Determination of the Geographical Origin of Hazelnuts (*Corylus avellana* L.) by Near-Infrared Spectroscopy (NIR) and a Low-Level Fusion with Nuclear Magnetic Resonance (NMR)

Navid Shakiba[a,b], Annika Gerdes[a,b], Nathalie Holz[a], Sören Wenck[b], René Bachmann[c], Tobias Schneider[a], Stephan Seifert[b], Markus Fischer[b], Thomas Hackl[a,b,]*

[a]Institute of Organic Chemistry, University of Hamburg, Martin-Luther-King-Platz 6, 20146 Hamburg, Germany,

[b]HAMBURG SCHOOL OF FOOD SCIENCE - Institute of Food Chemistry, University of Hamburg, Grindelallee 117, 20146 Hamburg, Germany

[c]Landeslabor Schleswig-Holstein, Max-Eyth-Straße 5, 24537 Neumünster, Germany

*Corresponding author: Tel.: +49-40 42838-2804; E-Mail: Thomas.Hackl@chemie.uni-hamburg.de

ABSTRACT

Fourier-transform near-infrared (FT-NIR) spectroscopy was used to determine the geographical origin of 233 hazelnut samples of various varieties from five different countries (Germany, France, Georgia, Italy, Turkey). The experimental determination of the geographical origin of hazelnuts is important, because there are usually large price differences between the producer countries and thus a risk of food fraud that should not be underestimated. The present work is a feasibility study using a low-cost method, as high-field NMR and UPLC-QTOF-MS have already been used for this question. Sample sets were split with repeated nested cross validation and an ensemble of discriminant classifiers with random subspaces was used to build the classification models. By using a preprocessing strategy consisting of multiplicative scatter correction, bucketing and the mean averaging of five measured spectra per sample, a test accuracy of 90.6 ± 3.9% was achieved, which rivals results obtained with much more expensive infrastructure. The application of the feature selection approach surrogate minimal depth showed that the successful classification is mainly caused by protein signals. In addition, a low-level data fusion of the NIR and NMR data was performed to assess how well the two methods complement each other. The data fusion was compared to a complementary approach, where the classification results based on the individual NIR and NMR models were jointly examined. The data fusion performed better than the individual methods with a test accuracy of 96.6 ± 2.8%. A comparison of the outliers in all classification models shows conspicuities in always the same samples, indicating that robust classification models are obtained.

KEYWORDS

Geographical Origin, NIR, NMR, Data Fusion, Hazelnut, Feature selection

1. INTRODUCTION

Hazelnuts (*Corylus avellana* L.) are a globally traded food with a production volume of approximately 1,125,000 t in 2019.[1] Turkey is the main producing country with a volume of 776,000 t in 2019, representing 69% of the world production. Other major producing countries are Italy, Azerbaijan, the USA, Chile, China and Georgia, although producer prices vary widely in some cases. For example, the price for a ton of hazelnuts from Georgia in 2019 was only 1550 USD/t, but in Italy it was 3600 USD/t, as these hazelnuts are considered to be of particularly high quality. Such a wide price range is bound to provide a financial incentive for food fraud, where hazelnuts from a low-price producing country are falsely declared with a different origin to increase profits.

Bachmann *et al.* (2018) and Klockmann *et al.* (2016) explored the issue of determining the geographical origin of hazelnuts using high-resolution instrumentation, ^{1}H NMR spectroscopy and ultraperformance liquid chromatography quadrupole time-of-flight mass spectrometry (UPLC-QTOF-MS) in combination with chemometric evaluation strategies.[2,3] These studies showed that it is possible to distinguish the origin of hazelnuts using metabolomics approaches. However, these tools are quite expensive and require a high level of scientific expertise, which is limiting, especially for smaller laboratories and small and medium-sized food companies.

Fourier-transform near-infrared spectroscopy offers a cost-effective way to determine the geographical origin of food, as has already been shown with various foods such as pistachio, wheat, almonds and walnuts.[4–7] In addition, NIR can be used for a wide range of food-related issues in the food sector, e.g. quality control of olive oil, determination of storage time of pork and identifying oxidation of vegetable oils.[8–10] Other advantages of NIR spectroscopy are the absence of hazardous chemicals, the non-destructive nature, a fast measurement time and the fact that no extraction is required. Two Italian research groups have already used NIR spectroscopy to distinguish 'Nocciola Romana', which carries a Protected Designation of Origin (PDO), from other hazelnuts; however, a holistic comparison of several countries of origin has not yet taken place.[11,12] Both studies examined whole, shelled hazelnuts in order to develop a non-destructive rapid method. Based on a comparison of different preparation techniques for NIR measurement to determine the geographical origin of almonds, we decided to analyze the samples after homogenization and freeze-drying, as we expected this approach to provide a higher information content and a better representation of the sample populations.[13] One aim of this study is therefore to investigate the ability of NIR spectroscopy to determine the geographical origin of hazelnut samples, as there is a need for such a low-cost analytical method. This could be used in industry for incoming goods inspection. To establish such a method, we compared various preprocessing strategies and classification approaches. In addition, surrogate minimal depth (SMD) was applied, a random forest based approach for feature selection and relation analysis that has already been used to study other vibrational spectroscopic data.[14–16]

The newly acquired NIR and the already existing NMR data were selected for low-level fusion due to their one-dimensionality and potentially complementary nature. Low-level fusion involves concatenation of the datasets with or without prior preprocessing methods.[17] In the case of hazelnuts as a matrix, the NIR captures mainly non-specific information on groups of substances with high concentration, e. g. lipids, carbohydrates and proteins, while the ^{1}H NMR measurement of the polar extract provides more specific information on substances such as organic acids, amino acids and specific carbohydrates. To the best of our knowledge, this is the first publication on the experimental determination of the geographical origin of food by combining NIR and high-field NMR data in a multiclass model. The aim of the low-level data fusion approach of the NIR and NMR data is to obtain a statistical model that is better than the individual methods.

2. Materials and Methods

2.1. Hazelnut Samples

In a previous study by our group 262 raw hazelnut samples were used for the determination of the geographical origin by means of ^{1}H NMR.[2] Authentic reference material was provided by partners, distributors and suppliers. Of these, only 233 samples could be used for NIR analysis, as the sample material for some samples has already been used up. Samples from a total of five countries were analyzed, with several samples coming from economically important growing regions. In this study, 27 German samples, 116 French samples, 15 Georgian samples, 37 Italian samples and 38 Turkish samples were used. The same samples were also taken for the ^{1}H NMR analysis and the low-level data fusion. More detailed information on origin and variety are given in the Supporting Information (Table S1).

