Discriminant models based on sensory evaluations: single assessors versus panel average

P. M. Granitto[1,2*], F. Biasioli[1], I. Endrizzi[1], F. Gasperi[1]

[1]IASMA Research Center – Agrifood Quality Department, Via E. Mach, 1 – 38010 San Michele all'Adige (TN) – Italy
[2] CIFASIS (CONICET-UNR-UPC) Bv. 27 de Febrero 210 bis – 2000 Rosario – Argentina.

Abstract:
Product classification based on sensory evaluations can play an important role in quality control or typicality assessment. Unfortunately its real world applications face the difficulties related to the cost of a proper sensory approach. To partially overcome these issues we propose to build discriminant models based on the evaluation of single assessors and develop an appropriate method to combine them. We compare this new strategy with the more traditional one based on the panel average. We consider as applicative examples two datasets obtained from the sensory assessment of diverse cheese typologies from North Italy by two different panels. Also, we apply diverse, innovative and noise resistant discriminant methods (Random Forest, Penalized Discriminant Analysis and discriminant Partial Least Squares) to show that our new strategy based on modeling each individual assessor is efficient and that this result is independent of the classifier being used. The main finding of our work is that using noise-resistant multivariate methods, product discrimination based on the combination of independent models built for each assessor is never worse than discrimination based on panel average and that the error reduction is higher in the case of low consonance between assessors. Experiments on the same datasets adding random uniform values (noise) with different intensities support these findings. We also discuss a demonstrative experiment using different sets of attributes for each assessor. Overall, our results suggest that, if the goal is product classification, the consonance among assessors or even the use of the same vocabulary seem not necessary, the key factor being the discrimination capability and repeatability of each judge.

Keywords: *sensory profiling, discriminant analysis, discriminant Partial Least Squares, Random Forest, Penalized Discriminant Analysis.*

[*]Corresponding author. Tel: +39-0461-615187; fax: +39-0461–650956. E- mail address: granitto@cifasis-conicet.gov.ar

1. Introduction

Quality control and product classification are often performed by discriminant methods based on chemical and physical data or chemometric evaluations (Reida, O'Donnell & Downey, 2006). In the case of food products, however, sensory characteristics (the ones eventually perceived by the consumers) play a key role. This fact motivates the potential interest on discrimination of food and beverages on the basis of sensory profiling data, not only for quality control but also for typicality assessment. The typicality of products and its certification underlines the fact that these products should present sensory characteristics comparable with a known standard. This is usually checked by comparing the intensity of single attributes (positive characteristics and defects) with given thresholds as discussed, e.g., by Peretz Elortondo, Ojeda, Albisu, Salmeròn, Etayo & Molina (2007) but the spreading of multivariate methods induces the use of classification models based on overall profiles (Cocchi, Rasmus, Durante, Mancini, Marchetti, Saccani, Sighinolfi & Ulrici, 2006). For example, panels of trained assessors are used to monitor olive oil (IOOC, 1996) or Scotch whisky (Jack & Steele, 2002). In both cases (and others) sensory control of quality is mandatory. As explained by Jack and Steele (2002), in this context the sensory evaluations must be carried out by highly trained and experienced panelists. But even in this case, asking a panelist directly if a product is typical (or not) is not an ideal approach. Panelists need a clear picture in their heads in terms of exactly which sensory properties constitute the typical product and which do not. This requires a vast experience of evaluating many different samples of the product. Even with such experience, the ability of the panelists to retain such large amounts of sensory information in their memories can be questionable. Also, panelists can be influenced by their own personal tastes, and products with less preferred characteristics may be incorrectly classified as ''not typical''. They may also be unwilling to commit, knowing the weight of their evaluations. In the same work, the authors explain that a more objective approach to the sensory evaluation in authenticity/typicality assessments can be obtained if the panel is limited to evaluate the sensory characteristics of the product and the classification (in typical or not, for example) is performed by a discriminant function (unknown to the panelists). But even this alternative approach requires the use of a highly trained and concordant panel. In this paper we discuss a new strategy to use the evaluations of a panel of assessors as a discriminant instrument between products (or qualities): the final decision, e.g. typical/non typical, is based on a model that exploits the discriminatory capability of each panel member, thus maximizing the overall performances even in the case of low consonance.

The most difficult and time consuming stage of sensory profiling is probably

the selection and training of a panel, in the attempt to reach an agreement in the use of scales and in the meaning of each attribute. But, even in successful cases, small differences between assessors are unavoidable (the assessor effect) and linear scaling and projection methods (Lea, Næs & Rødbotten, 2001, Martens & Martens, 2001) can only partly overcome the difficulties induced by this effect. The basic strategy in sensory profiling, after a proper training of the panel, is to consider (and treat) the remaining differences between assessors as normally distributed random values. Thus, evaluations from all the panelists are averaged, usually after a linear scaling to correct for minor differences in the use of the scale (Lea, Næs & Rødbotten, 2001). If there are systematic differences among assessors this averaging/scaling process is not optimal (even more if non-linear relations are present). More elaborated strategies for dealing with non-concordant assessors include general linear transformations, for example General Procrustes Analysis (Gower, 1975), or linear models including extra parameters to account for scale differences between assessors, for example Brockhoff models (Brockhoff, 1994). These strategies are useful in order to analyze the differences among assessors, but their performance in our scope (to use a panel as a discriminant instrument) is poor , as we will show later in Section 3.

