RUNNING TITLE

Chemometrics in food traceability

TITLE

The impact of chemometrics on food traceability

AUTHORS

Lucia Bertacchini, Marina Cocchi, Mario Li Vigni, Andrea Marchetti, Elisa Salvatore, Simona Sighinolfi, Michele Silvestri and Caterina Durante*

AFFILIATION

Department of Chemical and Geochemical Sciences, University of Modena and Reggio Emilia

Via Campi 183, 41125 Modena, Italy

*Corresponding authors: Caterina Durante, PhD

email: caterina.durante@unimore.it

ABSTRACT (MAX 250)

In the last decades, mankind has become totally aware about the importance of food quality: nowadays authentication and traceability are words of general use.

Food authentication verifies how much a food is in accordance with its label description and law and it could be considered a further guarantee for the quality and safety of a foodstuff.

The traceability of food could be considered an essential element in ensuring safety and high quality of food. The synergistic use of instrumental analytical techniques and chemometrics represents a promising way to obtain trustworthy results in the development of authenticity and traceability models. This chapter deals with the potentialities of chemometrics tools in resolving some real issues related to food traceability and authenticity. Particular attention will be paid to the use of some exploratory, classification and discrimination techniques.

In the first part of this chapter, a briefly description of European regulations (*Authenticity and Traceability: the European Union point of view*), and traceability and authenticity markers (*Authenticity and Traceability: a scientific point of view*) is reported. The second part is split into two sections: namely *Food Authenticity* and *Food Traceability applications*, where the main features and advantages of some chemometrics approaches are presented.

Introduction

We are what we eat and ..... we probably prefer to eat what we know….

In the last decades, mankind has become totally aware about the importance of food quality for different reasons [1]. According to ISO 9000:2005, quality is defined as the degree to which a set of inherent characteristics fulfils the requirements [2]. Broadly speaking, quality could be referred as the gap between how good food is and how better it could be. To consumer, quality might be considered the most important ingredient in food.

Food *authentication* verifies how much a food is in accordance with its label description and law [3] and it could be considered a further guarantee for the quality and safety of a foodstuff. In European Union, EU, the production and the commercialization of food are rightly protected and regulated by several rules [4-9] always according to quality and healthy criteria, as it is better explained in the following sessions.

Unfortunately, the various food scares and inappropriate risk management practises, that occurred around the world, such as Minamata disease, wine adulteration with methanol, dioxin in meat, bovine spongiform encephalopathy, BSE, avian influenza, etc., have highlighted the need to have as clear as possible declared information about the used raw materials, origin and processing of food. In recent years, EU has focused its efforts to support legal regulations [4-9] with scientific studies aimed at developing markers and analytical techniques to assess the geographical origin of foodstuff [10]. The main EU goals are directed to assure an honest marketing competitiveness and augment consumers' confidence by promoting and recognising the food bio-diversities and typicality through their well-established geographical origin. Moreover, it has also been stressed that highlighting the origin of food also represents an enhancement of the reputation of the product itself.

Nowadays, the synergistic use of instrumental analytical techniques and chemometrics represents the best way to obtain trustworthy results in the development of authenticity and traceability models [11-26].

Someone could ask: why chemometrics? Today, there are different reasons, which lead the researchers to use chemometrics. First of all, authenticity and traceability issues are multivariate; in fact, they cover many different aspects including the chemical and physical characterization, the adulteration, the discrimination, the control of production process and mislabelling.

In order to face these aspects, different analytical strategies, mainly based on instrumental techniques such as HPLC/MS, GLC/MS, UV-Vis, NMR, NIR, ICP/MS etc., are employed [11-27]. They yield to produce prodigious volumes of analytical data. Unfortunately, not always lot of data allows obtaining useful information. Thus, the versatility, flexibility and immediacy of chemometrics techniques may help in reducing the complexity of data. Generally, chemometrics tools help to extract useful information, which in turn, improves the interpretation and the presentation of the results.

In the first part of this chapter, a briefly description of European regulations (*Authenticity and Traceability: the European Union point of view*), and traceability and authenticity markers (*Authenticity and Traceability: a scientific point of view*) is reported. The second part is split into two sections: namely *Food Traceability* and *Food Authenticity applications*, where the main features and advantages of some chemometrics approaches are presented.

This chapter deals with the potentialities of chemometrics tools in resolving some real problems related to food traceability and authenticity. Since each chemometrics technique is widely explained in the previous chapters of the present book, only the application aspects are discussed in the present treatise. Particular attention will be paid only to the use of some exploratory, classification and discrimination techniques. Moreover, all instrumental details are outside the aim of this chapter and will not be presented here.

Authenticity and Traceability: the European Union point of view

Quality has been used, over the past years, mainly to describe subjective attributes such as beauty, goodness, expensiveness, freshness etc. Probably, the first *quality guru* was Joseph M. Juran that published his first quality-related article in 1935 [28]. In the manuscript, the author defined the "*quality*" as the "meeting or exceeding customer expectations." Some years later, quality was defined, according to ISO 9000:2000 [2]. The European Union has introduced many regulations to guarantee and protect the quality of the foodstuffs [EC 2081/1992, EC 2082/1992, EC 510/2006, EC 1898/2006, EC 628/2008, EC 178/2002]. The EU aim is not to limit innovation or to homogenize food products available on the European market but mainly to: (i) set fundamental norms of safety both on communitarian and international contexts; (ii) develop and implement quality excellence and, at the same time (iii) guarantee an high degree of food safety maintaining the peculiarities, tradition, of each food production.

One of the first European Union Protected Schemes came into force in 1992 (EC No 2081/92, 1992) and laid down a labelling system for the protection of food names on a geographical basis. In particular, the *Protected Designation of Origin* (PDO) and the *Protected Geographical Indication* (PGI) were introduced as quality markers in order to mainly protect traditional food. According to PDO denomination a product must be obtained, processed and prepared in a defined geographical area. It means that the quality or characteristics of the product are essentially linked to that area and to the skilfulness of producers (*savoir faire*). As *per* PGI, a food must be produced, processed or prepared in a geographical area. Thus, the characteristics, quality or reputation of food are attributable to that area. Afterwards, the Regulation EC 2082/1993 introduced two new brands (*Traditional Speciality Guaranteed*, TSG and *Certificate of Specific Character*, CSC) for specific character of agricultural products with the aim to protect the traditional recipes.

The efforts of the European Union in the safety and quality context can also be seen in the publication of many reports such as *From the farm to the fork* [29] and in the adoption of the *White paper on the food safety* [30] or the *Green paper* [31]. In this case, the Commission sets out the general principles on which European food safety policy should be based with the aim to achieve the highest possible level of health protection for the consumers.

However, a further assurance of quality and safety of food is also linked to its geographical origin. Aware of this fact, with EC 178/2002, the EU introduced the concept of traceability in the food chain and instituted the European Food Safety Authority, EFSA.

According to EC 178/2002, traceability is the ability to trace and follow the history of food by means of recorded objective identification procedures, starting from every level of the agro-food chain. The traceability of food along the food production is an essential element in ensuring safety and high quality of food.

Finally, the European Union continues to be always open with respect to these issues by also funding research projects on the topics of the geographical traceability and development of innovative analytical approaches for food authenticity within several European framework programs.

Authenticity and Traceability: a scientific point of view

The better way to assess the "quality" of food is eating it. Unfortunately, this is not always possible and the final response is subjective.

In the last decades, the interest of different researchers towards the use of analytical methodologies supporting the authenticity and traceability of food is more and more increasing. In particular, chemometrics tools combined with different analytical techniques are widely used to verify the authenticity of food. In the past two decades, large number of contributions including numerous review articles and books dealing with quality and authenticity control have been published [11-26].

