JOURNAL OF AGRICULTURAL AND FOOD CHEMISTRY

ARTICLE

pubs.acs.org/JAFC

Nuclear Magnetic Resonance Profiling of Wine Blends

Giovanna Imparato,[†] Elvio Di Paolo,[†] Angela Braca,[†] and Raffaele Lamanna*,[‡]

[†]Co.T.Ir., ss 16 nord 240, 66054 Vasto (Ch), Italy

[‡]ENEA, Centro Ricerche Trisaia, ss 106 Jonica Km 419.5, 75026 Rotondella (MT), Italy

S *Supporting Information*

ABSTRACT: Nuclear magnetic resonance (NMR) profiling is used for characterization of monocultivar binary wine mixtures. Classification and quantification of the relative amount of wine in the mixture are made in two steps. First, each sample is classified as a mixture of a determined type by solving the appropriate classification problem using NMR profiles. The relative amount of the two corresponding monovarietal wines is then evaluated by multilinear regression of a selected set of NMR variables. Linear discriminant analysis (LDA), used in the classification step, gives a very good separation among the different mixture classes. On the other hand, a single layer artificial neural network, used to solve the multilinear problem, gives the relative amount of wine type in the mixture with a precision of about 10%.

KEYWORDS: Food origin, NMR profiling, wine blends

INTRODUCTION

The identification of the origin of food materials is of great importance for both producers and consumers. Among the several approaches for the determination of food origin, molecular profiling is the most diffused. In particular, nuclear magnetic resonance (NMR) profiling has been successfully applied to the identification of botanical, zoological, and geographical origin of different foods.[1−14] NMR is a very suitable technique for the characterization of complex systems such as foodstuffs, because it allows one to determine simultaneously a high number of compounds. Actually, NMR profiles provide a quite exhaustive representation of the chemical composition of the sample without the need of extensive manipulations. However, the correlation between metabolic composition and food origin cannot be easily established, due to the natural variability in the chemical composition of the ingredients and to the various manipulation processes used in food preparation. This issue is usually addressed by using multivariate statistical techniques, which permit to extract information related to the food origin from the chemical noise.[4,15] Moreover, in some cases, foods of different origin are mixed together, making the identification process much more demanding.

Mixing of different grape cultivars is a crucial process in wine production. In fact, the flavor and the identity of great quality wines are determined by the variety pattern of grapes used. Variety blends are usually made at must level, but also, mixtures of monovarietal wines can be found.

The analysis of wine composition, in terms of grapes variety pattern, has different aspects. First, a wine sample has to be identified as a monovarietal wine or as a mixture. Second, the blend type has to be recognized according to the cultivars composing the mixture. Finally, the relative amount of monovarietal components has to be evaluated with an appropriate regression process.

The analysis of wine composition is, in general, a very difficult task if the type and the number of the varieties used in the blending process are unknown. However, in the case of binary mixtures, the problem can be successfully addressed by molecular profiling and suitable pattern recognition and regression approaches.

In this work, NMR was used to detect molecular profiles of binary mixtures of monovarietal Italian wines. In particular, blends having Montepulciano (Mont) monovarietal wines as base were created by successive additions of Merlot (Merl), Cabernet (Cab), and Sangiovese (Sang) wines. The obtained NMR profiles were used in a pattern recognition algorithm for the identification of the blend type and successively as inputs in a regression algorithm for the evaluation of the relative amount of each variety component. In particular, linear discriminant analysis (LDA) and an artificial neural single layer network (ANN) with linear activation function were used to identify the mixture type and the percentage of added wine in the Montepulciano base. The ANN allows the correct quantification of each wine component in the mixture with about 10% reliability.