In this paper we present a new and practical method to use sensory evaluations as discriminant instruments. We propose to develop an individual classifier (i.e. a discriminant model) for each single assessor and then to combine their outputs to produce an average prediction. In a sense, we simply change the order of the process: usually assessors' observations are averaged first and then a classifier is fitted; here we develop a classifier for each individual assessor and then look for a “consensus” of the discriminant functions.

The main potential problem with our new method is that evaluations made by individual assessors are usually too noisy to be successfully modeled with traditional statistical methods, as Linear or Quadratic Discriminant Analysis. However, in the last years several noise-resistant multivariate discriminant methods were developed (Hastie, Tibshirani, & Friedman, 2001). We selected three of these innovative methods (discriminant Partial Least Squares, Random Forest and Penalized Discriminant Analysis) to show that our new strategy, coupled with a noise-resistant classifier, can accurately discriminate between classes of products.

To evaluate this new strategy, we compared it with two versions of the more typical one consisting in taking the panel average before any discriminant modeling. We use as examples data from two different panels that evaluated hard and semi-hard cheeses from North-Italy. The selected case studies deal with the issue of assessing the typicality of local cheese productions. In both

studies we sampled the production over a long period of time because we expected non homogeneous characteristics. This is the well known variability related to technological aspects of the cheese making process (seasonal variation of milk characteristics and natural variability of the semi-industrial manufacturing process involved). Because of this, the data considered here as case studies has a particular setting, where the products were assessed several times during 12 to 15 months.

To check the robustness of the averaging methods, we also simulated a decrease in "assessor precision" by adding normally distributed random values (artificial noise) to their observations. We compared the different strategies again under these new conditions.

A potential advantage of our new method is that, in principle, it does not require panel agreement. It can even be applied to free-choice-profiling data. To evaluate this possibility, we finally performed an experiment using different subsets of attributes for each assessor.

2. Materials and Methods

2.1 Methods for combining evaluations

We compare three different methods for scaling and combining the evaluations of individual assessors. Figure 1 shows an outline of the three methods.

BAM (Brockhoff Averaging+Modeling):
In this method we first linearly transformed each score according to $Y_S=\alpha+\beta Y_0$, where Y_S is the new (scaled) score, Y_0 is the original score and α and β are the standardizing coefficients. For each attribute and assessor we determined the optimal α and β values according to the Brockhoff algorithm (Brockhoff, 1994), which finds in each case the values that produce the best consensus in the panel. The scaled scores were then averaged over assessor in order to obtain the panel consensus. Finally, discriminant functions were fitted to this scaled data.

SAM (Standardizing+Averaging+Modeling):
This method follows the most typical procedure in sensory analysis. Again, observations from all assessors were first standardized to reduce linear differences in the use of the scale. But, in this case, the α and β values were taken, respectively, as the mean value and the standard deviation of the observations for each attribute and assessor (in order to linearly transform the data to zero mean, unit variance). Thereafter, observations from all assessors were averaged (to produce the mean-panel or consensus values) and classifiers were fitted to this scaled data as a final step. This method is

very similar to the BAM previously described, the difference is that the scaling process is independent for each assessor.

MA (Modeling+Averaging):
This is the new strategy we propose to overcome difficulties related to differences between assessors. In this case we adjusted an individual classifier to the evaluations made by each assessor. Standardization is not required in this method, but was performed anyway to ensure the same range for all attributes, which improves the performance of the classifiers.
We then needed to “average” the predictions of the individual models to form a combined prediction. Most discriminant methods (including the three used in this work) classify samples by estimating the a posteriori probabilities of each product given the measured values of the attributes. Following the typical procedure in multiple classifier systems (Kuncheva, 2004), we estimated for all discriminant models the average posterior probability of each product and assigned each sample to the product with maximum average posterior probability.

2.2 Discriminant methods.

Over the last decade new and powerful statistical learning methods were developed, in particular related to ensemble methods and to the generalization of linear models (Hastie, Tibshirani & Friedman, 2001). For this work we selected a representative method from each of these two classes and we also implemented dPLS models, which are well known in sensory applications. We have successfully used these classifiers to discriminate strawberry cultivars on the basis of spectrometric data (Granitto, Biasioli, Aprea, Mott, Furlanello, Mark & Gasperi, 2007), finding that all of them showed similar performances for that task. We used implementations available as free packages for the R statistical environment software (R Development Core Team, 2005). In the following we briefly describe each classifier, mainly to show that they have very different basis. We highlight again that we are not very interested in the relative performance of these classifiers; we use them only to show that the differences between the three methods for combining evaluations are independent of the (powerful) classifier being used.