2.2. Sample Treatment

All hazelnut samples were treated according to Bachmann et al.[2] Hazelnut samples were frozen in liquid nitrogen before they were homogenized with a Grindomix GM 300 knife mill and dry ice was added. After evaporation of the dry ice, the samples could be directly used for NMR analysis. To prepare the samples for NIR measurement, the homogenized samples were then freeze-dried for 48 hours.

2.3. NIR spectroscopy

1.250 g (± 0.005 g) of the ground and freeze-dried hazelnut samples were thawed at 22 °C (± 2 °C) in closed glass vials (52.0 mm x 22 mm x 1.2 mm, Nipro Diagnostics Germany GmbH, Ratingen, Germany) preceding NIR measurement.

The NIR measurements were performed on a TANGO FT-NIR spectrometer (Bruker Optics, Bremen, Germany) equipped with an integrating sphere. Spectra were recorded in reflectance mode at room temperature (22 ± 2 °C), with the wavenumber range set to 11546-3949 cm^{-1} collecting 50 scans at a resolution of 2 cm^{-1}. Each sample was analyzed five times by shaking the lyophilisate in the glass vial between measurements.

2.4. NMR spectroscopy

The NMR spectra used for the data fusion were acquired by Bachmann *et al.* in an earlier study on a Bruker Avance III 400 MHz spectrometer (Bruker Biospin, Rheinstetten, Germany) operating at 400.13 MHz with the noesygppr1d pulse program.[2]

2.5. NIR spectra preprocessing

All preprocessing techniques were performed using MATLAB R2020b (The MathWorks Inc., Natick, MA, USA). Multiplicative scatter correction (MSC) was applied to ensure a good comparability between samples. MSC is a commonly used preprocessing step to normalize the data and remove artifacts from the samples by using the mean spectrum of the available data.[18] These artifacts are mostly due to differences in particle size of the powdered sample, which leads to non-uniform scattering effects.[19] After MSC, either no derivative, the first derivative or the second derivative was applied to the spectra. The approaches that used a derivative also utilized a Savitzky-Golay smoothing filter with a window size of 11 and a polynomial order of 2 to minimize the negative effects of a derivative on the signal-to-noise ratio.[18] Next, variable reduction was achieved by calculating the mean of five adjacent features into one bucket, leading to a reduction from 3720 variables (spectral range: 11538-3949 cm^{-1}) to 744 NIR-buckets. Finally, the mean or median of the five measured spectra per sample was determined. The classification results of the different preprocessing strategies were then compared.

2.6. NMR spectra preprocessing

The NMR data were processed with Topspin 3.2 (Bruker Biospins, Rheinstetten, Germany). A Fourier transformation with a line broadening factor of 0.3 was applied on the FIDs, then baseline corrected and phased. Integrals of signals and regions from the NMR spectra were determined manually in AMIX 3.9.14 (Bruker Biospins, Rheinstetten, Germany) as variable sized buckets and normalized to total intensity by scaling. A total of 222 NMR-buckets were defined for each sample. The mean and median of the triplicate measurement was determined.[2]

2.7. NIR-NMR low-level fusion

For the low-level fusion, the 744 NIR buckets of the best performing model were concatenated with the 222 NMR buckets, once mean and once median averaged.[17] Autoscaling was used as a scaling method for the fusion data.[20,21]

2.8. Multivariate data analysis

Multivariate data analysis was performed with MATLAB R2020b including the Classification Learner app (The MathWorks Inc., Natick, MA, USA). Samples were split into five equal parts with a stratified nested cross-validation stratified by geographic origin.[22] The internal validation was also split fivefold to avoid overfitting during model training. Repeated nested cross-validation (RNCV) was iterated five times to obtain an average result, resulting in 25 different corresponding training and test sets, with each sample being part of 20 training and 5 test sets, as different sample splits can lead to large differences in model accuracy. The Classification Learner app was used to determine which classifiers would be suitable for the NIR, NMR and low-level fusion data, and then the chosen classifier was used for automatic model training and subsequent validation. The ensemble of discriminant classifiers using random subspaces was the best performing method and was trained with 372 subspace dimensions and 30 learning cycles.[23] The test accuracies given are the mean of the test accuracies of all sample splits from the RNCV. The macro-F_1 score is calculated as the arithmetic mean of F_1 score of the five classes, formed from the harmonic mean of the class-wise precision and sensitivity. In addition, Fleiss' kappa was calculated to determine the degree of agreement of the classifiers in each model.[24]

2.9. Feature selection and relation analysis with surrogate minimal depth (SMD)

The software R in version 3.6.3 and the R package SurrogateMinimalDepth in version 0.2.0 (https://github.com/StephanSeifert/SurrogateMinimalDepth) were utilized for feature selection with the parameters ntree = 10000, mtry = 143, min.node.size = 1 and s = 149. In order to compensate for the class imbalance, case.weights were chosen accordingly meaning that samples from rare classes were sampled more frequently for training. Subsequently, the relation parameter mean adjusted agreement of the selected features was determined and depicted in a heatmap generated by the R package pheatmap in version 1.0.12. For the random forest classification, the R package ranger was applied with the above described parameters.[25] Since random forests provides an internal validation, no cross validation scheme had to be applied and all of the samples were utilized simultaneously.

3. RESULTS AND DISCUSSION

3.1. NIR-Spectroscopy

Hazelnuts are rich in fat (~61%), carbohydrates (~17%), protein (~15%) and have a water content of ~5%.[26] The fatty acid profile is dominated by oleic (72.8-83.5%), linoleic (7.6-16.6%) and palmitic (4.1-6.8%) acid and is similar to that of olive oil.[27] A model NIR spectrum of a hazelnut sample is shown in the Supporting Information (S2). The NIR spectrum shows strong similarities to those of other species of nuts, because of akin nutrient composition.[6,7] Due to the broad absorption and the overlapping of the signals of the individual substances of the complex matrix, no peaks in the spectra can be clearly assigned to specific metabolites. Instead, signals and regions in the spectra can usually be assigned to different molecular vibrations caused mainly by compound classes of macronutrients. The peak at 8550 cm^{-1} can be assigned to HC=CH (C-H second overtone) caused by unsaturated fatty acids. Other signals that can be associated with lipids are the second overtone of C-H (C-H, C-H_2, C-H_3) stretching vibrations between 8500-8000 cm^{-1}, the first overtone of C-H between 5900-5600 cm^{-1} and the combination bands of the methylenic CH_2 between 4500-4000 cm^{-1}.[11,12] The first overtone of N-H and O-H of proteins can be observed in the region between 7100-6100 cm^{-1}.[28] Another region related to proteins is between 4900-4600 cm^{-1}, caused by the combination band of peptide bonds.[28]