MATERIALS AND METHODS

Sample Preparation. Eight red wines (five Montepulciano d'Abruzzo, one Sangiovese, one Cabernet, and one Merlot), made in 2007, were provided by Crivea (Miglianico Ch) research center. All Montepulciano d'Abruzzo wines were added with Sangiovese, Cabernet, and Merlot in the following percentages: 10, 15, 20, 25, 35, 50, and 70%. In a second experiment, three Montepulciano d'Abruzzo and three Sangiovese wines, produced by the same Crivea research center in 2009, were mixed in the following amonts: 20, 40, 60, and 80%. All samples were stored at −21 °C until the moment of the analysis. Defrosted wines (0.9 mL) were placed into a NMR tube and added with 0.1 mL of D_2O (deuterium oxide) and 10 μL of a 0.06 M DSS (2,2-dimethyl-2-silapentane-5-sulfonic acid) solution. The sample pH was adjusted to

Received: February 10, 2011
Accepted: March 23, 2011
Revised: March 18, 2011
Published: March 23, 2011

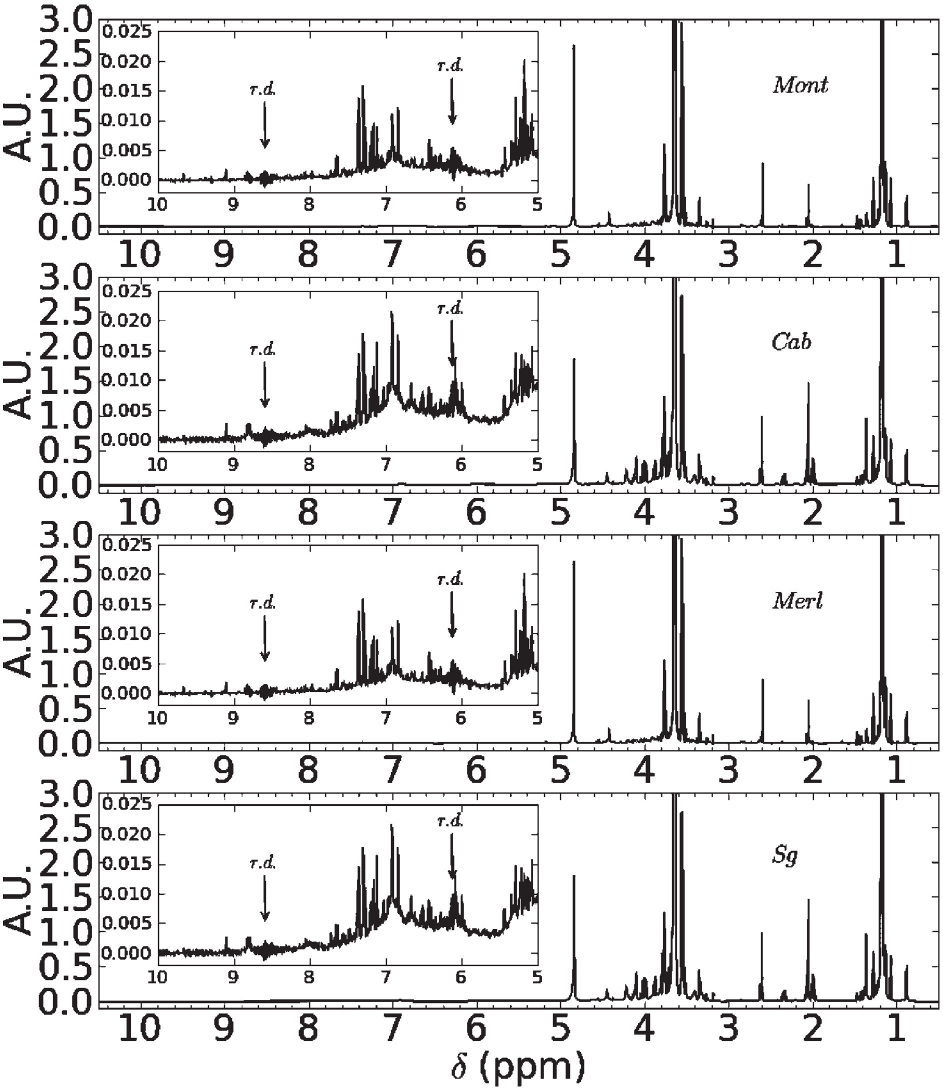

Figure 1. ^{1}H NMR spectra of pure wines used in the blends. The radiation damping artifacts are labeled by r.d. The intensity of NMR signal is expressed in arbitrary units (A.U.).