Discriminant Partial Least Squares (dPLS):
The dPLS algorithm has been extensively described and used in the chemometrics and sensory analysis literature (Wold, Sjöström & Eriksson, 2001, Martens & Martens, 2001). Basically, it has two steps, a PLS projection followed by the application of LDA in the projected subspace. The number of scores used in the projection step plays the role of a regularizing parameter, safeguarding against overfitting.

Random Forests (RF):
In a previous work (Granitto, Gasperi, Biasioli, Trainotti & Furlanello, 2007) we introduced the use of the RF algorithm in sensory analysis, describing its characteristics and possibilities. Basically, RF is an ensemble of decision trees created following a particularly efficient strategy aimed at increasing the diversity between the trees (Breiman, 2001). The combination of two different sources of diversity (fitting on bootstraps plus selecting at each node only from a subset of attributes) produces easy-to-build ensembles with very good performance as prediction tools. The RF algorithm has, in practice, only one free parameter: the number m of attributes made available at each node during the growing of trees. Following Breiman (2001), we set m to the square root of the total number of attributes M, which is the default value and usually gives near optimal results.

Penalized Discriminant Analysis (PDA):
PDA (Hastie, Buja & Tibshirani, 1995) is a regularized version of the traditional Linear Discriminant Analysis (LDA) (Ripley, 1996), more appropriate for situations with similar number of products and samples (i.e., in ill conditioned situations where LDA is prone to overfit). The method is based on recasting the LDA problem as a regularized linear regression one, and then to apply any of the many well-known techniques available for this task. We use standard Ridge Regression (Hastie, Tibshirani & Friedman, 2001), which has only one free parameter, the ridge constant lambda that penalizes high values of the fitted variables.

2.3. Evaluation method

We evaluated the three methods for combining evaluations by estimating their mean prediction error over the two datasets described below (Section 2.6). The mean prediction error (or classification error) is defined as the fraction of incorrectly predicted products for a given set of observations (i.e. the number of errors divided by number of samples in the set).

In order to produce an unbiased and accurate estimation of prediction errors we repeated 20 times a 5-fold internal cross validation procedure. Each time we split the 60 samples of each dataset at random in 5 subsets or folds, keeping products balanced. At each time, a fold was used as a test set and the remaining 4 as a learning (or training) set. This means that, each time, we used only the 48 samples in the learning set to:
i) normalize the observations according to one of the methods,
ii) choose the values of free parameters, like the value of the ridge constant for PDA or the number of scores kept by dPLS using an internal cross validation (over the learning set only),
iii) fit the three classifiers described before (RF, PDA and dPLS).

Finally, we used the obtained classifiers to predict the products corresponding to the samples left in the test set and to estimate the corresponding mean classification error. In all cases the test sets were also normalized using the α and β parameters estimated using the training set only. The same procedure was repeated for the three strategies for combining predictions (BAM, SAM and MA), using in all cases the same splits into learning and test sets. Results are always reported as mean values over the 100 estimations of mean discrimination errors on independent test sets.
It is worth mentioning that some assessors missed a few sessions, and for that reason some samples were not evaluated by all assessors. In those cases, when modeling individual assessors, we simply used all the available samples in the training set to adjust the models and estimated discrimination errors with the (sometimes reduced) test set.

2.4 Noise addition

To check the robustness of the averaging methods we added noise (uniform random values) to (raw) assessor evaluations and repeated the experimental method described before (2.4). We used four different noise levels. In all cases we added to each single observation a random number from a uniform distribution in [-a,a], with a = 10, 30, 60 and 100 respectively. This procedure was done before any scaling, with all data in the original [0, 100] scale.

2.5 Subsets of attributes

To evaluate the possibility of using different attributes for each assessor (to simulate the use of free-choice-profiling data) we performed two additional experiments. In both cases we basically repeated the full experimental method described in 2.4 for the MA method, but using for each assessor only subsets with 15 attributes (nearly half of the original sets). In the first case we simply selected the subset of attributes at random. In the second experiment, we used for each assessor the subsets of attributes with the highest discriminant power according to the ANOVA analysis (best subset for each training set).

2.6 Sensory assessment

We used data from two different panels to show with examples the potential of our method. In the next paragraphs we briefly describe the sensory assessment process (more details are beyond the scope of this paper).
Sensory profiling was carried out by two different panels of assessors, selected and trained according to specific procedures for sensory evaluation of hard and semi-hard cheese (Gallerani, Gasperi & Monetti, 2000, Lavanchy