Principal component analysis (PCA) is arguably the most widely used unsupervised method for reducing the complexity of metabolomics data while preserving variance as much as possible and revealing underlying class information.[29] The limitations of PCA as an exploratory method are that underlying patterns cannot be uncovered if the intragroup variance of the sample groups is greater than the intergroup variance.[30] The advantages of PCA include an initial unbiased look at the data to examine the extent to which samples are similar within and outside their groups and to identify potential outliers.[31] Figure 1A shows the PCA scores plot of the unprocessed samples, where the first principal component (PC) contains 85.0% of the variance and 8.3% the PC 2. The plot shows a cluster of all samples with no outliers or clear group separation. Nevertheless, the different groups show similarities. The French samples are in the center-left of the plot, while the German samples are below and the Georgian samples are above. The Italian and Turkish samples mostly scatter from the center to the right side of the plot along the first principal component (PC1). The PCA scores plot of the preprocessed data is shown in Figure 1B with PC 1 accounting for 60.2% of the variance and PC 2 representing 23.7%. The plot shows a coherent cluster with no outliers, but with less clear spatial allocations of the different groups. As expected, PCA cannot identify separate groups with respect to the origin of the samples. Hence, supervised multivariate analysis was performed to determine the geographical origin of the hazelnut samples.

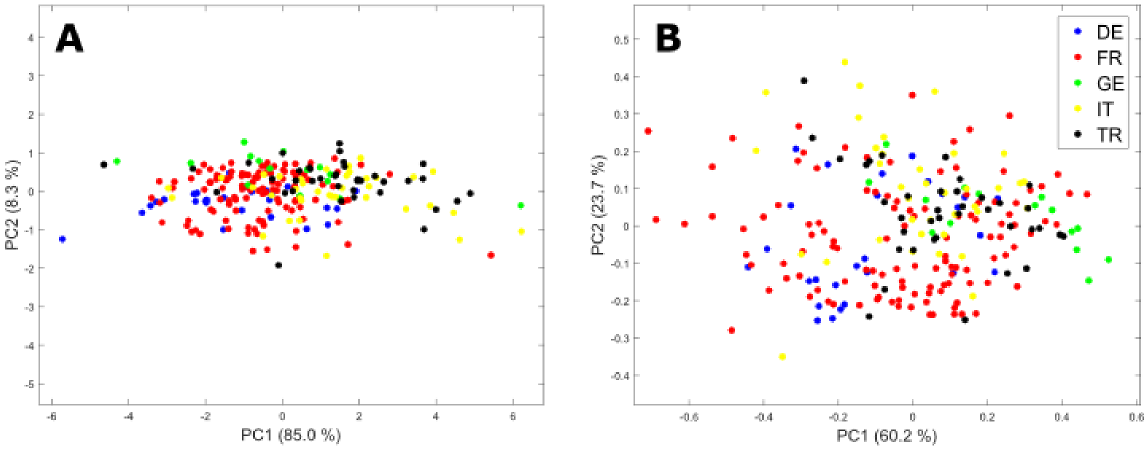


Figure 1. (A) PCA scores plot of unprocessed, mean averaged NIR spectra. **(B)** PCA scores plot of NIR spectra after MSC, bucketing and mean averaging.

Table 1. Test accuracy, macro-F1 score and Fleiss' kappa coefficient of the different preprocessing strategies of the NIR analysis attained by using an ensemble of discriminant classifiers using random subspaces.

Strategy	**Preprocessing**	**Test Accuracy**	**Macro-F_1**	**Fleiss' Kappa**
NIR-I	MSC – Mean	89.5 ± 4.3%	85.9%	87.4%
NIR-II	MSC – Median	84.2 ± 5.3%	80.1%	78.5%
NIR-III	MSC – Bucketing – Mean	90.6 ± 3.9%	88.1%	88.7%
NIR-IV	MSC – Bucketing – Median	81.3 ± 5.2%	76.7%	77.8%
NIR-V	MSC – 1. Derivative – Smoothing – Bucketing – Mean	76.6 ± 5.3%	70.5%	73.7%
NIR-VI	MSC – 1. Derivative – Smoothing – Bucketing – Median	75.1 ± 5.3%	68.5%	75.8%
NIR-VII	MSC – 2. Derivative – Smoothing – Bucketing – Mean	67.6 ± 4.8%	57.9%	67.7%
NIR-VIII	MSC – 2. Derivative – Smoothing – Bucketing – Median	68.6 ± 4.8%	56.9%	70.9%
NIR-IX	Cut – MSC – Bucketing – Mean	83.6 ± 3.8%	79.6%	77.7%
NIR-X	Cut – MSC – Bucketing – Median	75.2 ± 6.0%	70.6%	66.4%
NIR-SMD	MSC – Bucketing – Mean – SMD feature selection	86.4 ± 4.7%	82.1%	78.1%

For the application of supervised approaches different preprocessing strategies and different classifiers were applied and different parameters to assess their performance were utilized. The test accuracy is the most common measure for machine learning models. The macro-F_1 score provides information about the mean class-wise precision and sensitivity of a sample split.[32] Fleiss' kappa is a measure of inter-rater reliability for determining the homogeneity of the RNCV's ratings of the samples, regardless of whether they were correctly allocated or not.[33] The algorithm used for multivariate analysis was an ensemble of discriminant classifiers using random subspaces.[23] This algorithm showed the best results for strategies I-IV. However, strategies IV-VIII showed slightly better results with other classifiers (data not shown), but for clarity and comparability the same classifier was used for each strategy and the results are shown in Table 1. The mean spectra of preprocessing strategies I, V and VII are depicted in the Supporting Information (S3).