Several analytical strategies relying on instrumental techniques have been employed to deal with authenticity purposes. Briefly, these methods can be categorized into two types depending on their operating principle [16,27], namely: (i) physical-chemical techniques, such as spectroscopy, spectrometry, chromatography, pyrolysis, electronic nose, etc. and (ii) molecular biological approaches, such as DNA based methods. A recent study [27] shows that among the analytical methods, employed for obtaining an elemental fingerprint of the investigated food matrix, inductively coupled plasma (ICP)-based techniques are widely used (63% of cases) followed by atomic spectrometry (30%) and instrumental neutron-activation analysis (5%). Furthermore, it has been shown that the determination of stable-isotope ratio furnished good results as discriminator descriptor for different foodstuffs. In this case, isotope-ratio mass spectrometry has been used in 61% of cases followed by nuclear magnetic resonance (36%) and X-ray fluorescence spectrometry (3%) [27].

Chromatographic and spectroscopic techniques are within the most important tools used in food authentication. In particular, gas chromatography (GLC) and liquid chromatography (LC) are

widely used for the determination of both polar and non-polar compounds (volatile compounds in the case of GLC and organic acids, amino acids, polyphenols, etc. for LC). Almost all the infrared spectroscopic techniques**,** Near Infrared Spectroscopy, NIR, and Mid-Infrared Spectroscopy, MIR, are cheap, rapid and non-destructive. All of them are frequently used in combination with chemometrics.

DNA-based methods are also very important because the identification of species is often vital for scientific reasons as well as for mankind purposes, such as food allergy, cultural reasons or simply for personal preference [16].

As far as the choice of a chemometrics technique to be used is concerned, it strongly depends on the aim of the investigation. In the authentication context, the most general applications [16-21] are: (i) characterisation of foodstuffs, (ii) discrimination or classification into one or several categories, (iii) monitoring and controlling of production processes and (iv) identification of product adulteration, dilution and contamination. Wine is, by far, the food matrix for which chemometrics has been frequently called upon followed by cheese, olive oil, honey, meat, and so on [27].

On the other hand, the problem of food authenticity is strictly tied with the production chain traceability or, in the more complex case, with the geographical traceability of the product and/or its raw material. Unfortunately, tracing and/or tracking procedures, although they monitor the mass flow of incoming and outgoing materials, are often limited since they are not referred to objective data but mostly based on ‘declarations’ only supported by paper documents. Hence, it is of utmost importance and relevance to develop analytical tools able to ‘certify’ the provenience of food in order to accomplish food control and quality valorisation.

As regards the development of a geographical traceability models, it is worth to note that their robustness is also strictly related to the reality in which the food lies. In fact, typical products, with protection indication labels and characteristic of extensive geographical areas, require models developed on 'large scale', which take into account analytical parameters (indicators) and a systematic knowledge of the investigated foodstuffs as well as of their territory of origin (*mapping*). Food, for which the disciplinary regulations involve "more restricted" areas and/or productions linked to the brand of the producers, requires the development of more detailed traceability models characterised by a less uncertainty level of the investigated indicators. In both the cases, the use of chemometrics techniques results imperative.

Generally speaking, the used analytical parameters, namely traceability indicators, can be distinguished into *primary or direct* and *secondary or in-direct* [11]. *Primary indicators,* i.e.

elemental composition, radiogenic isotope ratio and stable isotope ratio can be directly linked to the same determinations in soil samples or referred to particular geographical area. On the other hand, s*econdary indicators* i.e. variables related to food composition/making procedure, can be regarded as foodstuff fingerprint. These variables may allow discrimination from food products of different geographic origin but cannot be directly linked to the production zone.

Actually in traceability studies, there are mainly three different analytical approaches that are adopted in order to link the food commodities to the territory of provenance. In the first one, measurements of traceability indicators are carried out on a representative number of food samples and a possible relationship with their geographical origin is found by using multivariate models. These models are based on the existence of a calibration set, "training samples", where the geographical origin and the authenticity is somehow guaranteed, while geographical, geological, and climatic reference information are taken from paper documentations [20, 32-35]. In the second approach, the primary indicators are directly measured both on the investigated food as well as in some reference samples of their soils of origin [36-38]. In this case, no-systematic selection/sampling of investigated soils is used. In the third approach, a representative soil sampling is considered as well taking into account climate, geographical and geological features referred to extended macro-areas (i.e. 100÷200 km^2) [10]. The choice of one approach over another mainly depends on the aims of the research and on the posed question too. In any case, it is of utmost importance to build models that take into account the cause / effect relationship between the monitored variables.

Therefore, in the majority of the cases, robust models can be produced by the aid of chemometrics tools that operate, in a dynamic bidirectional process, with the instrumental analytical techniques.

Far away from typing the word end to this argument, in the present chapter some applicative examples are presented and discussed.

Food Traceability applications

Chemometrics approaches for soil sampling planning in traceability studies

Introduction

A suitable planning of sampling is the first and always crucial issue in scientific investigations. In particular, when dealing with development of geographical traceability models for PDO (Protected Designation of Origin) food, representativeness of sampling is certainly an imperative in order to obtain robust models and establish a food-territory link.

Before starting any sampling, several parameters should be set, i.e. the number of samples, the sampling sites, the frequency of sampling, the procedures and so on, always according to the aim of the study. For instance, when the goal is to relate food to soil, it is of utmost importance to obtain an *a priori* knowledge about the different geological features of the territory, since some direct traceability indicators, such as radiogenic isotope signature and trace element composition, are also influenced by the soil characteristics. Furthermore, the productivity related to the investigated area should be considered as well, in order to correctly weight the sampling. In other words, combined use of productivity and geological information should give a clear and complete description of investigated area, important for obtaining trustworthy results. Unfortunately, the involved variables are very often numerous and characterized by different nature. Thus, it emerges the need to use analytical tools able to simultaneously take into account all of them for achieving a correct sampling.

As far as soils sampling is concerned, conventional strategies suggest a scheme, regular and circular grids, systematic and non-systematic patterns, unaligned random sampling [39] etc., for collecting samples usually within limited areas. Other approaches, mainly based on multivariate techniques, are able to consider different characteristics of investigated area, such as production area, density and so on [40,41].

In this work, multivariate characterization of soils, i.e. principal component analysis (PCA) for the evaluation of the soils characterization data (geological features, winegrowing coverage, grape varieties and yearly productions), and experimental design (onion design) applied on the resulting PCA latent variables, were synergistic used. In particular, a sub-set of representative soil samples was selected for setting up traceability models of two PDO oenological products coming from Modena district, namely *Lambrusco* wines and *Aceto Balsamico Tradizionale di Modena* (ABTM).

Their stringent production regulations [42, 43] allow the grape cultivation and the productive steps only within the district, whose area is estimated in 2700 km^2, of which 90 km^2 are grape-cultivated.
In this case, the investigated area is quite extended and cannot be considered geologically homogeneous. Furthermore, a punctual sampling and examination is not feasible, considering the total number of samples to be analyzed.
Finally, a reduced sub-set of sites was considered to investigate variability of chemical features of soils sampled among different farms, within the same farm and as function of sampling depth. The aim is to establish the proper number of samples to be collected for each farm and the depths to investigate. This goal was achieved by the combined use of X-ray powder diffraction technique, which produces a fingerprint related to the composition and morphological structure of the soil, and chemometrics [26].

Selection of representative soil samples

The representative sampling sites were chosen among all farms (around 4600) enrolled in Wine-Vine Register of the Modena district. Information about the total extension of the farms, their relative area in each municipality and all the cultivated grape varieties was considered. In particular, the productivity-related variables were firstly used to obtain an *a priori* screening of the farms. The farms with none of the grape varieties listed in the production regulations of ABTM or *Lambrusco* wines are discarded. The four widest producers for each municipality and the others differing at least for one grape variety were always taken into account.
The application of these criteria led to the identification of 705 farms, 466 located in the alluvial plane zone (**A**) and 239 in the Apennines margin (**B**).
Afterwards, the spatial coordinates X and Y (referred to a UTM ED 1950 system) and geo-morphological, pedological and lithological data were considered and codified in order to be used for the sampling sites selection, as extensively described in previous studies [44] and summarised in **Table 1**.