et al.,1993, Berodier, Lavanchy, Zannoni, Casals, Herrero & Adamo, 1997 and Murray & Delahunty, 2000).
The first panel (8 assessors) described six typical “Nostrani” cheeses interesting as possible candidates of PDOs (Protected Designation of Origin) : “Puzzone di Moena”, “Spressa delle Giudicarie”, “Vezzena”, “Nostrano del Primiero”, “Nostrano della Val di Non” and “Nostrano della Val di Sole”. All of them are made with raw bovine milk, in six different cheese factories located at Trentino area (Gasperi, Biasioli, Framondino & Endrizzi, 2004). Sixty samples, 10 for each product, have been collected during a period of 15 months, two samples every three months, in order to cover the possible time-variability of the local production.
The second panel (9 assessors) worked on four “Grana” cheeses with different origin and ripening degree. “Parmigiano Reggiano” is made in a restricted region of the Po valley, including the provinces of Reggio Emilia, Parma and, partly, of Modena, Mantova and Bologna. “Grana Padano”, on the other side, is produced in a large area of North Italy. The other two, “Grana Trentino” and “Grana Trentino Giovane”, are varieties of “Grana Padano” produced exclusively at Trentino province, being different only in their ripening degree. These Italian hard cheeses are made from raw bovine milk, partly skimmed by creaming, with the addition of a natural whey starter (Battistoni & Corradini, 1993). Sixty samples collected during 12 months (15 for each class, each one from a different cheese factory), with ripening degrees representative of the products present on the market, have been analyzed in this case.
Both panels, indicated here as “Nostrani” and “Grana” developed profile protocols containing 35 and 30 attributes, respectively (Granitto et al, 2007a), according to the consensus method. The intensity of each attribute was evaluated on a 100 mm unstructured scale anchored at each extreme. At each session the panel evaluated a set of 6 samples, one sample of each product, presented in an order balanced for assessor, sample and presentation. The same cheeses were replicated twice in successive sessions (with a week held between them). Measurements from both replications were averaged since ANOVA analysis of both datasets (Lea et al., 2001) showed that there is no replication effect.
The final “Nostrani” dataset comprises 60 evaluations over 35 attributes and 6 products. The “Grana” dataset also comprises 60 evaluations over 4 products and 30 attributes.

3. Results

3.1 Panelists' concordance and discriminant capabilities

In Table 1 we show the values of Pearson's Correlation Coefficient (r) between panel means and each assessor for the "Nostrani" dataset. In the third column we list mean r values (over the 35 attributes) for each assessor. All 8 assessors show high mean r values, with a minimum of 0.59 and an average of 0.67. Evaluating the correlation of single attributes (fourth and fifth columns, showing respectively the number of attributes with assessor-panel correlation <0.4 and <0.2), again all assessors show good performances, being only assessors #4 and #7 slightly worse than average. In Table 2 we show the corresponding results for the "Grana" dataset. This panel shows a different behavior, with a low minimum r value of 0.37 and an average of 0.53. Assessors #2 and #3 in particular have very low correlation with the panel. The analysis of the correlation of single attributes suggest that this panel has lower internal agreement than the "Nostrani" one.

In the same tables we also compare the performance of discriminant models adjusted over the evaluations of individual assessors (i.e. each discriminant model predicts which is the evaluated product using only the observations made by one of the assessors). For the "Nostrani" dataset (Table 1), the three classifiers have similar performances in average. The worst results are shown by two assessors who missed some sessions as indicated in the second column (assessors #4 and #6). The same analysis in the "Grana" dataset (Table 2) shows that, again, the three classifiers have similar average performances, with differences only for particular cases, such as assessors #3 or #8. Overall, the nine assessors show similar discriminant capabilities. Two of them, #5 and #6, show a slightly better performance than the average and other two, #2 and #3 a worse one. These latter two assessors missed some evaluation sessions, which could explain their low performance (discriminant models were adjusted using reduced training sets in these cases).

3.2 Averaging strategies

Tables 3 and 4 compare the results of the three strategies for panel averaging previously described. In both tables the three classifiers (RF, PDA and dPLS) show very similar results, which supports the conclusion that the performance of the methods for combining assessors is independent of the classifier being use, providing that the classifier is efficient (for comparison, the MA method with a classical LDA classifier on the "Nostrani" dataset gives a mean discrimination error of 0.41). Therefore, we will analyze only the mean value of the three innovative classifiers from here on.

All the results in Tables 3 and 4 are clearly better than those obtained by individual assessors (Tables 1 and 2) indicating that averaging (in any way) the evaluations of several assessors effectively produces better discrimination, as expected.
In both cases there is a decrease in mean classification error from the BAM method to SAM and finally to MA, which suggests that general linear scaling cannot explain completely the differences between assessors. The decreases in error are moderate for the "Nostrani" dataset but very important for the "Grana" dataset, which showed a lower concordance between assessors. On this last dataset there is a 35% reduction in discrimination error from SAM to MA, and a 45% reduction from BAM to MA.

3.3 Noise addition

In Figure 2 we compare the three combination methods after the addition of uniform noise. On the “Grana” dataset (left panel) the differences between the three methods are clear with all but the highest noise levels. Only the addition of very high uniform noise deteriorates the relative performance of the new MA strategy compared to SAM. The same qualitative results can be observed for the “Nostrani” dataset (right panel).

3.4 Modeling with subsets of attributes

In Table 5 we show the results of the demonstrative experiment with subsets of attributes (15 best or random attributes). On both panels, the use of subsets of 15 attributes for each individual assessor produces discriminant results only slightly worse than those obtained using all the attributes, and equal or better than those obtained with the traditional SAM or BAM methods with the full set of attributes (Tables 3 and 4). For comparison, we also included in Table 5 the discriminant errors produced by the SAM mean panel in the same conditions (15 best or random attributes), which show a clear deterioration in discrimination performance. We repeated these experiments with different subset lengths (data not shown) obtaining similar qualitative results.