4.00 ± 0.05 by addition of suitable small amounts of a 1 M D_2O phosphate buffer at pD = 6.5.

NMR. Proton NMR spectra were recorded on a Bruker 600 Avance spectrometer operating at 600.13 MHz at 298 K. The spectra were collected with a 90° pulse of 9.45 s, relaxation delay of 1.5 s, and 256 scans. The strong water signal were suppressed by presaturation during the relaxation delay. The time domain data (FID) were Fourier transformed without apodization and were carefully phase and baseline corrected (Topspin1.3, Bruker). Each spectrum was divided in segments containing NMR signals only, by applying, on the average spectrum, a running windowing median algorithm for baseline/peak discrimination.[16] In each segment, the NMR signal was reduced in buckets[17] of equal width (0.0122 ppm) except at the segment terminal ends, where a reduced width was allowed to avoid unnecessary inclusion of baseline points. For the same reason, when the length of a segment was shorter than 0.0122 ppm, the bucket width was adapted to the segment length. To avoid statistical distortions, the suppressed water signal region was always excluded from the bucket spectrum. In addition, the citric acid region, which presents a strong concentration dependence of line positions, and ethanol resonances were also excluded from the bucket spectrum.

Statistical Methods. The problem of determining the mixture composition in terms of monocultivar wine components has to be addressed in two steps. The first step is the recognition of the sample as a mixture and the identification of its varietal components. This is a typical classification problem. The second step is the determination of the relative amount of each varietal component in the mixture, which is a typical regression problem. Both regression and classification problems can be seen as a particular case of function approximation.[18] In the classification problem, the probability of membership is approximated by a suitable discriminant function, which provides a discrete variable associated to the sample belonging class. In the regression problem, the relationship between the input and the output variables is approximated by a regression function according to the experimental data. In this case, the outputs are continuous variables. The simplest choice of the discriminant function is a linear combination of the input variables in which the parameters of the model are represented by the coefficients of the linear combination. LDA belongs to this class of linear classification models, and in the present work, it was used in the classification step of the wine mixtures.

In regression problems, the model function depends on the relationship between dependent and independent variables. In case this relationship

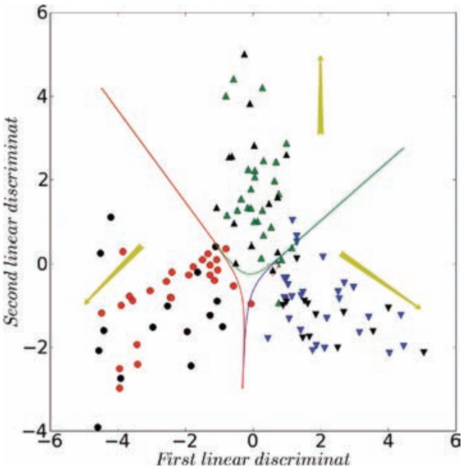

Figure 2. LDA of different mixtures: Cab-Mont (red circles), Merl-Mont (blue down triangles), and Sang-Mont (green up triangles). An approximation of the separation boundaries for each class is reported as a solid line of color corresponding to that of class symbols.

is not explicitly identified, an artificial neural network may be used to approximate the regression function. However, in the case of wine mixtures, the relationship between the metabolite concentrations and the relative amount of each wine component in the mixture is expected to be linear. This multilinear regression problem is anyway equivalent to an artificial single layer neural network with a linear activation function, which is represented by:

$$y_k = g\left(\sum_j w_{jk} x_j\right) \quad (1)$$

where x is the input vector, y is the output vector, and g is the activation function. In the NMR spectrum, the variable x_j is represented by a single bucket, while y represents the relative percentage of monocultivar wines in the mixture.