PLEASE, INSERT HERE TABLE 1

The data was arranged in two different bi-dimensional matrices of 466×45 and 239×44 (soils × variables) dimensions, for samples belonging to area **A** and **B** respectively, since the information obtained for the two areas was different.

Given the multivariate nature of the investigated system, Principal Component Analysis (PCA) was applied to each autoscaled matrix and suitable models were chosen taking into account the explained variance (R^2), the eigenvalues as well as the information held in the loadings.

Eleven and seven principal components (PCs) were used for building models for **A** (49.82% of total data variance) and **B** (45.51% of total data variance) matrices, respectively.

The obtained score values were used for the selection of samples by means of experimental design techniques, in order to achieve the maximum coverage and uniformity. In particular, a linear D-optimal Onion design [45] was chosen since the experimental domain is quite extended and not-regular. Moreover, the design, splitting the initial space into a number of concentric subsets ('shells' or 'layers'), allows to optimise the space division (using G-efficiency criterion) and the sampling selection (by means of linear D-optimal model). The G-efficiency criterion [46] compares the performance of D-optimal designs with different number of design runs, where each design has been computed according to D-optimality (maximization of the determinant of the information matrix referring to the candidate set).

Seventy and thirty producer fields (from which soils will be sampled) were set as initial request for **A** and **B** matrices, respectively; a linear model was fit and the G-efficiency criterion was used to select the best samples set.

A 5- and 3- layers onion designs were performed for **A** and **B** matrices, respectively and a careful layer filling was planned in order to achieve the best coverage of the space.

Since the samples located further from the centre of the PC space could be considered more peculiar with respect to the others, it was decided to choose as much samples as possible for the outer layers, taking into account the G-efficiency as well. **Figure 1** reports the proposed number of samples for each layer, the corresponding G-efficiency and the final number of samples selected used for filling layer samples, for **A** and **B** matrices, respectively.

PLEASE, INSERT HERE FIGURE 1

The selected sampling sites are highlighted within the score plots of the first three PCs for **A** and **B** areas in **Figure 2** and **Figure 3,** respectively.

PLEASE, INSERT HERE FIGURE 2

PLEASE, INSERT HERE FIGURE 3

Finally, all the selected fields were reported on pedological map of Modena district, using their X and Y coordinates and they resulted to be uniformly distributed for both the in-plane and hill regions (figure not reported).

Evaluation of soils variability

To evaluate the soils variability and establish the suitable sampling conditions, the attention was focused on four farms representative of the Modena district; three of these (A, B and D) are located in the in-plain region, while the other one (C) in the hill area. Three sampling sites for A and D and five for B and C were chosen and soil samples were collected at five different depths, starting from 10 cm up to 60 cm (a = 10-20 cm, b = 20-30 cm, c = 30-40 cm, d = 40-50 cm, e = 50-60 cm), giving a total number of 80 samples (16 sampling points x 5 depths).

All the samples were analysed by means of X-ray diffraction (XRD) of powder and the collected diffractograms consist of 5200 data point covering the region from 5° to 92° 2θ.

Due to the complexity of the signals, some preprocessing treatments were needed before data analysis.

At first, an in house routine, based on the wavelet transform [47,48] and developed in Matlab, was used to reduce noise and correct baseline trend. In fact, instrumental noise can be considered as a high frequency contribution while the background as a low frequency one. Thus, each signal was decomposed by using the discrete wavelet transform (DWT) at decomposition level ten with a daubechies 5 wavelet filter. The approximation coefficients of the 10^{th} level were set to zero, since they account for baseline trend; whilst a thresholding of detail coefficients for all the levels was applied using a global threshold value obtained by a wavelet coefficients selection rule (Birgé-Massart penalization method [49]) on the basis of the standard deviation of first decomposition level detail coefficients [50]. Then, the XRD spectra were reconstructed applying inverse wavelet transform (IDWT).

Moreover, alignment of the signals resulted necessary to avoid discrimination between samples not imputable to real differences. In fact, shifts of the peaks were also noticed in replicated signals of the same samples and could be due to the not perfectly reproducible handmade loading of sample into the measuring cell and to the instrumental drift.

The alignment procedure was performed using the icoshift algorithm [51]; after a preliminary alignment of the whole spectrum (coshift procedure), an interval alignment (intervals chosen in order to obtain the best alignment) was carried out using one dataset signal as alignment target. The

target was selected among the spectra that the program suggested to be as more similar as possible to the medium signal.

Finally, a blockscaling procedure [52], called "block-adjusted non-scaled data", was used to allow peaks with minor intensity to contribute to the model without altering the relative scale of variables belonging to the same block.

The data were finally arranged in a 80×5200 matrix and analysed by means of PCA, in order to understand the relationships among samples and investigate the presence of trends or clusters. A 2 PCs model, with 78.26% of explained variance, was chosen.

PLEASE, INSERT HERE FIGURE 4

Looking at the scores plot (**Figure 4**), it is possible to see in-plain samples (producers A, B and D) well separated from hill samples (producer C) on the second principal component and grouped in three clusters on the base of the field of origin. These clusters are quite compact as regards B and D samples, while A samples present a slightly greater scatter.

On the other hand, hill samples differentiate along the first principal component; in particular, samples coming from holes 1 and 3, all sampling depths, are located at negative values of PC1, whilst all the samples from holes 4 and 5 are at positive values of PC1. Soil samples coming from sampling point 2 are divided in an upper part (a and b depths) at positive values of PC1 and a lower one (c, d and e depths) at negative values. This variability is highlighted along PC1, which explains most of the variance; thus, hill samples result to be characterised by a great complexity and heterogeneity, in agreement with texture analysis of soils.

PLEASE, INSERT HERE FIGURE 5

Observing the loadings plot (**Figure 5**) and considering a preliminary identification of diffraction peaks, it is possible to identify the phases that mainly influence the discrimination among different samples. The intra site variability of hill samples, shown on PC1, is mainly due to a different presence of quartz and calcite; while the distinction among hill and in-plain samples, highlighted on PC2, is probably caused by a different clay composition of the soils.

Summary

In this study, a suitable approach based on chemometrics techniques was proposed for the selection of soils sampling sites in a context of geographical traceability of food. In particular, D-Optimal Onion design was chosen since it is widely used for mapping and planning purposes and it allows to achieve the maximum coverage and uniformity of selected samples in the whole domain. An efficient mapping of inspected geographical area was obtained ensuring coverage of farms characterised by main production of investigated food and insisting on soils with different geological features.

Once identified the sampling sites, X-ray diffraction analysis coupled with chemometrics techniques were used and allowed to investigate the soil complexity and assess the proper number and type of samples to be collected for the extensive sampling of the investigated area. In particular, this study allowed observing the distinction between hill and in-plain samples.

For soil sampling plan, it is of utmost importance to take into account the complexity of hill samples, which present lot of variability as regard not only the different sampling points but sometimes also in the sampling depths. In fact, the hill region is characterised by mixture of soils with different origin and composition. On the other hand, in-plain soils result to be more homogeneous, in particular as regards the sampling depth.

On the strength of these results, only upper and lower fractions seem to be sufficient to describe in-plain soils, thus reducing the total number of samples to be further analysed.

Geographical traceability of raw materials for PDO and PGI oenological products.

Introduction

Aceto Balsamico Tradizionale di Modena (ABTM) and Aceto Balsamico di Modena (ABM) are two of the most well-known and appreciated Italian foods.

In 2000, ABTM received the PDO certification for its typical production and the well defined geographic area of production [43]. Briefly, it is obtained by acetic and alcoholic fermentation of cooked musts of selected grapes coming from restricted areas of Emilia Romagna indicated in the ABTM regulation. It ages at least 12 years in a series of casks of different woods and capacity. In order to compensate both annual spilling, i.e. for marketing, and natural evaporation occurring during the summer, a certain amount of vinegar is transferred from one cask to the next starting from the cask containing the oldest product by the so called *topping up procedure*. Besides ageing, the marketing of the product requires the approval from a Panel of Master Tasters, educated by a competent authority.