4. Discussion

At the level of individual assessor the panels seem equivalent in some aspects (Tables 1 and 2): they both present a couple of assessors who clearly outperform average results (in correlation and discrimination) and another couple who show lower than average capabilities. In both panels those assessors missed some sessions. However (and relevant to our work), the “Grana” panel has, in general, a lower concordance between the assessors.

Correlation with the panel and discriminant capabilities are not always equivalent. Assessor #6 in the “Grana” panel is an interesting example, showing a relatively low r (0.52) but the best discriminant results. It is clear that this assessor is evaluating the samples in a different (but consistent) way. The possibility of using this information without any necessity of concordance is at the basis of the better performances of the MA method.

The analysis of discriminant errors of single assessors is an interesting by-product of the proposed MA methodology, useful even for classical panel applications. Potentially, it allows the monitoring of panel and assessor performances in a different way, detecting cases (as the previously discussed one) where consistent and reliable assessors do not agree with the panel.

In this work we applied three classifiers based on very different principles: ensemble methods, generalization of linear models and partial least squares classification. In all our experiments (Tables 1 to 5, Figure 2), the three classifiers showed very similar discrimination capabilities for the datasets considered. This suggests that our findings do not depend on the discriminant method applied (providing that the method is efficient).

On the “Nostrani” dataset the use of the new MA strategy produced only a small improvement in discrimination performance over the SAM method, while for the “Grana” dataset there is a one third reduction in discriminant error. It seems evident that some assessors have strong differences in the use of the scale in the latter case. The new MA method can easily deal with this problem, producing mean classification errors lower than the two more typical strategies (BAM and SAM) for both datasets.

The poor performance of the BAM method is probably the result of an "overfitting" issue. Brockhoff models (or GPA) are useful tools for understanding and describing the differences between assessors, but they probably need to be regularized (in a statistical sense) in some way to be useful for discrimination purposes.

Experiments on the same datasets after the addition to the profile data of random uniform noise with different intensities indicate that these results stand in almost all cases. Only for very noisy data (when we add noise in the same range of the original scale) models based on traditional panel average work clearly better. In those cases, however, discrimination errors are too high to be of any practical use.

Our demonstrative experiment with subsets of features indicates that the MA method is less affected by the use of different subsets of attributes for each assessor. This latter finding (even if additional experiments are needed

to confirm the result) suggests that our methodology, relying only on the consistency of single judges, could be possibly extended to free choice profiling data.

Concluding, the main finding of our work is that, regardless of the classification method used, discrimination based on the combination of independent models built for each assessor with modern multivariate methods is better than the discrimination obtained using panel average. This opens an interesting prospective because it indicates that, if the goal is product discrimination, the best strategy is also the less time consuming. The key factor in this case is the discrimination capability and repeatability of each individual member of the panel. The consonance among single assessors or even the use of the same vocabulary could not be necessary, simplifying considerably the training process. We envisage the possibility to simply extend this method to the case of free choice profiling data. Our method provides also, as a by-product, a direct way to assess panelists' performances for product discrimination.

Acknowledgments

Work partially supported by PAT projects MIROP and SAMPPA and ANPCyT grant PICT 11-15132. We thanks Guillermo Hough for useful comments on a previous manuscript.

Bibliography

Battistoni, B., & Corradini, C. (1993). Italian Cheese. In: Cheese, chemistry, physics and microbiology, ed. P.F. Fox, Chapman & Hall, London, 1993.

Berodier, F., Lavanchy, P., Zannoni, M., Casals, J., Herrero, L., & Adamo, C. (1997). Guide d'Evaluation Olfacto-Gustative des Fromages a Pate Dure et Semi-dure. Food Science and Technology / Lebensmittel-Wissenschaft und-Technologie 30, 653-664.

Breiman, L., (2001). Random Forests. Machine Learning, 45(1), 5-32.

Brockhoff, P & Skovgaard, Ib M. (1994). Modelling individual differences between assessors in sensory evaluations. Food Quality & Preference, 5, 215-224.

Cocchi M, Rasmus B., Durante C., Mancini D., Marchetti A., Saccani F., Sighinolfi S., Ulrici A. (2006) Analysis of sensory data fo Aceto Balsamico Tradizionale di Modena (ABTM) of different ageing by application of PARAFAC models. Food Quality & Preference, 17, 419-428

Gallerani G., Gasperi F., & Monetti A. (2000). Judge selection for hard and semi-hard cheese sensory evaluation. Food Quality & Preference, 11, 465-474.

Gasperi F., Biasioli F., Framondino V., & Endrizzi I. (2004). Ruolo dell'analisi sensoriale nella definizione delle caratteristiche dei prodotti tipici: l'esempio dei formaggi trentini / The role of sensory analysis in the characterization of traditional products: the case study of the cheese from Trentino. Sci. Tecn. Latt.- Cas , 55, 345-364.

Gower, J. C. (1975). Generalized Procrustes analysis. Psycometrika, 40, 33-51.