Through the bucketing procedure, the whole spectrum was reduced to 409 variables, which were still too many as compared to the number of samples to avoid overfitting phenomena. To further reduce the number of variables, retaining only those significant for individuating the origin of the wine samples, the analysis of variance (ANOVA) was used in combination with LDA to maximize the ability of LDA in predicting unknown samples.[19−22]

For this purpose, each set of wine samples was divided into a training and a validation set by including in the last set one-third of randomly selected samples. By using the whole set of samples, a progressive number of variables were selected according to their F of Fisher, the variables with the highest F were retained. The selected variables were used to build a linear discriminant model for the training set. The obtained model was then used to make prediction on the corresponding validation set, and the number of correctly predicted samples was determined as a function of the number of variables included in the model. The best model order, that is, the optimal number of variables to be used in the pattern recognition algorithm, was then chosen as the number of variables that give the maximum of correctly predicted samples in the validation set relative to the specific classification problem. This procedure of model selection gives the best compromise between the complexity of the model and its final predictive ability.[22]

RESULTS AND DISCUSSION

Figure 1 shows the proton NMR spectra of the four mono-varietal wines used as components for the wine mixtures. The most intense signals are associated with ethanol resonances, while the water signal is quite low due to presaturation signal suppression. Small artifacts appear in the spectra at a periodic distance from ethanol resonances equal to the frequency difference between the triplet and the quadruplet of ethanol. These harmonic peaks are probably due to radiation damping[23] and actually progressively disappears after sample dilution. In any case, when present, these signals were removed during the bucketing procedure. The ethanol content of a wine is strongly related to several factors such us grape ripening, sugar content, seasonality, etc. and then is not suitable for discriminating between wines made by different grape varieties. For this reason, its resonances were removed from the set of signals used for cultivar discrimination. Multiple solvent signal suppression also may be a solution to remove ethanol signals and to attenuate radiation damping effects.

Let us consider as a single set all of the samples belonging to a particular binary mixture independently of the relative amount of single cultivar components. In such a way, three sets are created, respectively, for Cab-Mont, Sang-Mont, and Merl-Mont binary mixtures. To each class actually belongs all of the mixtures with a relative concentration ranging from 10 to 100% of the secondary component (Cab, Merl, and Sang).

Figure 2 shows the LDA on these three sets in which a validation subset was created by randomly choosing one-third of the total number of samples. In the figure, the training set samples are plotted with colored symbols, while the validated samples are shown in black. The samples in the validation set, which were attributed to the wrong class, are surrounded by a bigger symbol representing their real belonging class. An approximation of the separation boundaries is reported in the figure for each class.[24] The arrows indicate the direction of increasing amount of the secondary wine in the Montepulciano-based mixtures. Actually, the samples close to the figure center are those with higher similarity due to the high percentage of base wine. The analysis was performed on the 27 NMR buckets having the highest value of F of Fisher relative to the three mixture classes. The number of variables included was determined according to the maximum prediction ability of the LDA model on the validation set. The prediction ability was estimated in the 95% of successfully predicted samples during validation procedure. K-fold cross-validation, with $k = 10$, confirms a success percentage of 95 ± 7%.

Once the nature of the mixture was established, the relative amount of wines present can be evaluated by solving the appropriate regression problem. The results of an artificial single layer neural network with linear activation function are reported in Figures 3, 4, and 5, respectively, for Cab-Mont, Merl-Mont, and Sang-Mont blends.[25] In the x-axis, the real amount of the second component of the mixture is reported, while the y-axis represents the ANN average-predicted values. Error bars show the standard deviation of the leave-one-out validation test. As shown by the linear regression of the ANN data (solid line), the correspondence between the actual mixture content and the ANN predicted value is very high (correlation coefficient $R > 0.99$).

Figure 6 shows the prediction of the amount of added wine in a Montepulciano base when all of the samples (Cabernet, Merlot, and Sangiovese) were considered. Actually, ANN was trained by considering each wine addition, in the base wine, independently of the cultivar. This picture shows that the percentage of base wine can be determined independently of the nature of contaminating wines, suggesting the possibility that also more complex wine mixtures could be quantified by NMR profiling.