The adverse effects on the signal-to-noise ratio due to smoothing and the use of the derivative of the spectra is reflected in the relatively poor results of preprocessing strategies V-VIII. Strategies VII and VIII, which use the second derivative, have a test accuracy of 67.6 ± 4.8% and 68.6 ± 4.8%, respectively. Including the information from the confusion matrices shows an even worse picture. Macro-F_1 scores of 57.9% and 56.9% for strategies VII and VIII, respectively, show a more dramatic decline to their test accuracies compared to other preprocessing strategies. This is probably due to the fact that the French sample group has the highest number of samples and a large proportion of the samples from other countries are misclassified as French. The strategies with the first derivative already show significantly better results with a test accuracy of 76.6 ± 5.3% for strategy V and 75.1 ± 5.3% for strategy VI. The macro-F_1 scores are also much closer to test accuracies. Since additive effects are observed in the spectrum, the use of the first derivative is reasonable in theory. In the practice of this study, however, the negative effects of the derivative on the signal-to-noise ratio may have led to poorer predictive performance of the models. The NIR preprocessing strategies IX and X cut off the spectrum above the wavenumber of 9000 cm^{-1}. This is a common preprocessing step as this region is usually not very information-rich and contains mainly bands from the third and the fourth overtone vibrations.[28,34] Both strategies lead to a decrease in test accuracy of 7.0% and 6.1% compared to strategies III and IV, which contain features from the whole spectrum. This suggests that the bands in this region are relevant for the research question. Although all strategies forgoing the derivative show good model performance, the results of strategies I and III show the highest test accuracy of 89.5 ± 4.3% and 90.6 ± 3.9%, a macro-F_1 score of 85.9% and 88.1% and a Fleiss' kappa coefficient of 87.4% and 88.7%. NIR strategy I only used MSC and mean averaging as preprocessing steps, while NIR-III also used bucketing of the variables. Although strategy I shows similar results to those of strategy III, this preprocessing approach is not pursued further because bucketing ensures a more robust model, reduces the risk of overfitting and requires less computing time. In summary, the best preprocessing method is one of the simplest. Forgoing a part of the spectra and using derivatives resulted in a loss of information and thus lower classification accuracies. Consistent with all preprocessing strategies except those using the second derivative is that mean averaging lead to a higher test accuracy. The advantage of the median is its robustness and protection against outliers but using the mean average can improve the spectral resolution, leading to better classification results.

Predicted Class

		DE	FR	GE	IT	TR	Sensitivity
True Class	DE	20.8	5		0.4	0.8	77.0%
	FR	1.6	111.8	1	1.6		96.4%
	GE		1.2	11.8	0.4	1.6	78.7%
	IT		5.8	0.8	30.4		82.2%
	TR		1.6		0.2	36.2	95.3%
Precision		92.9%	89.2%	86.8%	92.1%	93.8%	90.6%

Figure 2. Confusion matrix of NIR preprocessing strategy III. The values given correspond to the mean of the five runs of the RNCV. Mean classification accuracy (90.6%), precision and sensitivity scores of the classes are also given. Confusion matrices of the other NIR preprocessing strategies are in the Supporting Information (S5). [Color]

The good classification results show the impact of geographical influences on the macronutrient profile of hazelnut, despite variable factors such as post-harvest processing and different harvest years, making NIR spectroscopy well suited for determining the geographical origin. An external classification accuracy of 90.6 ± 3.9% for a multiclass model with five classes is impressive in the field of geographical origin determination of food using NIR (figure 2). To put this result in relation to other publications in this field: A classification model determining the geographical origin of walnuts from seven countries achieved a classification accuracy of 77.0 ± 1.6% using a linear discriminant analysis.[7] Another study that investigated the geographical origin of almonds obtained a classification accuracy of 80.3 ± 1.5% when comparing six countries of origin with a support vector machine model.[6] Less complex models than that one, comparing only two sample groups of hazelnuts, have been developed by Moscetti et al. (2014) and Biancolillo et al. (2018).[11,12] Moscetti et al. (2014) compared ‘Nocciola Romana’ hazelnuts, which have a Protected Designation of Origin (PDO) indication, with hazelnuts of the ‘Tonda di Giffoni’ and ‘Barrettona’ cultivars and reported a classification accuracy of 95.5% using a support vector machine algorithm.[35] Biancolillo et al. (2018) investigated a similar question by comparing the ‘Nocciola Romana’ PDO with ‘other’ hazelnuts originating from Italy or the USA, resulting in a correct classification rate of 93.9% for ‘Nocciola Romana’ and 95.1% for ‘others’ by partial least square discriminant analysis.

In order to identify features that are responsible for this successful classification and to analyze their relationships, the feature selection approach surrogate minimal depth (SMD) was applied to the NIR data of preprocessing strategy III. Unlike other feature selection techniques, SMD does not evaluate the importance of the features individually, but by including their relations with each other.[14] SMD selected 245 of 744 buckets, and the high number of selected features can be explained by the fact that the bands in the NIR spectrum are quite broad and many features belong to the same signal. To obtain a more comprehensive interpretation of the important features, the mean adjusted agreement, a relation parameter that takes into account the mutual association to the result, is also obtained by SMD. The results of this relation analysis are presented in a heat map (Figure 3A) and in a spectrum colored according to the respective clusters of the relation analysis (Figure 3B). The heat map shows six distinct clusters that mainly contain neighboring buckets confirming the conclusion previously drawn from the high number of selected features. Somewhat surprisingly, four of the clusters are located in the wavenumber range between 11300-8700 cm^{-1} and contain only low intensity signals. Cluster 1 (red), cluster 2 (blue) and cluster 3 (purple), which show moderate to strong relations to each other, are in the region between 11500-10200 cm^{-1}, which can be assigned to the third overtone of C-H from methyl and methylene. However, cluster 3 (purple) contains features in the range of 10600-10200 cm^{-1}, where also bands of the N-H stretch-second overtone are found. The importance of the spectral regions represented by the clusters 1-4 is also confirmed by the results of the NIR preprocessing strategy IX, which did not include the buckets over 9000 cm^{-1} and gave a lower classification accuracy of 83.6 ± 3.8% (Table 1). However, the fact that this accuracy is still quite high shows that the features in clusters 5 (green) and 6 (gray) are even more important for classification. Moreover, these clusters show very interesting relations between the distinct regions 7000-6000 cm^{-1}, 4800-4700 cm^{-1} and 4400-4300 cm^{-1}. Signals in the region between 7000-6000 cm^{-1} can be assigned to the first overtone of N-H from the peptide bond and side chains of amino acids as well as the first overtone of O-H.[28,36] N-H combination from proteins, C-H/C=O lipid associated and O-H combination bands are in the region between 4800-4700 cm^{-1}.[28] Signals in the region between 4410-4390 cm^{-1} can be ascribed to C=O and N-H in α-helix and β-sheet structures in peptides.[28] Since all of these related bands can be assigned to functional groups of proteins, we can conclude that the successful classification is caused by different protein compositions of the hazelnut samples.