Aceto Balsamico di Modena (ABM) received the PGI certification in 2009 [53]. Its raw materials are mainly concentrated or cooked musts, wine vinegar and caramel and its production is regulated by the respective regulation.

In this study, the possibility to develop analytical models able to discriminate vinegar raw materials, namely concentrated or cooked musts, coming from Emilia Romagna was investigated.

In fact, the geographical provenance of musts is important in the case of ABTM, and could be an added value for the production of ABM.

For these purposes, 67 concentrated musts coming from grape juices of several geographical areas were investigated. **Table 2** reports some information about their geographical origin and grape varieties.

PLEASE, INSERT HERE TABLE 2

Each sample was analyzed by means of Near Infrared (NIR) spectroscopy and Attenuated Total Reflectance Mid Infrared (ATR-MIR) spectroscopy. In particular, NIR and MIR spectra were collected in the spectral regions from 10000 cm^{-1} to 4150 cm^{-1} (with a resolution of 4 cm^{-1}) and from 4000 cm^{-1} to 600 cm^{-1} (with a resolution of 4 cm^{-1}), respectively. These techniques are very common in food and raw materials characterization, since they are fast, non-destructive, low cost and do not require a preliminary sample preparation [54, 55]. Moreover, multivariate data analysis

is extensively and fruitfully applied on NIR and MIR signals with the aim to extract and visualise relevant information [56, 57].

When dealing with spectral data, a deeper understanding of the signals is necessary, since relevant information is often mixed with many uninformative sources of variation that may affect part or the whole signal domain. In this context, NIR and MIR spectra were pre-processed, in order to remove the variability sources due to both the physical nature of sample and environmental-experimental conditions of the measurement. In particular, a second derivative transform (Savitzky-Golay [58]) was applied to NIR signals in order to reduce parallel shifts and slope changes of the baseline. The raw and pre-processed NIR spectra are reported in **Figure 6a** and **6b,** respectively.

PLEASE, INSERT FIGURE 6a AND FIGURE 6b

A 2° order polynomial Savitzky-Golay smoothing followed by Standard Normal Variate (SNV) [59] was applied to MIR spectra to reduce baseline shift and instrumental noise (**Figure 7**).

PLEASE, INSERT FIGURE 7

Classification models based on Near Infrared (NIR) spectroscopy

A one class SIMCA model was evaluated in order to distinguish Emilia Romagna concentrated musts from the other ones. Emilia Romagna samples were split in training (31 samples) and test (15 samples) sets. Furthermore, test set includes samples coming from Apulia, Argentina and Spain (21 samples) in order to evaluate the specificity of SIMCA model.

Table 3 reports the number of chosen latent variables (LVs), explained variance of SIMCA model together with sensitivity (percentage of the objects belonging to the modelled class rightly accepted by the model) and specificity (percentage of the objects belonging to the other classes rightly rejected by the model) values for both training and test sets.

PLEASE, INSERT TABLE 3

As far as training set is concerned, the reported values highlight an optimal sensitivity of the model, since all samples coming from Emilia Romagna are correctly classified. On the other side, sensitivity of test set is not sufficient since only the 53 % of Emilia Romagna samples are well classified.

In order to improve the classification ability of the model, variables selection step was performed.

The selection of variables could separate relevant information from *unwanted* variability and at the same time allows data compression, i.e. more parsimonious models, simplification or improvement of model interpretation and so on. Although many approaches can be used for features selection, in this work, a wavelet-based supervised feature selection/classification algorithm, WPTER [12], was applied. The best performing model was obtained using a daubechies 10 wavelet, a maximum decomposition level equal to 10, between-class/within-class variance ratio criterion for the thresholding operation and the percentage of selected coefficients equal to 2%. Six wavelet coefficients were selected, belonging to the fourth, fifth, sixth, eighth and ninth levels of decomposition.

Few spectral regions were selected (figure not shown), around 12000-8800 cm^{-1} range, probably selected for background/baseline correction, 8700-7890 cm^{-1} and 7600-6260 cm^{-1}, where CH and OH overtones are present, respectively.

The performance of SIMCA model built on WPTER selected region, are reported in **Table 4**. With respect to SIMCA results, before feature selection, it is possible to note that both sensitivity and specificity calculated on test set are significantly improved.

PLEASE, INSERT TABLE 4

NIR spectra were also analysed by means of PLS-DA analysis. In this case, extra Emilia Romagna samples (class 2) were also split in training (15 samples) and test (6 samples) set due to the characteristics of PLS-DA algorithm.

Table 5 lists the sensitivity and specificity values for both training and test sets of Emilia-Romagna class. PLS-DA model shows good classification ability on training set, while the sensitivity toward the test set still remains unsatisfactory. Nevertheless, PLS-DA model seems to perform better than SIMCA, in particular, as regard specificity.

PLEASE, INSERT TABLE 5

Classification models based on Mid Infrared (MIR) spectroscopy

One class SIMCA model of Emilia Romagna was built using the same training and test sets of the previous NIR data analysis.

Specificity and sensitivity of model are reported in **Table 6**.

PLEASE, INSERT TABLE 6

Sensitivity and specificity of MIR based SIMCA model are slightly better than those obtained from NIR signals and the results still improve by application of PLS-DA analysis (**Table 7**)

PLEASE, INSERT TABLE 7

Summary

The synergistic use of chemometrics, Mid and Near infrared spectroscopy has been proved to be a valid tool for geographical traceability discrimination of concentrated musts. In particular, the two used classification methods, i.e. SIMCA, PLS-DA, provide slightly different results and **Table 8** shows a summary, in terms of sensitivity and specificity before and after features selection with WPTER algorithm, for Emilia Romagna class models.

PLEASE, INSERT TABLE 8

The different results between SIMCA and PLS-DA, are probably related to presence of high variability spectra regions common to all samples, independently from their different geographic origins. In this case, since SIMCA builds disjoint class models, this information leads to low specificity, while PLS-DA, being a discriminant technique, gives to this common regions low weights in the PLS regression model.

Food Authenticity applications

Study of grape juice heating process in a context of quality control of food

Introduction

Cooked grape must is the starting raw material for the production of ABTM, according to its production European Regulation [43]. It is obtained by heating grape juice in uncovered pans with direct fire. During this process, different chemical reactions can occur, mainly involving grape sugars such as glucose and fructose. Sugars show a complex chemical reactivity dependent from temperature of reaction, pH as well as the presence of oxidant and reducing agents. In particular, the heating of monosaccharides, in weak acidic media, leads to a multistep process with elimination of water and formation of furanic derivatives. In this work, multivariate data analysis techniques are used to extract useful information on critical steps of heating process in order to obtain cooked musts characterized by a low content of furfurals. Although the presence of furfurals in cooked must confers peculiar organoleptic characteristics, it might represent a negative aspect for a safety point of view. In fact, European Food Safety Authority (EFSA) established an acceptable daily intake, ADI, equals to 0.5 mg/kg of body weight for furfural [60,61].

In this study, nine cooking processes of grape juices coming from three different wine cellars were monitored. Must samples were regularly taken during the whole heating process, from the raw materials to the final product, thereafter named reduced cooked must.

PLEASE, INSERT TABLE 9

Table 9 summarizes some information on musts sampled for each cooking process (labelled from C1 to C9) of different wine cellar producers (kept anonymous, simply labelled as A, B and C). More details about the conduction of the heating process are reported in literature [62]. The total number of collected samples was 122.

For all samples, eleven parameters were monitored, namely, temperature (T), refractive index (n_D), density (d), total acidity (AcT), water content (H_2O), 5-(hydroxymethyl) furfural (5-HMF), furfural

(Furf), glucose (Glu), fructose (Fru), tartaric (Tar) and malic (Mal) acid concentrations referred to 'dry' matter.