For each particular mixture, the ANN was trained using as inputs 10 buckets with the highest value of F of Fisher

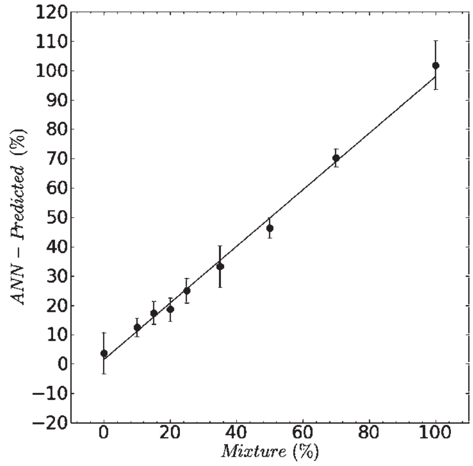

Figure 3. ANN prediction of the relative amount of Cabernet in Montepulciano wines vs the actual composition of the blend. The solid line represents the linear regression of the data by $y = \alpha + \beta x$ with $\alpha = 1.5 \pm 1$, and $\beta = 0.96 \pm 0.03$. The regression correlation coefficient is $R = 0.997$.

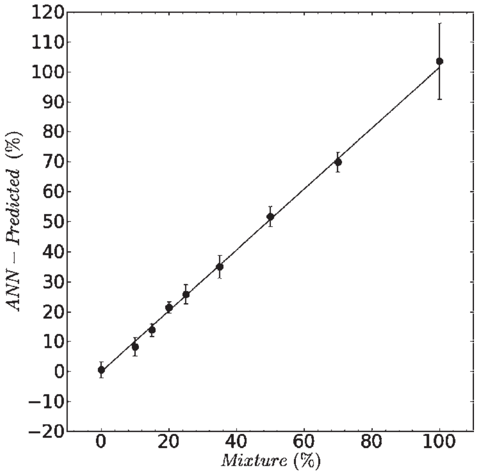

Figure 4. ANN prediction of the relative amount of Merlot in Montepulciano wines vs the actual composition of the blend. The solid line represents the linear regression of the data by $y = \alpha + \beta x$ with $\alpha = -0.1 \pm 0.6$, and $\beta = 1.02 \pm 0.02$. The regression correlation coefficient is $R = 0.999$.

determined by ANOVA of the spectra relative to the corresponding mixture only. The chemical shifts of the selected variables are reported on the correspondent spectrum in Figure 7. Circles represent the selected variables used in the classification problem, while triangles correspond to the variables used in the regression step. Each mixture has a proper set of discriminating variables with very different discriminant power. The lowest F values are found in the Merl−Mont mixtures, indicating a great similarity between Merlot and Montepulciano wines. On the other hand, the highest F value (600) is displayed by the signal at 2.21 ppm in Sang−Mont mixtures, which is tentatively assigned to the methyl group of acetamide. However, even if this signal is removed from the bucket spectrum, a good prediction performance is achieved by ANN (see the Supporting Information).

The results until now displayed refer to mixtures in which only one sample of contaminating wine is added to the five wines used as base. In a second experiment, the effects of the natural

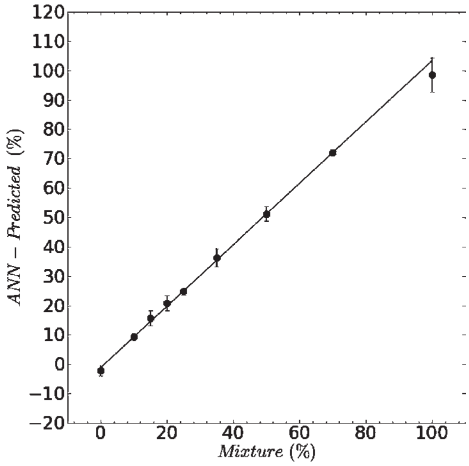

Figure 5. ANN prediction of the relative amount of Sangiovese in Montepulciano wines vs the actual composition of the blend. The solid line represents the linear regression of the data by $y = \alpha + \beta x$ with $\alpha = -1.2 \pm 0.2$, and $\beta = 1.045 \pm 0.007$. The regression correlation coefficient is $R = 0.999$.