Granitto, P.M., Gasperi, F., Biasioli, F. & Furlanello, C. (2007a). Predictive modeling and feature selection for sensory analysis. International Journal of Pattern Recognition and Artificial Intelligence , submitted.

Granitto, P.M., Gasperi, F., Biasioli, F., Trainotti, E., & Furlanello, C. (2007b). Modern data mining tools in descriptive sensory analysis: a case study with a Random Forest approach. Food Quality & Preference, 18, 681-689.

Granitto, P.M., Biasioli, F., Aprea, E., Mott, D., Furlanello, C., Mark, T.D., & Gasperi, F. (2007c). Coupling Proton Transfer Reaction-Mass Spectrometry with data mining techniques: classification of strawberry cultivars. Sensors and Actuators B, 121:2, 379-385.

Hastie, T., Buja, A., & Tibshirani, R. (1995). Penalized Discriminant Analysis, Annals of Statistics, 23, 73-102.

Hastie, T., Tibshirani, R., & Friedman, J.H. (2001). The elements of statistical learning. Springer-Verlag: New York.

IOOC (International Olive Oil Council) (1996). Sensory Analysis of Olive Oil. http://www.internationaloliveoil.org, accesed 17/05/07.

Jack, F. R., & Steele, G. M. (2002). Modeling the sensory characteristics of Scotch whisky using neural networks—a novel tool for generic protection. Food Quality and Preference, 13, 163–172.

Kuncheva, L.I. (2004). Combining Pattern Classifiers. Methods and Algorithms, Wiley: New York.

Lavanchy, P., Bérodier, F., Zannoni, M., Noël, Y., Adamo, C., Squella, J., & Herrero, L. (1993). L'Evaluation Sensorielle de la Texture des Fromages à Pâte Dure ou Semi-dure. Etude Interlaboratoires. Food Science and Technology / Lebensmittel-Wissenschaft und-Technologie 26 (1), 59-686

Lea, P., Næs, T. & Rødbotten, M. (2001). Analysis of Variance for Sensory Data. Wiley: New York.

Martens, H. & Martens, M. (2001). Multivariate Analysis of Quality: An Introduction. Wiley: New York.

Murray, J. M., & Delahunty, C. M. (2000). Mapping consumer preference for the sensory and packaging attributes of Cheddar cheese. Food Quality and Preference 11, 419-435.

Pérez Elortondo F. J., Ojeda M., Albisu M., Salmeròn J., Etayo I., Molina M. (2007) Food quality certification; An approach for the development of accredited certification methods. Food Quality & Preference, 18, 425-439

R Development Core Team (2005). R: A language and environment for statistical computing. R Foundation for Statistical Computing: Vienna. (http://www.R-project.org.)

Reida, L.M., O'Donnell, C.P., & Downey, G. (2006). Recent technological advances for the determination of food authenticity. Trends in Food Science & Technology, 17, 344–353.

Ripley, B.D. (1996). Pattern Recognition and Neural Networks. Cambridge University Press: Cambridge.

Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics, Chemometrics and Intelligent Laboratory Systems, 58, 109-130.

Assessor	S	Correlation			Discriminant Analysis			
		Mean	<0.4	<0.2	RF	PDA	dPLS	Mean
1	60	0.66	1	0	0.38	0.41	0.41	0.40
2	56	0.68	2	0	0.34	0.31	0.37	0.34
3	60	0.75	0	0	0.36	0.33	0.41	0.37
4	40	0.59	5	2	0.58	0.59	0.63	0.60
5	60	0.68	1	0	0.39	0.35	0.36	0.37
6	48	0.64	3	1	0.56	0.53	0.53	0.54
7	54	0.61	5	2	0.38	0.42	0.38	0.39
8	60	0.70	2	0	0.33	0.30	0.35	0.33
Max		0.75	5	2	0.58	0.59	0.63	0.60
Min		0.59	0	0	0.33	0.30	0.35	0.33
Mean		0.67	3.0	0.6	0.42	0.40	0.43	0.42

Table 1: Evaluation of individual assessor's performance for the “Nostrani” dataset. The second column (S) shows the number of samples evaluated by each assessor. For the correlation analysis the “Mean” column shows the mean assessor-panel correlation (taken over all attributes) and columns labeled “<0.4” and “<0.2” show the number of attributes with assessor-panel correlation lower than that value. In the case of Discriminant Analysis, results are mean discrimination errors over 100 test sets. RF, PDA and dPLS are the 3 classifiers described in the text, “Mean” column shows the mean of the 3 classifiers.