Furfural and 5-(hydroxymethyl) furfural (5-HMF) were measured and considered as quality indicators. In particular, fructose and glucose undergo degradation phenomena involving furfurals formation, tartaric and malic acids are the main organic acids in grape; finally, the loss of water has to be monitored since it represents a critical step for furfurals formation. Refractive index, density and totally acidity, were measured in order to characterize the bulk of the system as well.

Sugar and acids content were determined with Gas Chromatographic technique (GC-FID, Varian 3400GC provided with a flame ionization detector). The analytical procedure was described in detail in literature [62,63]. Quantification was performed by means of the internal standard method and the calculation of response factor, by repeated injection of multiple standard solutions. Associated uncertainty and recovery of the method were calculated [62] too.

The determination of furfurals species was carried out by means of liquid chromatography (HPLC-DAD, Beckman system Gold apparatus, equipped with a single piston pump, injection valve, 20 mL sample loop and diode array detector). The analytical procedure was widely described in detail in previous work [62].

Principal component analysis (PCA) and parallel factor analysis (PARAFAC) were used to investigate the role of different technological parameters/cooking strategies adopted by different producers on the quality of final cooked musts.

Principal Component Analysis (PCA) results

As regards samples, looking at the list in **Table 9** (122 samples), it emerges that some samples, such as grape juices and must samples used for refilling, are common within some different cooking processes of the same producer. Hence, in the data set used for PCA, these common samples are repeated, due to the type of scaling adopted, obtaining a total of 132 samples. In particular, the different available variables for each process were separately arranged in a bi-dimensional matrix (samples × variables). Each subset of data was separately mean centered. After that, the mean centered data matrices were assembled into a unique one (132 rows × 11 columns) and successively autoscaled. In this way, it was possible to highlight the evolving trend inside each heating process, maintaining 'unchanged' the differences among the various cooking processes. A three-principal

components model was chosen taking into account the explained variance (R^2=87.18%) and the eigenvalues vs. number of components (SCREE plot not reported).

Figure 8 reports the scores plot of the first principal component and highlights the variability inside each cooking process as function of the heating time.

PLEASE, INSERT FIGURE 8

In particular, C2, C3 and C4 processes, with an heating over 40 hours, show PC1 scores varying of about 15 units, while the other ones (under 20 hours of heating) covering almost 4 units interval. Furthermore, PC1 orders samples according to number of hours of heating, assigning negative values to grape juices as well as intermediate cooked musts, and positive values to rather cooked musts and final products. The trends observed for C6÷C9 processes are less regular, probably due to refilling operation with partially cooked musts. Moreover, must samples used for the refilling procedures in the processes performed by producer C (i.e. R1, in cooking processes C6 and C7; R1 and R2 in cooking processes C8 and C9) have PC1 score values higher than the samples taken just before refilling. This could be explained considering the different capacity and technology of pans used in must cooking process with respect to the small pans where must samples used for refilling were heated. Finally, PC1 highlights a strange trend of C1 process, which, although the total heating time was around 43 hours, seems more similar to C6÷C9 processes. It was shown that this process had some problems with heating apparatus of the pan.

By analyzing PC1 loadings, **Figure 9**, it is possible to obtain information about the variables responsible of the trend shown in **Figure 8**.

PLEASE, INSERT FIGURE 9

In particular, it emerges that all variables, except the water content (H_2O%), present positive values. By the synergistic analysis of both the figures, it is possible to note that the more cooked samples of C2÷C4 processes are mainly discriminated from the other cooked musts for their lower water content and higher n_D, d, Total Acidity values and chemical components contents (*vice versa* for juices and intermediate samples).

The second and third principal component scores plot (figure not shown) mainly highlights the peculiar behaviour of some samples inside each heating process.

Parallel Factor Analysis (PARAFAC) results

Taking into account the trilinear structure of the data (**Figure 10**), characterized by three distinct sources of systematic variability, i.e. cooking time, monitored variables and different cooking processes, PARAFAC analysis was carried out as data display tool in order to highlight cooking processes similarity/ dissimilarity and the evolution of the different parameters during the cooking process.

In order to monitor processes, only the cooking times (hours) reported in **Figure 10** were considered. For C6÷C9, the final products were considered as cooked 20 hours. Hence, the data were arranged in a three-dimensional array, with samples cooked at different hours in the first mode, the monitored variables in the second mode and the cooking processes in the third one, obtaining a three-dimensional array of 9×11×9 dimensions. Samples used for the refilling are not considered in this analysis.

PLEASE, INSERT FIGURE 10

Before PARAFAC analysis, the data array was centered across the first mode (cooking processes), in order to remove the offsets, and scaled to unit variance within the second mode.

A two factors PARAFAC model showed a reasonable compromise of acceptable residual sum of squares, number of iterations, core consistency and reproducibility of the replicate PARAFAC runs.

The scores plot of the first two factors (F1 vs. F2) for the first mode, (**Figure 11a**), shows the evolution over the time of cooking process. In particular, the first factor distinguishes the first phase of cooking process (samples heated for 0 to 8 hours with negative score values) from the other ones (positive score values).

PLEASE, INSERT FIGURE 11a, 11b and 11c

Furthermore, reduced cooked musts, i.e. samples cooked at 20 hours, are clearly different from the others. The second factor, F2, mainly highlights the difference between the crude juice and other samples.

Loadings plot of the first two factors for the second mode (**Figure 11b**), i.e. variables mode, highlights the different trends of chemical-physical parameters: water content decreases with the evolution of the cooking process (it lies in the same region of crude juice in **Figure 11a**), while all the other compositional parameters, including furfurals, increase (they are in same area corresponding to final product in **Figure 11a**). Temperature is not influent in F1, and has the highest F2 value.

Finally, the loadings plot of the first two factors for the third mode (**Figure 11c**), i.e., cooking processes mode, mainly discriminates the processes according to their producers. The second factor, F2, mainly distinguishes C3 from the other ones.

For a synergistic combination of all the three loading plots, it emerges that the conduction of C3 and C2 processes with higher temperature (around 90°C) allows achieving cooked must with high concentration of chemical compounds, including furfurals compounds. Thus, it emerges that drastic conditions (high temperature values and extended time of cooking) strongly raise furfurals formation. In fact, sugar dehydration reaction occurs in a more pronounced and, in worse cases, uncontrolled way, if the loss of water takes place in presence of high sugars content and acid conditions.

Summary

In this study, it emerges how the use of explorative analysis, as PCA and PARAFAC, can really improve the visualization of the data and the interpretation of results, even in presence of variables of different nature.

This work is very important in an authenticity context, since it represents a first attempt to give to producers useful knowledge about the used heating process. This could allow obtaining products in full compliance with not only the traditional procedure but also with quality and safety assurance.

Study of sensory and compositional profiles during the ageing process of Aceto Balsamico Tradizionale di Modena (ABTM)

Introduction

In this application, it is investigated the possibility to develop a regression model in order to correlate the volatile fraction of different aged ABTM samples with sensory attributes [64]. In particular, the goal of the work is twofold: a) following the variation of both compositional and sensory variables with ageing and b) finding possible ageing markers.

To get an overall view of the product and the ageing process, different aged samples were analysed and many sources of variability such as different cask series, organic composition (i.e. volatile fraction), sensorial attributes and the judgment of several panellists were taken into account.

In particular, Partial Least Square Regression (PLS) and multi-way method, such as N-PLS, were applied to study the relationships between the volatile fractions, sampled and characterized by using HS-SPME/GC techniques, and the sensory attributes obtained by expert panellists [65,66].

The possibility to correctly predict the sensorial attributes on the basis of the GC signals is relevant for authentication tasks. 'Aroma' is one of the properties of this product mainly transformed during the ageing phase, due to microbial and chemical reactions. Thus, experimental data could be used to support the panel test evaluation, since the volatile organic compounds, which concur to the olfactory characteristics of ABTM, have a fundamental importance in the sensorial evaluation of the product.