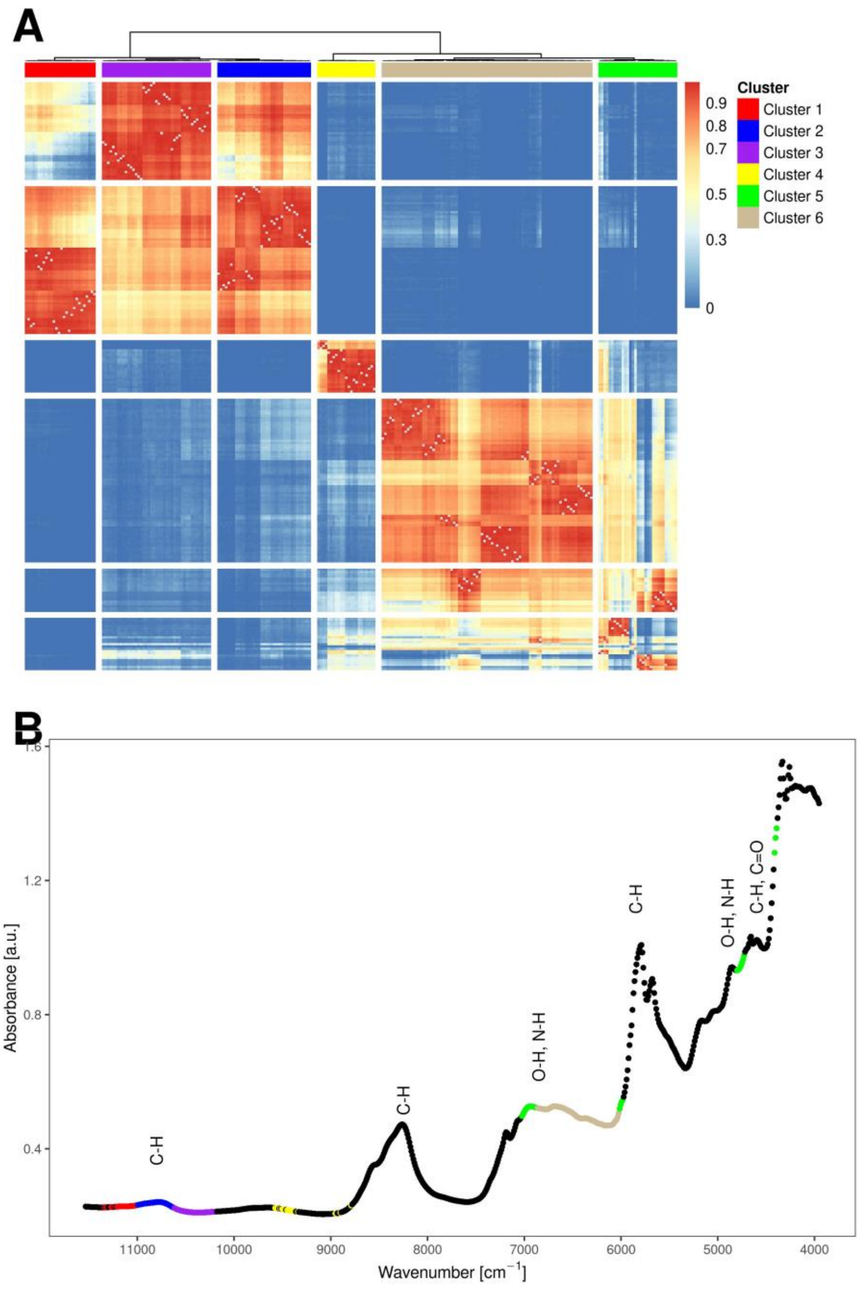


Figure 3. Results of the SMD feature selection and relation analysis: A Heatmap of the relation parameter mean adjusted agreement of the 245 selected features (A), as well as an example NIR hazelnut spectrum with buckets colored according to the respective associated clusters (B) are shown.

It is also worth noting that the random forest analysis on which SMD is based on has a much lower classification accuracy of 60.1% compared to the previous analysis (Supporting Information, Figure S5.11). The main difficulty of the random forest model was to distinguish between similar groups of Germany and France, and Turkey and Italy. To assess how much of the relevant information is contained in the significant buckets, we repeated the RNCV and classification using only these features resulting in a classification accuracy of 86.4 ± 4.7%. As this is only 4.2% less than the model with the best performance using all features, this shows that the 245 selected buckets carry the most important information for classification and that the SMD performs well for feature selection, even though the random forest analysis shows comparatively poor results for classification. The averaged NIR spectra of samples from the five countries of origin (Supporting Information, Figure S4) were overlaid to check for differences in the spectra. In agreement with the feature selection results, the spectral regions 11500-10000 cm^{-1}, 7000-6000 cm^{-1} and 5100-4500 cm^{-1} show the largest differences. The region between 4400-4000 cm^{-1} also shows differences, but the signals in this region were susceptible to matrix effects, as evident by the inconsistencies of this region across the five measurements of a single sample.