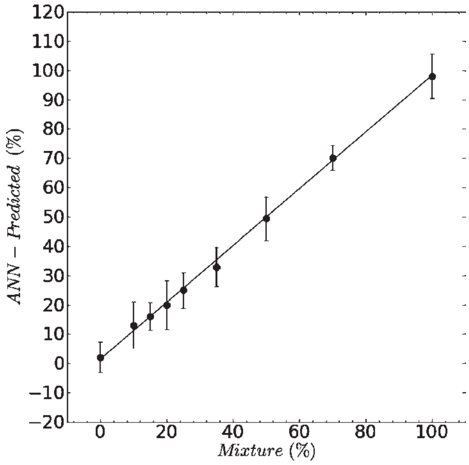

Figure 6. ANN prediction of the relative amount of Cabernet, Merlot, and Sangiovese in Montepulciano wines vs the actual composition of the blend. The solid line represents the linear regression of the data by $y = \alpha + \beta x$ with $\alpha = 1.5 \pm 0.6$, and $\beta = 0.97 \pm 0.01$. The regression correlation coefficient is $R = 0.999$.

variability of the second wine component were analyzed. In Figure 8, the ANN quantification of a mixture of three Montepuciano and three Sangiovese wines is reported. The ability to predict the unknown amount of Sangiovese wine in the mixture is still good even if larger error bars are present. The increment of the error bars in Figure 8, as compared to that of other ANN results, is actually due to the increase of mixture variability related to the inclusion of three contaminating wines but also to the natural variability of the samples related to the cultivation year. In fact, despite in Figure 6 contaminating wines of different cultivars were considered, the validation procedure gives error bars smaller than that of Figure 8. Actually, because the wine used in the second experiment was produced 2 years after the wines of the first experiment, some care has to be used in comparing the two data sets. In fact, the chemical compositions of grapes and wines strongly depend on the cultivation season.

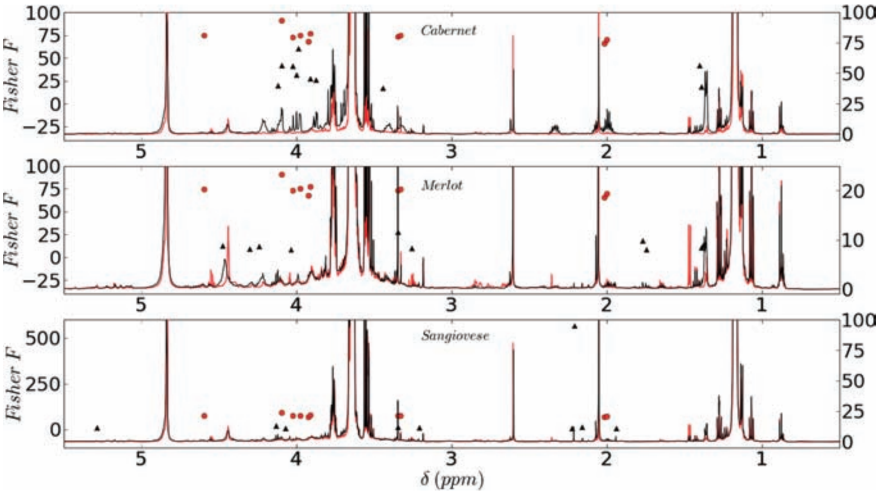

Figure 7. *F* of Fisher of selected variables (left *y*-axes) together with reference NMR spectra (right *y*-axes) for each second blend component. Black triangles are the selected variables used as input in ANN algorithm for each mixture quantification. Red circles represent the selected variable used in LDA pattern recognition of mixture type. The spectra of Montepulciano base (red line) and of secondary blend component (black line) are shown for comparison.