Thirty-six ABTM samples were investigated, sampling six different series of casks coming from six different producers (each series comprising six casks of different capacity), **Table 10**.

PLEASE, INSERT TABLE 10

An expert panel consisting of eight panellists was used for the sensorial evaluation of all the samples, determining sixteen variables, namely four *visual*, five *smell/aroma*, six *texture* and one called *final sensation*.

The visual attributes are *density* (despite the name, it is linked to consistency/viscosity), *colour intensity* and *clearness* (established by looking at the transparency of the product through a candle

light).

Smell/aroma comprising four attributes namely, *frankness* defined as the absence of interferences in the perception of the flavour; *refinement,* which evaluates the presence of balsamic flavours, *intensity/persistency* related to noise perception of intensity and persistency of the balsamic flavours and *acidity*.

The five texture attributes are *fullness* which indicates how rich the sensation is perceived in the mouth, similar to '*corpo*' for wine; *intensity* of the taste; *texture* which is the quality of the perceived taste; *harmony* which evaluates the equilibrium between the acid and sweet tastes and *taste acidity*.

Final sensation represents a sort of overall perception of the sample received by the panellist.

Finally, for each class of attributes, the three partial averages were taken into account, as well.

The score attributed to each product is based on a total of 400 points distributed as follow: 15% to the visual attributes; 30% to the aroma attributes; 45% to the texture attributes and 40 points are reserved for *final sensation*. The visual scores range from 1 to 20 points, aroma from 1 to 30, texture from 1 to 36 and final sensation from 1 to 40. A detailed description of the evaluation method and panellists training has been reported in literature [65].

Afterwards, the ABTM volatile fraction was sampled by using a SPME fibre [67] and analyzed by Gas Chromatography with flame ionization detector (GC-FID), as reported in a previous work [64]. The chromatograms were acquired at constant sampling rate of 2.5 Hz, for a total time of 63 minutes (9451 points). A fibre blank was registered at the beginning of every session of experiments. The overall repeatability was evaluated by analyzing a control sample (vinegar sample) during each day of session.

Unfold-PLS results

As far as X-block is concerned, the data were organised in a bi-dimensional matrix by taking the ABTM samples coming from the six producers on rows and the GC-signals on columns, obtaining an X-matrix of 36×9451 dimensions. The data were mean centered.

The sensorial data were organized in a bi-dimensional matrix (Y-block) too, with the 36 samples on rows and the average scores of the 8 judges for each of the 16 sensory attributes on columns,

obtaining a 36×16 matrix dimension. The data were mean centered, since all the sensorial attributes have a similar range of variability.

The number of significant PLS latent variables (LVs) was chosen by considering two approaches: first, the commonly used leave one out cross validation (LOO) procedure [68] and the second called leave one producer out (LOP). In this latter approach, a producer at time was left out from the model and considered as test set. The procedure was done six times, obtaining six different root mean square errors in prediction (RMSEP-LOP).

PLS model, considering all the producers, was built with 5 LVs on the basis of the minimum RMSECV-LOO (9.23) as well as of the Y explained variance in fit (89.21%) and in cross validation (64.33%). Considering the overall RMSEP-LOP values, obtained separately for each producer as function of the sensory attributes (**Figure 12**), it is possible to obtain information about the prediction capability of the model. In particular, visual parameters (**Figure 12a**) show low error values but a lower variability (i.e. standard deviation) as well.

PLEASE, INSERT FIGURE 12a, 12b, 12c

On the contrary, aroma (**Figure 12b**) and texture parameters (**Figure 12c**) are better modelled since in general their associated RMSEP-LOP values are below the corresponding standard deviation. In particular, aroma values present a lower error range than texture ones. This is quite consistent with the nature of **X** data, since the ABTM volatile fraction is analysed.

From a synergistic analysis of all figures, the results highlight a good prediction for the 2nd and 5th producers; furthermore, their absence seems to decrease the robustness of the model increasing the RMSEP-LOO values.

Going on the plot of PLS regression coefficients, it is possible to obtain information about the chromatographic regions, which mainly contribute to the regression model performance. For the sake of clarity, only the plot of the PLS regression coefficients relative to the average of the *aroma* attributes was reported (**Figure 13**).

PLEASE, INSERT FIGURE 13 HERE

From this plot, it emerges that the model mainly keeps information from ethanol (Rt: 5.1 min), ethyl acetate (Rt: 9.8 min), acetic acid (Rt: 11.1) and furfurals (Rt: 21.2 min) compounds. From all the other parameters, almost the same chromatographic regions are important as well.

N-PLS results

As far as N-PLS case is concerned, the chromatographic data were firstly compressed by using principal component analysis (PCA). In particular, data were organised in a matrix of 36 (6 producers×6 casks) × 9451 dimensions and then compressed by singular value decomposition to 30 components, with 30 equal to the rank of matricized data. This reduction merely represents a computational shortcut to speed up calculations. The 30 derived score vectors were used instead of the original GC signals and data were rearranged into a 3-way matrix with producers in the first mode, PC-scores in the second mode and cask samples in the third one, yielding a 6×30×6 **X** array.

The choice of defining the producers as mode 1 is motivated by the applicability of the model. In fact, in the future, it is feasible to think to monitor/predict new producers.

As regards Y-array, the average of the eight judges' assessment for each sensorial attribute was considered. Thus, the data were arranged in a 3-way array with producers in the first mode, sensorial attributes in the second one and casks in the third one, obtaining a 6×16×6 **Y** array.

Both **X** and **Y** arrays were centered across the first mode, in analogy with the two way case.

Also in this case, root mean squares errors in leave one out cross (RMSECV-LOO) and in leave one producer out (RMSEP-LOP) validations were investigated.

The presence of some individual samples and/or variables, mainly influential for a given model, was evaluated through the calculation of Leverage [69]. Finally, the residual structure of **Y**-array was analyzed by inspecting the sum of squares of residuals (SSres) plot for each Y-mode.

The model was built with one latent variable (**Y** explained variance in cross validation, LOO, 84.38%). Analogously to the unfold-PLS case, the robustness and the predictive capability of the model was tested by Leaving One Producer Out (RMSEP-LOP) procedure. In all the six cases, the models built with one latent variable gave the best results in terms of lowest root mean squares cross-validation error. Considering the RMSEP-LOP values for each sensory parameter for the different models (data not reported), they have the same trend of the respective unfolding analysis with numerical values smaller than the previous one.

In order to get an overview of the variables, which mainly influence the prediction ability of the regression model, the loading matrix, corresponding to the second mode, was considered, back-transforming it into the original domain by multiplying it with the variable loading matrix of PCA decompression. In this way, it was possible to represent the loadings plot for GC-variable mode as function of the retention time. The same volatile compounds of the PLS case resulted to be relevant for predicting the sensorial parameters.

Leverage values for mode 1 and 3, ha and hc, and respective residual sum of squares (SSres), **Figure 14**, allow to gain combined information about the vinegar samples and/or producers which are well modeled (small SSres) and at the same time yield a unique contribution (high ha or hc).

PLEASE, INSERT FIGURE 14a, 14b and 14c HERE

In particular, from **Figure 14a and 14c**, 3rd, 4th and 5th producers have the largest ha and small residuals, hence contribute positively to the model, whilst the 3rd producer is influential but not well fitted. As far as the sample mode is concerned, **Figure 14b and 14d**, the youngest samples (number 6) is the most influential but it is not well fitted (high residuals values). On the other side, older samples (numbers 1 and 2) are not so influential in modelling the sensorial data but they are well fitted.

Summary

Both unfolded and three-dimensional regression models gave satisfactory predictions of almost all the sensorial parameters, which characterised the quality of ABTM samples. In particular, the aromatic parameters are better modelled, probably due to nature of the employed analytical data, i.e. volatile organic compounds.

Furthermore, both models highlighted a main contribution on the prediction ability given by chemical compounds characteristic of the fermentation and the bio-oxidation processes which take place in the younger casks (samples 5 and 6), such as ethanol, ethyl acetate, and acetic acid, on the other hand by compounds which are produced during the must cooking procedure (furfural).