Assessor	S	Correlation			Discriminant Analysis			
		Mean	<0.4	<0.2	RF	PDA	dPLS	Mean
1	60	0.53	10	1	0.33	0.36	0.35	0.35
2	48	0.46	12	4	0.47	0.48	0.50	0.49
3	32	0.37	17	7	0.45	0.53	0.44	0.47
4	60	0.51	12	1	0.40	0.43	0.38	0.41
5	60	0.59	3	0	0.30	0.32	0.32	0.31
6	60	0.52	9	3	0.31	0.27	0.28	0.29
7	60	0.58	6	0	0.37	0.39	0.35	0.37
8	60	0.57	7	1	0.36	0.46	0.41	0.41
9	60	0.58	5	1	0.40	0.39	0.42	0.40
Max		0.59	17	7	0.47	0.53	0.50	0.49
Min		0.37	3	0	0.30	0.27	0.28	0.29
Mean		0.52	9.0	2.8	0.38	0.41	0.39	0.39

Table 2: Evaluation of individual assessor's performance for the “Grana” dataset. Columns are the same as in Table 1.

Method	RF	PDA	dPLS	Mean
BAM	0.26	0.26	0.28	0.27
SAM	0.22	0.21	0.28	0.24
MA	0.21	0.21	0.23	0.22

Table 3: Comparison of the three different methods for panel averaging (discussed on the text) on the “Nostrani” dataset. Rows correspond to the averaging methods, columns to the different classifiers and the mean value of them. Results are mean discrimination errors over 100 test sets (lower results are better).

Method	RF	PDA	dPLS	Mean
BAM	0.33	0.39	0.35	0.35
SAM	0.29	0.29	0.29	0.29
MA	0.20	0.17	0.18	0.19

Table 4: Comparison of the three different methods for panel averaging on the “Grana” dataset. Details are similar to Table 3.

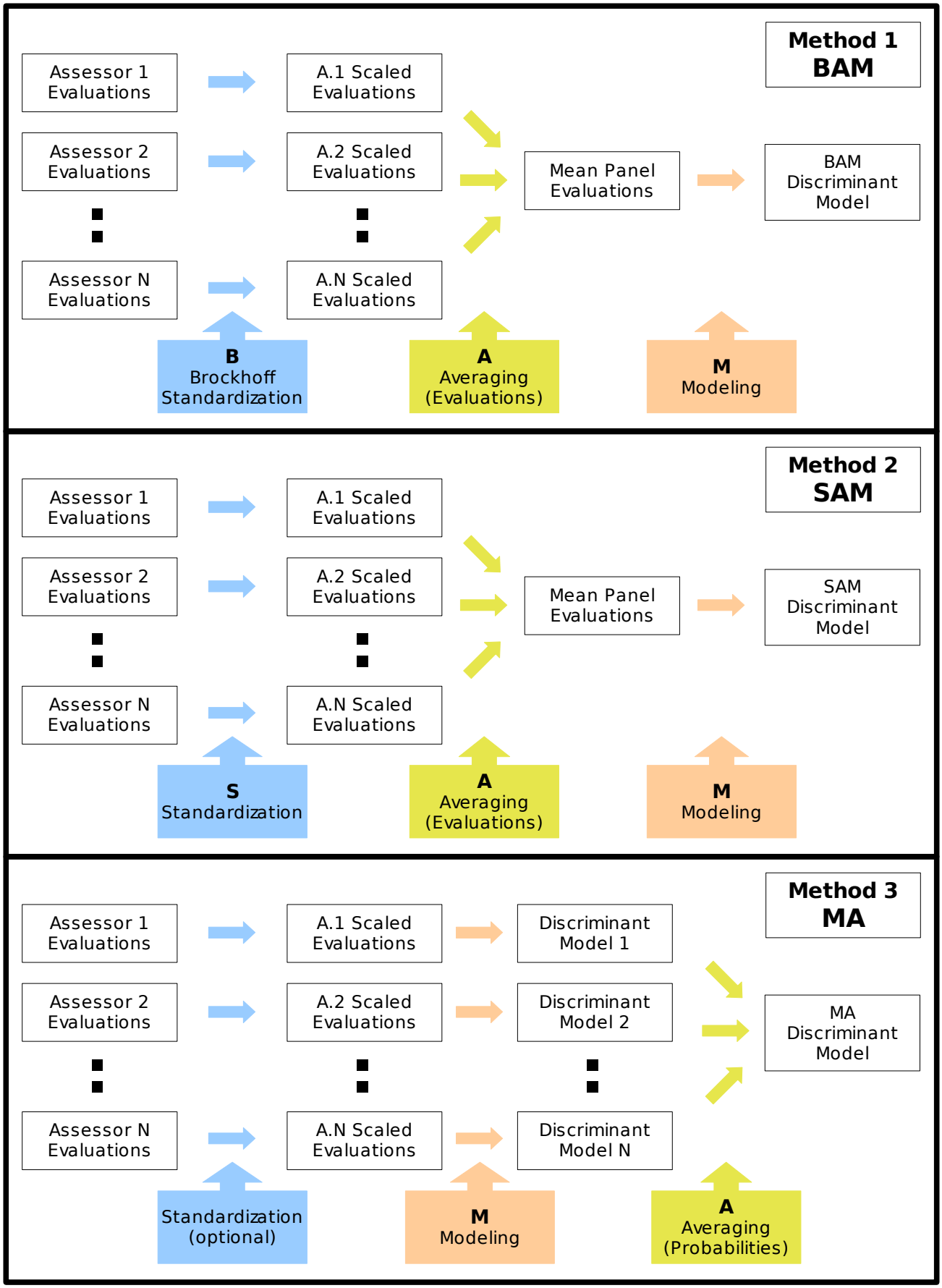
Method	Selection	Grana	Nostrani
MA	Random	0,22	0,27
	ANOVA	0,21	0,26
SAM	Random	0,36	0,32
	ANOVA	0,30	0,26