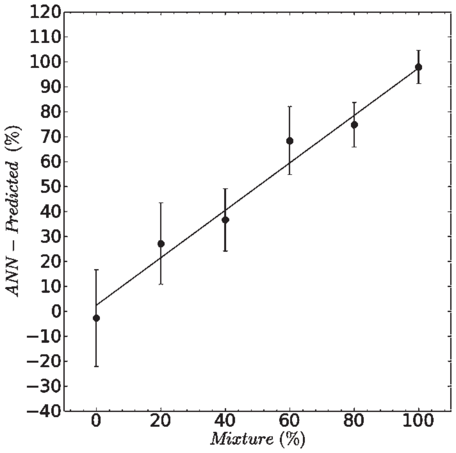

Figure 8. ANN prediction of the relative amount of three Sangiovese in three Montepulciano wines vs the actual composition of the blend in the second experiment. The solid line represents the linear regression of the data by $y = \alpha + \beta x$ with $\alpha = 2.3 \pm 4.3$, and $\beta = 0.95 \pm 0.05$. The regression correlation coefficient is $R = 0.988$.

In any case, whatever is the natural variability for the year of interest, the method proposed is able to evaluate the unknown amount of added wine in a binary mixture, estimating also the reliability of the prediction by suitable error bar calculations. In addition, the reliability of the ANN prediction might be increased by training the network by a suitably high number of know samples. Because the performance of the method depends on the year of wine production, new predictive ANN models have to be constructed for each production season. Also, the spectral components, which are responsible for the discrimination success, are different in each production season due to variability of the production conditions.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information. Data analysis and figures of ANN prediction and *F* of Fisher of selected variables. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: raffaele.lamanna@enea.it.

■ ACKNOWLEDGMENT

The helpful collaboration with the CRIVEA (Centro per la ricerca vitivinicola ed enologica in Abruzzo) is acknowledged.

■ REFERENCES

(1) Brescia, M. A.; Kosir, I. J.; Caldarola, V.; Kidric, J.; Sacco, A. Chemometric classification of Apulian and Slovenian wines using ^{1}H NMR and ICP-OES together with HPICE data. *J. Agric. Food Chem.* **2003**, *51*, 21–26.

(2) Nilsson, M.; Duarte, I. F.; Almeida, C.; Delgadillo, I.; Goodfellow, B. J.; Gil, A. M.; Morris, G. A. High-resolution NMR and diffusion-ordered spectroscopy of port wine. *J. Agric. Food Chem.* **2004**, *52*, 3736–3743.

(3) Ogrinc, N.; Kosir, I. J.; Spangenberg, J. E.; Kidric, J. The application of NMR and MS methods for detection of adulteration of wine, fruit juices, and olive oil. A review. *Anal. Bioanal. Chem.* **2003**, *376*, 424–430.

(4) Pereira, G.; Gaudillere, J.; Van Leeuwen, C.; Hilbert, G.; Lavialle, O.; Maucourt, M.; Deborde, C.; Moing, A.; Rolin, D. 1H NMR and chemometrics to characterize mature grape berries in four wine-growing areas in Bordeaux, France. *J. Agric. Food Chem.* **2005**, *53*, 6382–6389.

(5) Du, Y.-Y.; Bai, G.-Y.; Zhang, X.; Liu, M.-L. Classification of wines based on combination of 1H NMR Spectroscopy and principal component analysis. *Chin. J. Chem.* **2007**, *25*, 930–936.

(6) Consonni, R.; Cagliani, L. R. Geographical characterization of polyfloral and acacia honeys by nuclear magnetic resonance and chemometrics. *J. Agric. Food Chem.* **2008**, *56*, 6873–6880.

(7) Mannina, L.; Segre, A. High resolution nuclear magnetic resonance: From chemical structure to food authenticity. *Grasa Aceites* **2002**, *53*, 22–33.