The relative efficiency of both unfold-PLS and N-PLS has been compared in order to understand how the nature of data, i.e. the different source of variability could influence the performance of a

model. The use of N-PLS model led to a more parsimonious (1 LV) model and showed, in general, lower RMSECV values and a higher value of explained variance in prediction.

Notwithstanding, it has to be remarked that both methods allows a straightforward interpretation of the whole data set and about interrelations between sensorial and chemical parameters.

Characterisation and classification of Ligurian extra virgin olive oil

Introduction

Extra Virgin Olive Oil (EVOO) belongs to the superior olive oil category and it is solely obtained from fruit of olive tree, *Olea Europera*, by mechanical means.

Although the production of olive oil is concentrated in the Mediterranean countries area, the cultivation of olive trees is spreading in other many countries. The increasing consumption of EVOO is related to its peculiar properties, like seasoning of food as well as to its healthy benefits. Nevertheless, as result of agricultural traditions, local extraction and blending practices, EVOO may be quite different in taste and quality depending on its geographical origin with consequent differences in price within the same category. In the context of developing an analytical methodology able to assess the quality and authenticity of EVOO samples [20,70,71,72], this study was focused on the characterisation of the whole aroma fraction and on the development of analytical tools able to distinguish EVOO samples coming from 'Riviera Ligure-Riviera dei fiori' (Northern of Italy) [73] from other Mediterranean ones. Liguria EVOO, in which cultivar 'Taggiasca' prevails, was designed with PDO certification and represents one of the most highly esteemed and valuable European EVOO, since it is characterised by delicate, sweet, slightly pungent and green aroma. In particular, aroma is one of the most important parameters in the estimation of the quality of an EVOO sample, hence a great number of researches has been done to evaluate inter and intra relationships between the sensory notes and the concentration of the organic compounds.

In this framework, the analysis of the volatile fraction was performed using Head Space Solid Phase Micro Extraction (HS-SPME) [67] coupled with GC-MS system for the extraction and chromatographic separation and the identification of volatile organic compounds. The obtained GC-MS signals (Total Ion Current) were processed by SIMCA analysis.

EVOO samples with certified geographical origin, object of this study, were obtained from Consortium. They belong to different olive cultivars and come from different geographical areas, namely Liguria (Northern of Italy), Apulia (Southern of Italy), Greece, Tunisia and Spain. Since the main commercial interest is to distinguish Liguria EVOOs from the rest of olive oils, the data set was split in two classes, Liguria and Not Liguria oils. SIMCA analysis was applied in order to build a classification model able to rightly classify the samples belonging to a category and correctly to reject the other ones. Furthermore, as second step, the data set was divided in three classes too, namely Liguria, Apulia and Foreign samples.

Extra virgin olive oil samples

Seventy-two EVOO samples, produced from olives of different cultivars and harvested in 2003, were analysed by means of HS-SPME/GC-MS technique. They came from five different Mediterranean countries. In particular, 22 samples were from Apulia (mainly Ogliarola and Coratina cultivars), 21 from Liguria (mainly Taggiasca cultivar), 12 from Greece (Koroneika and Athinoia cultivars), 10 from Spain (Arbequina cultivar) and 7 samples from Tunisia (Chemlali cultivar).

The instrumental signals were arranged in bi-dimensional matrix with as many rows as samples (72) and as many columns as GC-points (1680) recorded during data acquisition (retention time, Rt: 72 min).
Before data analysis, the first 4.5 minutes and the last 7 ones of the signals were cut because there were no peaks at all, giving a GC vector of 1404 points for each sample.
In **Figure 15**, the average of the EVOO chromatograms for each of the five different geographical proveniences was reported.

PLEASE, INSERT FIGURE 15 HERE

In general, extracted volatile fraction includes a large number of hydrocarbons, aldehydes, alcohols, ketones, esters and other minor compounds. It is possible to observe that the average Liguria signal reports the lowest intensities of the different chromatographic peaks, followed by the Spanish one, perfectly in agreement with their much more delicate, sweet and slightly astringent flavour with respect to the other ones. In particular, the chromatographic peaks, which seem to characterise the volatile fraction of Liguria olive oil, are trans-2-hexenal (Rt: 34-34.5 min), and hexanal (Rt: 28-29 min), both C6 linear unsaturated aldehydes characteristics of high quality virgin olive oils.

As far as classification analysis is concerned, firstly, the 72 samples were split into two categories, Liguria and not Liguria (NL) samples, aiming to find a predictive classification rule to discriminate Liguria samples. Then, the samples were divided into three categories: Liguria, Apulia and Foreign (Greece, Spain and Tunisia) for investigating which category could mainly 'overlapping' with Liguria one.

The different data sets were separately and randomly split into training (for building a calibration model) and test sets (for validating it) as schematised in **Table 11**.

PLEASE, INSERT TABLE 11 HERE

The data matrices were separately mean centered and a model was built for each class (Liguria and Not Liguria and Liguria, Apulia, Foreign). The number of principal components (PCs) was chosen according to the best compromise among the minimum root mean square error in cross validation (RMSECV), sensibility and specificity of each model.

As first step, SIMCA was used to perform the classification among Liguria and Not Liguria classes. Twelve PCs were chosen for Not-Liguria class model, while three PCs for the Liguria one. The SIMCA results reported in **Table 12** highlight an excellent sensibility, meaning a right classification of both training and test sets, as well as an excellent specificity for Liguria samples.

PLEASE, INSERT TABLE 12 HERE

As concern Not Liguria class, almost all training set (except three samples) is well modelled. However, considering the test set samples, the sensibility of the model fairly decreases (68%). Eight Liguria samples fall inside the confidence limits of the model giving a very low specificity value (43%).

Afterwards, in the second step, as previously explained, samples were divided in three categories, Liguria, Apulia and Foreign (Greece, Spain and Tunisia). Internal and external cross validation were done following the same procedure of the previous case (random group and test set samples). Three PCs were retained for both Liguria and Apulia classes, while seven PCs gave the best results for the Foreign model. As far as Liguria class model is concerned, the results were unchanged with respect to the previous case. Indeed, all the training and test samples were in the confidence limits of the model (sensibility of 100%) and all the external samples, belonging to Apulia and Foreign classes, were rightly rejected (specificity of 100%). Furthermore, an improvement was evident in the performance of Apulia and Foreign classes, as well. In particular, Apulia class model was totally able to model both training and test set, showing a sensibility of 100%. Six Foreign samples and no

Liguria ones fall inside the Apulia model confidence limits, giving a relatively high specificity (83%).

Finally, as regards Foreign class, all the training set was inside the confidence limits and only two samples, belonging to test set, were outside (sensibility: 83%). At the same time, the model became more specific with respect to the previous step, since only four Liguria samples were not rejected by Foreign class model (specificity: 91%).

Summary

The aim to discriminate Liguria class from the other ones was successfully accomplished, showing as the use of HS-SPME/GC-MS technique coupled with chemometrics analysis can be an useful tool for the solution of such challenging classification problem in the control of food quality research.

References

[1] W. Van Rijswjik, L.J. Frewer, D. Menozzi and G. Faioli, Consumer perception of traceability: A cross-national comparison of the associated benefits. Food Qual. Prefer., 19 (2008) 452-464.

[2] ISO9000:2000 and ISO9000:2005, Quality management systems. (www.iso.org).

[3] J. Dennis, Recent developments in food authentication. Analyst, 123 (1998) 151-156.

[4] EC 2081/1992, Regulation (EC) n° 2081/92 of 14 July 1992.

[5] EC 2082/1992, Regulation (EC) n° 2082/92 of 14 July 1992.

[6] EC 178/2002 Regulation (EC) n° 178/2002 of 28 January2002.

[7] EC 510/2006, Regulation (EC) n° 510/2006 of 20 March 2006.