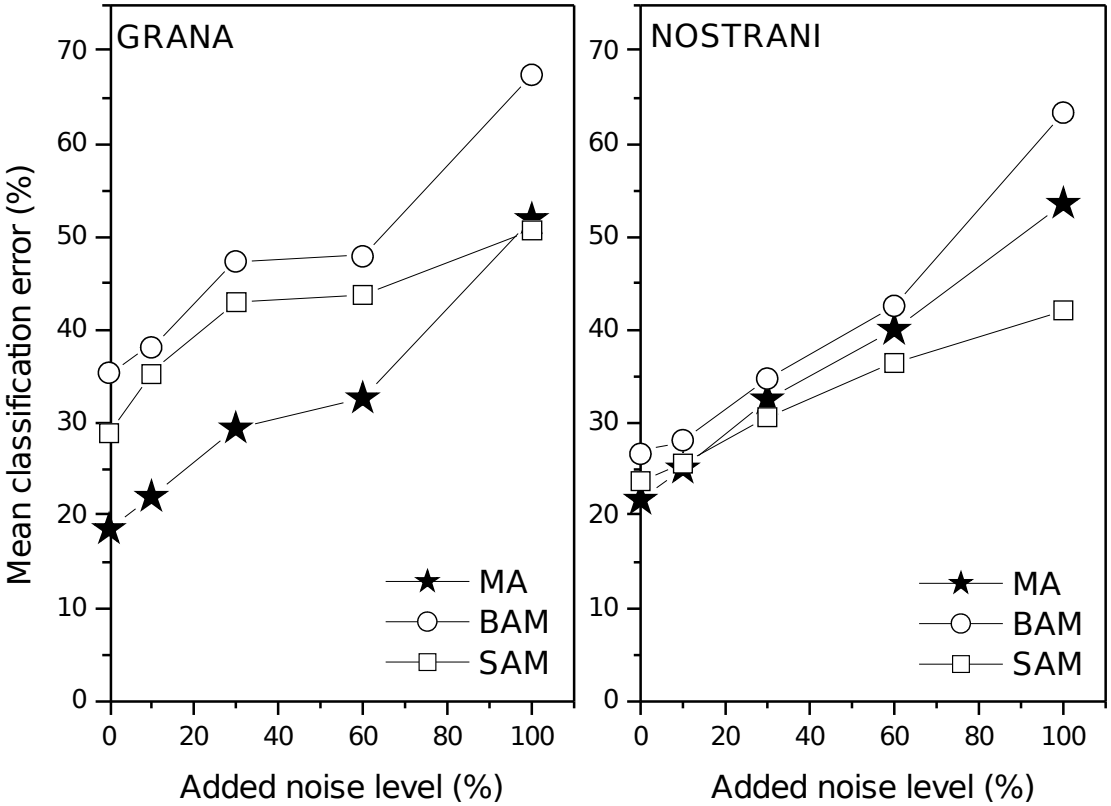
Table 5: Discriminant results using subsets of 15 attributes selected in different ways to fit the classifiers. Results are mean discrimination errors over the 3 classifiers (RF, PDA and dPLS) and 100 test sets.

Figure captions:

Fig 1: Summary of the three methods for the combination of assessors' evaluations.

Fig 2: Effect of noise addition. Left panel: “Grana” dataset. Right panel: “Nostrani” dataset. Results are mean values over the 3 different discriminant methods for each noise level and averaging strategy.


Method 1
BAM
Assessor 1 Evaluations
Assessor 2 Evaluations
Assessor N Evaluations
A.1 Scaled Evaluations
A.2 Scaled Evaluations
A.N Scaled Evaluations
Mean Panel Evaluations
BAM Discriminant Model
B Brockhoff Standardization
A Averaging (Evaluations)
M Modeling
Method 2
SAM
Assessor 1 Evaluations
Assessor 2 Evaluations
Assessor N Evaluations
A.1 Scaled Evaluations
A.2 Scaled Evaluations
A.N Scaled Evaluations
Mean Panel Evaluations
SAM Discriminant Model
S Standardization
A Averaging (Evaluations)
M Modeling
Method 3
MA
Assessor 1 Evaluations
Assessor 2 Evaluations
Assessor N Evaluations
A.1 Scaled Evaluations
A.2 Scaled Evaluations
A.N Scaled Evaluations
Discriminant Model 1
Discriminant Model 2
Discriminant Model N
MA Discriminant Model
Standardization (optional)
M Modeling
A Averaging (Probabilities)


GRANA
NOSTRANI
Mean classification error (%)
Added noise level (%)
Added noise level (%)
MA
BAM
SAM
MA
BAM
SAM
0
10
20
30
40
50
60
70
0
20
40
60
80
100