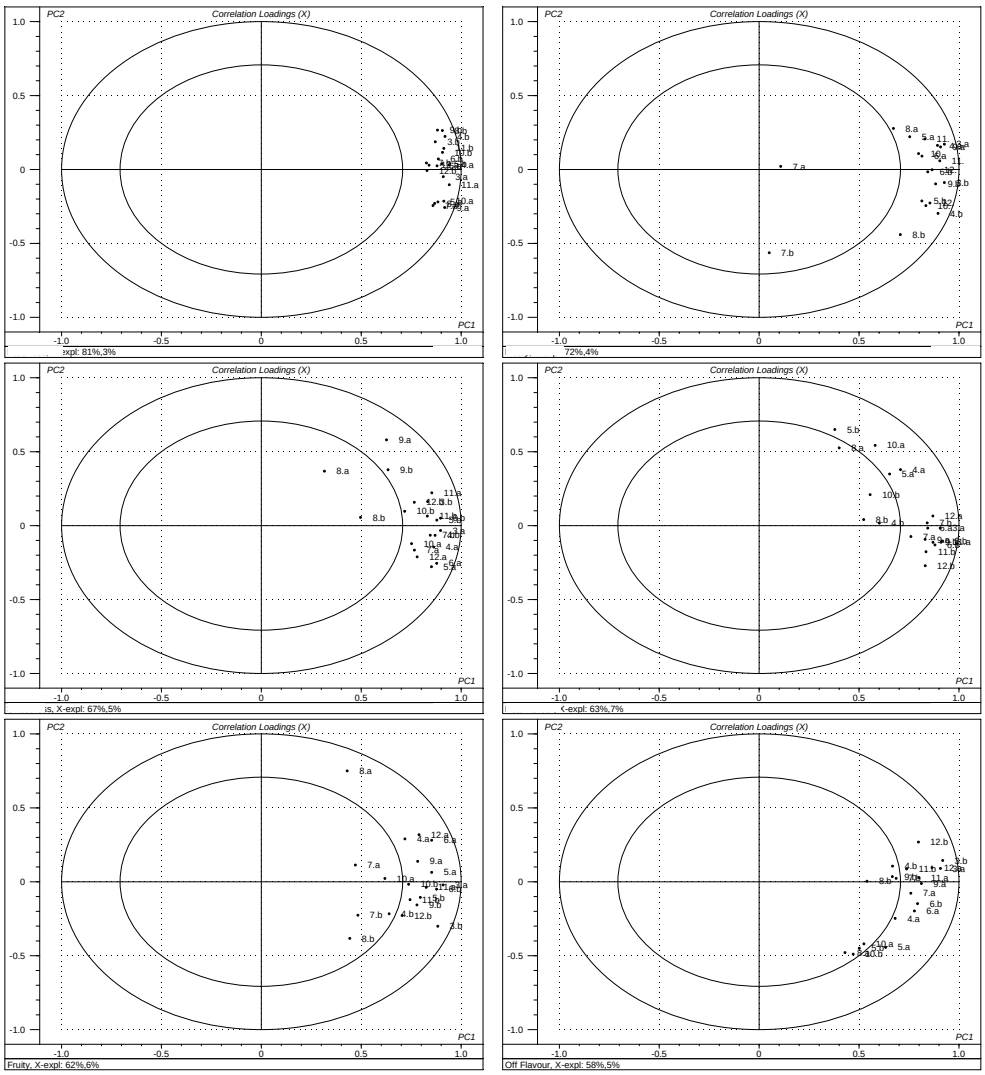

Figure 6: Consonance analysis for the six attributes, correlation loading plots. The plots are shown in the order from the strongest to the lowest panel agreement. *Top left*: Hardness (81% agreement), *Top right*: Mealy (72%), *Middle left*: Sweetness (67%), *Middle right*: Pea flavour (63%), *Bottom left*: Fruity (62%), *Bottom right*: Off flavour (58%). Assessor 7 needs further training for mealy and fruity, assessor 8 for sweetness and fruity. The panel needs further training in off-flavour.