(8) Mannina, L.; Calcagni, C.; Rossi, E.; Segre, A. Review about olive oil characterization using high-field nuclear magnetic resonance. *Ann. Chim.* **2003**, *93*, 97–103.

(9) Brescia, M. A.; Di Martino, G.; Fares, C.; Di Fonzo, N.; Platani, C.; Ghelli, S.; Reniero, F.; Sacco, A. Characterization of Italian durum wheat

semolina by means of chemical analytical and spectroscopic determinations. *Cereal Chem.* **2002**, *79*, 238–242.

(10) Brescia, M. A.; Sgaramella, A.; Ghelli, S.; Sacco, A. 1H HR-MAS NMR and isotopic investigation of bread and flour samples produced in southern Italy. *J. Sci. Food Agric.* **2003**, *83*, 1463–1468.

(11) Capron, X.; Smeyers-Verbeke, J.; Massart, D. L. Multivariate determination of the geographical origin of wines from four different countries. *Food Chem.* **2007**, *101*, 1585–1597.

(12) Donarski, J. A.; Jones, S. A.; Charlton, A. J. Application of cryoprobe 1H nuclear magnetic resonance spectroscopy and multivariate analysis for the verification of Corsican honey. *J. Agric. Food Chem.* **2008**, *56*, 5451–5456.

(13) Gil, A. M.; Duarte, I. F.; Godejohann, M.; Braumann, U.; Maraschin, M.; Spraul, M. Characterization of the aromatic composition of some liquid foods by nuclear magnetic resonance spectrometry and liquid chromatography with nuclear magnetic resonance and mass spectrometric detection. *Anal. Chim. Acta* **2003**, *488*, 35–51.

(14) Amato, M. E.; Ansanelli, G.; Fisichella, S.; Lamanna, R.; Scarlata, G.; Sobolev, A. P.; Segre, A. Wheat flour enzymatic amylolysis monitored by in situ 1H NMR spectroscopy. *J. Agric. Food Chem.* **2004**, *52*, 823–831.

(15) Larsen, F. H.; van den Berg, F.; Engelsen, S. B. An exploratory chemometric study of 1H NMR spectra of table wines. *J. Chemom.* **2006**, *20*, 198–208.

(16) Torgrip, R. J. O.; Aberg, M.; Karlberg, B.; Jacobsson, S. P. Peak alignment using reduced set mapping. *J. Chemom.* **2003**, *17*, 573–582.

(17) Lamanna, R. Bucketing is performed by home made tnmr 1.0 software available from the author (R.L.) under GNU license.

(18) Bishop, C. M. *Neural Networks for Pattern Recognition*; Oxford University Press: New York, 2000.

(19) Berrueta, L. A.; Alonso-Salces, R. M.; Heberger, K. Supervised pattern recognition in food analysis. *J. Chromatogr. A* **2007**, *1158*, 196–214.

(20) Guyon, I.; Elisseef, A. An Introduction to Variable and Feature Selection. *J. Machine Learn. Res.* **2003**, *3*, 1157–1182.

(21) Marini, F.; Magri, A.; Balestrieri, E.; Fabretti, F.; Marini, D. Supervised pattern recognition applied to the discrimination of the floral origin of six types of Italian honey samples. *Anal. Chim. Acta* **2004**, *515*, 117–125.

(22) Lamanna, R.; Cattivelli, L.; Miglietta, M. L.; Troccoli, A. Geographical origin of durum wheat studied by ^{1}H-NMR profiling. *Magn. Reson. Chem.* **2010**, *49*, 1–5.

(23) Peng, L.; Cai, S. H.; Fu, R. Q.; Ye, C. H.; Chen, Z. Harmonic peaks in 1D NMR spectra induced by radiation damping fields. *Chem. Phys. Lett.* **2009**, *479*, 165–170.

(24) Venables, W. N.; Ripley, B. D. *Modern Applied Statistics with S*; Springer: New York, 2002.

(25) Lamanna, R. nnet R package is used for this calculation inside the home made software statnmrqt4 0.1 available from the author (R.L.) under GNU license.