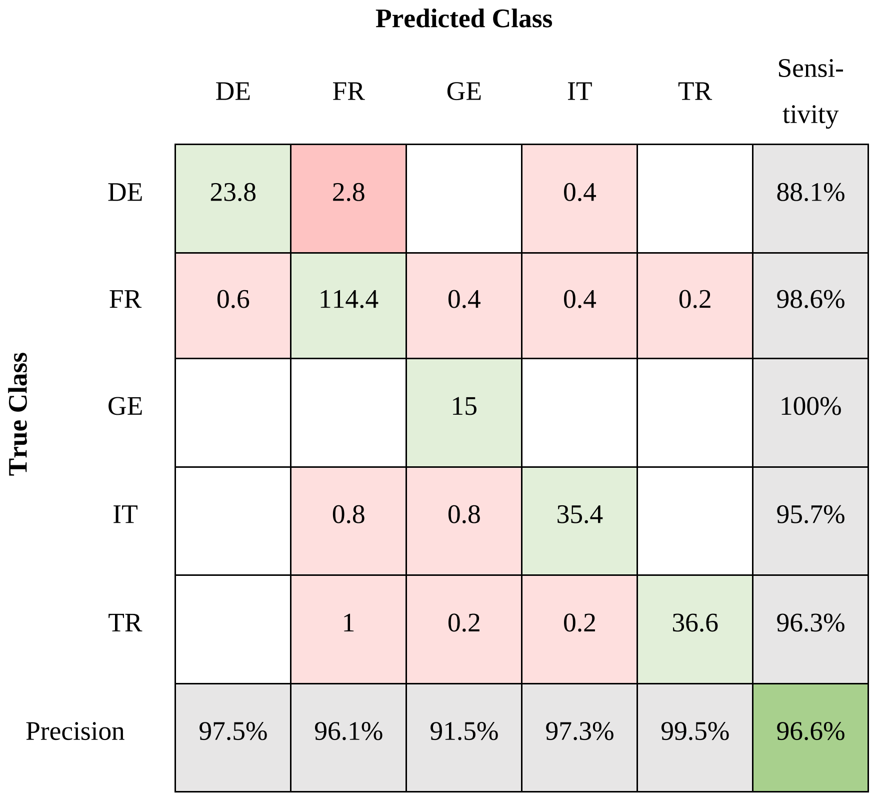
3.2. NIR-NMR-Data Fusion

In the original publication, which used ^{1}H NMR spectroscopy to determine the geographical origin of hazelnuts based on the polar metabolome, 262 hazelnut samples were divided into a training set containing two-thirds of the samples (172) and a test set with one-third of the samples (90).[2] Subsequently, several classification algorithms were trained with the training set and the test set was used to externally evaluate model performance. As with the results of the NIR spectroscopy, an ensemble of discriminant classifiers with a random subspace algorithm showed the best results. This is quite interesting because despite different observed features, both datasets appear to have similar underlying structures in the processed data. The model achieved a cross validation accuracy of 91% for the training set and an accuracy of 96% for the test set. Due to the use of a smaller number of samples and a stratified repeated nested cross validation to capture the variance of different sample splits, the NMR classification results were recalculated using the same external splits into training and test sets. The NMR data were Fourier transformed, baseline corrected, phased and 222 regions were defined as variable sized buckets, which were normalized to total intensity by scaling.[2] The NMR measurements were performed in triplicate, so the median and mean were compared for averaging. The classification results of the NMR and Fusion preprocessing strategies are shown in Table 2.

Table 2. Test accuracy, macro-F_1 score and Fleiss' kappa coefficient of the recalculated NMR analysis and the different NIR-NMR-Fusion approaches obtained using an ensemble of discriminant classifiers using random subspaces.

Strategy	Preprocessing	Test Accuracy	Macro-F_1	Fleiss' Kappa
NMR-I	NMR – Mean	94.3 ± 3.2%	93.6%	93.2%
NMR-II	NMR – Median	95.1 ± 3.0%	94.4%	94.2%
Fusion-I	NIR-III + NMR-Mean Fusion	96.1 ± 2.7%	95.3%	95.1%
Fusion-II	NIR-III + NMR-Median Fusion	96.1 ± 3.2%	95.4%	94.0%
Fusion-III	NIR-III + NMR-Mean Fusion – Autoscaled	96.6 ± 2.8%	96.0%	96.4%
Fusion-IV	NIR-III + NMR-Median Fusion – Autoscaled	96.0 ± 2.7%	95.4%	95.8%

The results of the mean and median were quite similar. Median averaging of the NMR data yielded a test accuracy of 95.1 ± 3.0%, a macro-F_1 score of 94.4% and a Fleiss' Kappa coefficient of 94.2%. NMR-I, using the mean of the NMR data, achieved a 0.8% lower classification rate and similarly low values for the macro-F_1 score and the Fleiss' Kappa coefficient. Due to the similar results of the two NMR spectroscopy strategies, both datasets were tested in a low-level data fusion with the NIR spectroscopy strategy III data. In this case, all buckets used for NIR, and NMR analysis were combined in a matrix resulting in 966 features. The dataset was then scaled using autoscaling, resulting in a standard deviation of one for each feature.[37] Fusion-I combined the data of NIR-III and NMR-I, while Fusion-II used the median averaged NMR buckets. Both fusions yielded similar results, with a test accuracy of 96.1% and only small differences in the other measures. Test accuracy of Fusion-I is 1% higher compared to NMR-II, which only uses the median averaged NMR buckets, indicating a better classification performance of the model. Fusion-III and Fusion-IV subjected the datasets from Fusion-I and -II to autoscaling. Fusion-IV used the fused dataset of NIR strategy III and the median averaged NMR buckets leading to a test accuracy of 96.0 ± 2.7% thus showing almost the same results as Fusion-I and -II. Fusion-III instead used mean averaged NMR buckets in the fusion and yielded a test accuracy of 96.6 ± 2.8%, a macro-F1 score of 96.7% and a Fleiss' kappa coefficient of 96.4%, thus showing the best results for each measure out of all examined models. Compared to the NMR-II approach, test accuracy increased by 1.5%. Such an increase is quite large considering that the statistical measures of all models are higher than 90%. A further comparison of NMR-II (Supporting Information, Figure S6.2) and Fusion-III (Figure 4) shows improvements in the classification of the German, French and Italian samples, while the accuracy for Georgian and Turkish samples remains the same.



True Class \ Predicted Class	DE	FR	GE	IT	TR	Sensitivity
DE	23.8	2.8		0.4		88.1%
FR	0.6	114.4	0.4	0.4	0.2	98.6%
GE			15			100%
IT		0.8	0.8	35.4		95.7%
TR		1	0.2	0.2	36.6	96.3%
Precision	97.5%	96.1%	91.5%	97.3%	99.5%	96.6%

Figure 4. Confusion matrix of Fusion-III. The values correspond to the mean of the five runs of the RNCV and the mean classification accuracy (96.6%), precision and sensitivity scores of the classes are also shown. Confusion matrices of the NMR strategies and the other data fusions are shown in the Supporting Information (S6).

The individual allocations of the classification model of Fusion-III (Supporting Information, Table S7) were examined to obtain information about problematic samples. In total, 13 samples were misclassified at least once, and four samples were misclassified in each split of RNCV. These were two German samples, one from Bavaria and one from Rhineland-Palatinate, one French sample and one Turkish sample. These samples were also misclassified at least once in the individual models of NIR-III and NMR-II, but only the Turkish sample was falsely classified in all RNCV splits. One of the misclassified German samples was from Rhineland-Palatinate, which was the only sample from this federal state. It was classified three times as a French sample and twice as an Italian sample. The other German sample is of mixed variety from the municipality of Aiglsbach in Bavaria, which has always been misclassified as French. As there are five samples from Aiglsbach, the reason for these misclassifications is probably not the lack of enough samples for training but the individual composition of this sample. For a similar reason, a Turkish mixed variety sample from the Düzce region was always misclassified as French even though there are 13 samples with the same characteristics that were not misclassified once. The models did not show clear results regarding the French sample, which was misclassified in all five splits. It was classified twice as Georgian and Italian sample and once as Turkish sample. Again, these misclassifications are probably due to the individual composition of this sample, as there were seven other samples of the Pauetet variety from the Midi-Pyrénées region, which were always correctly classified.

For comparison, a complementary approach, i. e. cross-checking the NIR and NMR classification results with respect to misclassified samples, was performed using the best performing models based on NIR-III and NMR-II (Supporting Information, Table S7). The results of the two models show a large overlap in the classification of the samples. 192 of the 233 samples analyzed were correctly classified by both models. A total of 41 samples were misclassified at least once in the RNCV by the NIR approach and 21 samples by the NMR approach. Again, the models overlap, as 11 samples were misclassified at least once by both models. This demonstrates that both classification models perform well and misclassify similar samples. However, it is also shown that the models also provide complementary information. In total, only two samples were incorrectly classified in all five splits, all as French, of the RNCV and in both models indicating the conservative nature of this approach. One of these samples is a Turkish sample of a mixed variety from the province of Düzce and the other is of the ‘Tonda di Giffoni’ variety from the Campania region in Italy, which has a Protected Geographical Indication (PGI). The Turkish sample is the same one that Fusion-III misclassified. Of all 233 samples, eight are of the ‘Tonda di Giffoni’ variety from France and three from Italy. This suggests that the metabolome of the misclassified Italian ‘Tondi di Giffoni’ sample may be more influenced by the cultivar than by environmental factors, so more samples of this cultivar from Italy are needed to adequately train the models. Of the 11 samples that were misclassified at least once in the complementary approach, eight were also misclassified at least once in Fusion-III, indicating the similarity of the results. Another factor to consider at this point is the fact that these samples were obtained from suppliers and in principle there is the possibility of a mix-up, even if it is very unlikely.

The question remains whether a data fusion should be used when determining the geographical origin of hazelnuts. If both methods have already been used, data fusion may even give better results than the methods on their own. In this case, the fusion and subsequent autoscaling of the NIR dataset, which used MSC, bucketing and mean, with the NMR dataset, which used the mean of the buckets, gave the best individual model for the hazelnut geographical origin question. However, the complementary approach of analyzing samples sequentially using NIR and NMR spectroscopy proves to be a more conservative and reliable method. This method would also be suitable for transfer to industry, where NIR analysis is used as a level 1 analysis and conspicuous samples are subsequently analyzed in an analytical laboratory using NMR as a level 2 analysis.

CRediT authorship contribution statement

Navid Shakiba: Conceptualization, Investigation, Visualization, Methodology, Writing – original draft

Annika Gerdes: Investigation, Writing – review & editing

Nathalie Holz: Investigation, Writing – review & editing

Sören Wenck: Investigation, Visualization, Writing – review & editing

René Bachmann: Investigation, Writing – review & editing

Tobias Schneider: Software, Writing – review & editing

Stephan Seifert: Supervision, Writing – review & editing

Markus Fischer: Supervision, Resources, Writing – review & editing

Thomas Hackl: Conceptualization, Supervision, Resources, Writing – review & editing

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank Maike Arndt for providing expertise and assistance. We would also like to thank the hazelnut suppliers for the samples.

Funding

The NMR part of the present project of the Research Association of the German Food Industry (FEI) was funded by the German Federation of Industrial Research Associations (AiF project: 20506N) within the framework of the program for the promotion of joint industrial research (IGF) of the Federal Ministry of Economic Affairs and Energy (BMWi) on the basis of a resolution of the German Bundestag.

The NIR part including the data fusion was developed within the project “Food Profiling – Development of analytical tools for experimental verification of the origin and identy of food”. This project (funding code: 2816500914) is funded by the Federal Ministry of Food and Agriculture (BMEL) based on a resolution of the German Bundestag. Project funding was provided by the Federal Agency for Agriculte and Food (BLE) within the framework of the Innovation Promotion Program.

Appendix: Supporting Information

Supporting Information of this article can be found online at

4. References

[1] Food and Agriculture Organization of the United Nations, Production of Crops, 2021. http://www.fao.org/faostat/en/#data (accessed 25 February 2021).

[2] R. Bachmann, S. Klockmann, J. Haerdter, M. Fischer, T. Hackl, 1H NMR Spectroscopy for Determination of the Geographical Origin of Hazelnuts, J. Agric. Food Chem. 66 (2018) 11873–11879. https://doi.org/10.1021/acs.jafc.8b03724.

[3] S. Klockmann, E. Reiner, R. Bachmann, T. Hackl, M. Fischer, Food Fingerprinting: Metabolomic Approaches for Geographical Origin Discrimination of Hazelnuts (Corylus avellana) by UPLC-QTOF-MS, J. Agric. Food Chem. 64 (2016) 9253–9262. https://doi.org/10.1021/acs.jafc.6b04433.

[4] R. Vitale, M. Bevilacqua, R. Bucci, A.D. Magrì, A.L. Magrì, F. Marini, A rapid and non-invasive method for authenticating the origin of pistachio samples by NIR spectroscopy and chemometrics, Chemometrics and Intelligent Laboratory Systems 121 (2013) 90–99. https://doi.org/10.1016/j.chemolab.2012.11.019.

[5] H. Zhao, B. Guo, Y. Wei, B. Zhang, Near infrared reflectance spectroscopy for determination of the geographical origin of wheat, Food Chem. 138 (2013) 1902–1907. https://doi.org/10.1016/j.foodchem.2012.11.037.

[6] M. Arndt, M. Rurik, A. Drees, C. Ahlers, S. Feldmann, O. Kohlbacher, M. Fischer, Food authentication: Determination of the geographical origin of almonds (Prunus dulcis Mill.) via near-infrared spectroscopy, Microchemical Journal 160 (2021) 105702. https://doi.org/10.1016/j.microc.2020.105702.

[7] M. Arndt, A. Drees, C. Ahlers, M. Fischer, Determination of the Geographical Origin of Walnuts (Juglans regia L.) Using Near-Infrared Spectroscopy and Chemometrics, Foods 9 (2020). https://doi.org/10.3390/foods9121860.

[8] R.J. Mailer, Rapid evaluation of olive oil quality by NIR reflectance spectroscopy, Journal of the American Oil Chemists' Society 81 (2004) 823–827. https://doi.org/10.1007/s11746-004-0986-4.

[9] Q. Chen, J. Cai, X. Wan, J. Zhao, Application of linear/non-linear classification algorithms in discrimination of pork storage time using Fourier transform near infrared (FT-NIR) spectroscopy, LWT - Food Science and Technology 44 (2011) 2053–2058. https://doi.org/10.1016/j.lwt.2011.05.015.

[10] G. Yildiz, R.L. Wehling, S.L. Cuppett, Method for Determining Oxidation of Vegetable Oils by Near-Infrared Spectroscopy, Journal of the American Oil Chemists' Society 78 (2001) 495–502. https://doi.org/10.1007/s11746-001-0292-1.

[11] R. Moscetti, E. Radicetti, D. Monarca, M. Cecchini, R. Massantini, Near infrared spectroscopy is suitable for the classification of hazelnuts according to Protected Designation of Origin, J. Sci. Food Agric. 95 (2015) 2619–2625. https://doi.org/10.1002/jsfa.6992.

[12] A. Biancolillo, S. de Luca, S. Bassi, L. Roudier, R. Bucci, A.D. Magrì, F. Marini, Authentication of an Italian PDO hazelnut ("Nocciola Romana") by NIR spectroscopy, Environ. Sci. Pollut. Res. Int. 25 (2018) 28780–28786. https://doi.org/10.1007/s11356-018-1755-2.

[13] M. Arndt, M. Rurik, A. Drees, K. Bigdowski, O. Kohlbacher, M. Fischer, Comparison of different sample preparation techniques for NIR screening and their influence on the geographical origin determination of almonds (Prunus dulcis MILL.), Food Control 115 (2020) 107302. https://doi.org/10.1016/j.foodcont.2020.107302.

[14] S. Seifert, S. Gundlach, S. Szymczak, Surrogate minimal depth as an importance measure for variables in random forests, Bioinformatics 35 (2019) 3663–3671. https://doi.org/10.1093/bioinformatics/btz149.

[15] S. Seifert, Application of random forest based approaches to surface-enhanced Raman scattering data, Sci. Rep. 10 (2020) 5436. https://doi.org/10.1038/s41598-020-62338-8.

[16] V. Živanović, S. Seifert, D. Drescher, P. Schrade, S. Werner, P. Guttmann, G.P. Szekeres, S. Bachmann, G. Schneider, C. Arenz, J. Kneipp, Optical Nanosensing of Lipid Accumulation due to Enzyme Inhibition in Live Cells, ACS Nano 13 (2019) 9363–9375. https://doi.org/10.1021/acsnano.9b04001.

[17] E. Borràs, J. Ferré, R. Boqué, M. Mestres, L. Aceña, O. Busto, Data fusion methodologies for food and beverage authentication and quality assessment - a review, Anal. Chim. Acta 891 (2015) 1–14. https://doi.org/10.1016/j.aca.2015.04.042.

[18] Å. Rinnan, F. van den Berg, S.B. Engelsen, Review of the most common pre-processing techniques for near-infrared spectra, TrAC Trends in Analytical Chemistry 28 (2009) 1201–1222. https://doi.org/10.1016/j.trac.2009.07.007.

[19] T. Isaksson, T. Næs, The Effect of Multiplicative Scatter Correction (MSC) and Linearity Improvement in NIR Spectroscopy, Applied Spectroscopy 42 (1988) 1273–1284. https://doi.org/10.1366/0003702884429869.

[20] J.E. Jackson, A User's Guide To Principal Components, John Wiley & Sons, 1991.

[21] J. Meurs, scaledata, 2021. https://github.com/jorismeurs/scaledata (accessed 28 January 2021).

[22] S. Watermann, C. Schmitt, T. Schneider, T. Hackl, Comparison of Regular, Pure Shift, and Fast 2D NMR Experiments for Determination of the Geographical Origin of Walnuts, Metabolites 11 (2021). https://doi.org/10.3390/metabo11010039.

[23] T.K. Ho, The random subspace method for constructing decision forests, IEEE Trans. Pattern Anal. Machine Intell. 20 (1998) 832–844. https://doi.org/10.1109/34.709601.

[24] J.L. Fleiss, B. Levin, M.C. Paik, Statistical Methods for Rates and Proportions, third ed., Wiley, 2003.

[25] M.N. Wright, A. Ziegler, ranger A Fast Implementation of Random Forests for High Dimensional Data in C++ and R, J. Stat. Soft. 77 (2017). https://doi.org/10.18637/jss.v077.i01.

[26] U.S. Department of Agriculture, USDA Food and Nutrient Database for Dietary Studies 2017-2018. http://www.ars.usda.gov/nea/bhnrc/fsrg (accessed 25 February 2021).

[27] P.L. Benitez-Sánchez, M. Len-Camacho, R. Aparicio, A comprehensive study of hazelnut oil composition with comparisons to other vegetable oils, particularly olive oil, European Food Research and Technology 218 (2003) 13–19. https://doi.org/10.1007/s00217-003-0766-4.

[28] J. Workman, L. Weyer, Practical guide and spectral atlas for interpretive near-infrared spectroscopy, second ed., CRC Press, Boca Raton, FL, 2012.

[29] B. Worley, R. Powers, Multivariate Analysis in Metabolomics, Curr. Metabolomics 1 (2013) 92–107. https://doi.org/10.2174/2213235X11301010092.

[30] S. Guo, P. Rösch, J. Popp, T. Bocklitz, Modified PCA and PLS: Towards a better classification in Raman spectroscopy–based biological applications, Journal of Chemometrics 34 (2020). https://doi.org/10.1002/cem.3202.

[31] F. Gharibnezhad, L.E. Mujica, J. Rodellar, Applying robust variant of Principal Component Analysis as a damage detector in the presence of outliers, Mechanical Systems and Signal Processing 50-51 (2015) 467–479. https://doi.org/10.1016/j.ymssp.2014.05.032.

[32] A. Tharwat, Classification assessment methods, ACI 17 (2021) 168–192. https://doi.org/10.1016/j.aci.2018.08.003.

[33] T.R. Nichols, P.M. Wisner, G. Cripe, L. Gulabchand, Putting the Kappa Statistic to Use, Qual Assur J 13 (2010) 57–61. https://doi.org/10.1002/qaj.481.

[34] T. Segelke, S. Schelm, C. Ahlers, M. Fischer, Food Authentication: Truffle (Tuber spp.) Species Differentiation by FT-NIR and Chemometrics, Foods 9 (2020). https://doi.org/10.3390/foods9070922.

[35] European Commission, Regulation (EC) No 510/2006 'Nocciola Romana' PDO, Official Journal of the European Union (2008).

[36] E.W. Ciurczak, B. Igne, J. Workman, D.A. Burns (Eds.), Handbook of near-infrared analysis, CRC Press/Taylor & Francis Group, Boca Raton, 2021.

[37] R.A. van den Berg, H.C.J. Hoefsloot, J.A. Westerhuis, A.K. Smilde, M.J. van der Werf, Centering, scaling, and transformations: improving the biological information content of metabolomics data, BMC Genomics 7 (2006) 142. https://doi.org/10.1186/1471-2164-7-142.