[8] EC 1898/2006, Regulation (EC) n° 1898/2006 of 14 December 2006.

[9] EC 628/2008, Regulation (EC) n° 628/2008 of 2 July 2008.

[10] TRACE: "Tracing Food Commodities in Europe", N°FP6-2003-FOOD-2-A 006942 (2005-2009).

[11] M. Lees (Ed.), Food Authenticity and traceability. Woodhead Publishing Ltd., Cambridge, 2003.

[12] M. Cocchi, C. Durante, G. Foca, D. Manzini, A. Marchetti, and A. Ulrici. Application of wavelet based algorithm on HS-SPME/GC signals for the classification of balsamic vinegars. Chemometr. Intell. Lab., 71 (2004) 129-140.

[13] S.D. Kelly, K. Heaton and J. Hoogewerff, Tracing the geographical origin of food: the application of multi-element and multi-isotope analysis. Trends Food Sci. Tech., 16 (2005) 555-567.

[14] X. Capron, J. Smeyers-Verbeke and D.L. Massart. Multivariate determination of the geographical origin of wines from different countries. Food Chem., 101 (2007) 1585-159.

[15] M. Cocchi, C. Durante, A. Marchetti, C. Armanino, M. Casale. Characterization and discrimination of different aged 'Aceto Balsamico Tradizionale di Modena' products by head space mass spectrometry and chemometrics. Anal. Chim. Acta, 589 (2007) 96-104.

[16] Da-Wen Sun (Ed.), Modern Techniques for Food Authentication. Academic Press is an imprint of Elsevier, 2008.

[17] T. Cajka, J. Hajslova, F. Pudil and K. Riddellova. Traceability of honey origin based on volatiles pattern processing by artificial neural networks. J. Chromatogr. A, 1216 (2009) 1458-1462.

[18] A. Gonzalvez, S. Armenta and M. de la Guardia. Trace-element composition and stable-isotope ratio for discrimination of foods with Protected Designation of Origin. Trends Anal. Chem., 28 (2009) 1295-1311.

[19] T. Woodcock, G. Downey and C. P. O'Donnell. Near infrared spectral fingerprinting for confirmation of claimed PDO provenance of honey. Food Chem., 114 (2009) 742-746.

[20] L. Mannina, F. Marini, M. Gobbino, A.P. Sobolev and D. Capitani. NMR and chemometrics in tracing European olive oils: the case study of Ligurian samples. Talanta , 80 (2010) 2141-2148.

[21] E. Mattarucchi, M. Stocchero, J.M. Moreno-Rojas, G. Giordano, F. Reniero, and C. Guillou. Authentication of Trappist Beers by LC-MS Fingerprints and Multivariate Data Analysis. J. Agr. Food Chem., 58 (2010) 12089-12095.

[22] A. Schellenberg, S. Chmielus, C. Schlicht, F. Camin, M. Perini et al. Multi-element stable isotope ratios (H,C,N,S) of honey from different European regions. Food Chem., 121 (2010) 770-777.

[23] D.G. Asfaha, C.R. Quètel, F. Thomas, M. Horacek, B. Wimmer et al. Combining isotopic signatures of n(^{87}Sr)/n(^{86}Sr) and light stable elements (C, N, O, S) with multi-elemental profiling for the authentication of provenance of European cereal samples. J. Cereal Sci. 53 (2011) 170-177.

[24] R.D. Di Paola-Naranjo, M.V. Baroni, N.S. Podio, H.R. Rubinstein, M.P. Fabani et al. Fingerprints for main varieties of Argentinean wines: Terroir differentiation by inorganic, organic, and stable isotopic analyses coupled to chemometrics. J. Agr. Food Chem., 59 (2011) 7854-7865.

[25] P. Oliveri and G. Downey, Multivariate class modeling for the verification of food-authenticity claims. Trends Anal. Chem., 35 (2012) 74-86.

[26] L. Bertacchini, C. Durante, A. Marchetti, S. Sighinolfi, M. Silvestri and M. Cocchi, Use of X-ray diffraction technique and chemometrics to aid soil sampling strategies in traceability studies. Talanta 98 (2012) 178-184.

[27] B. Peres, N. Barlet, G. Loiseau and D. Montet. Review of the current methods of analytical traceability allowing determination of the origin of foodstuffs. Food Control 18 (2007) 228–235.

[28] J. M. Juran, 'Inspectors' Errors in Quality Control' in Mechanical Engineering, originally presented at American Society of Mechanical Engineers (ASME) conference.

[29] http://europa.eu/pol/food/index_it.htm

[30] http://ec.europa.eu/dgs/health_consumer/library/pub/pub06_en.pdf

[31] http://europa.eu/documentation/official-docs/green-papers/index_en.htm

[32] F. Camin, L. Bontempo, K. Heinrich, M. Horacek, S. Kelly, C. Schlicht, F. Thomas, F. Monahan, J. Hoogewerff and A. Rossmann, Multi-element (H,C,N,S) stable isotope characteristics of lamb meat from different European regions, Anal. Bioanal. Chem. 389 (2007) 309-320.

[33] T. Cajkaa, K. Riddellovaa, E. Klimankovaa, M. Cernaa, F. Pudila and J. Hajslova, Traceability of olive oil based on volatiles pattern and multivariate analysis, Food Chem. 121 (2010) 282-289.

[34] N. Grošelj, G. van der Veer, M. Tušar, M. Vračko and M. Novič, Verification of the geological origin of bottled mineral water using artificial neural networks, Food Chem. 118 (2010) 941-947.

[35] I. Stanimirova, B. Üstün, T. Cajka, K. Riddelova, J. Hajslova, L.M.C. Buydens and B. Walczak, Tracing the geographical origin of honeys based on volatile compounds profiles assessment using pattern recognition techniques, Food Chem. 118 (2010) 171-176.

[36] C. Marisa, R. Almeida and M.T.S.D. Vasconcelos, Multi-element composition of wines and their precursors including provenance soil and their potentialities as fingerprints of wine origin, J. Agric. Food Chem. 51 (2003) 4788–4798.

[37] S. Swoboda, M. Brunner, S.F. Boulyga, P. Galler, M. Horacek and T. Prohaska, Identification of Marchfeld asparagus using Sr isotope ratio measurements by MC-ICP-MS, Anal. Bioanal. Chem. 390 (2008) 487-494.

[38] M. Brunner, R. Katona, Z. Stefànka and T. Prohaska, Determination of the geographical origin of processed spice using multielement and isotopic pattern on the example of Szegedi paprika, Eur. Food Res. Technol. 231 (2010) 623-634.

[39] Margesin, R., Schinner, F, Manual for soil analysis: monitoring and assessing soil bioremediation. 2005, Springer, Heidelberg, Germany, first edition.

[40] Food Authentication by chemical profiling. Trace booklet.

[41] K. Schlesier, C. Fauhl-Hassek, M. Forina, V. Cotea, E. Kocsi, R. Schoula, F. van Jaarsveld and R. Wittkowski, Characterization and determination of the geographical origin of wines. Part I: overview. Eur. Food Res. Technol., 230 (2009) 1-13.

[42] Production Regulation for "Lambrusco" wines, Decree of July, 27th 2009, published on the Italian Official Journal n° 187-188, August 2009.

[43] ABTM Production Regulation, EC 813/2000, Regulation (EC) n° 813/2000 of 17 April 2000.

[44] S. Totaro, P. Coratza, C. Durante, G. Foca, M. Li Vigni, A. Marchetti, M. Marchetti and M. Cocchi, Soils sampling planning in traceability studies by means of Experimental Design approaches. Submitted.

[45] L. Eriksson, E. Johansson, N. Kettaneh-Wold, C.Wikström and S. Wold, Design of Experiments - Principles and Applications, third ed., Umetrics Academy, Umeå Sweden, 2008.

[46] L.A. Sarabia and M.C. Ortiz in: S. Brown, R. Tauler and B. Walczak (Eds), Comprehensive Chemometrics, Chemical and Biochemical Data, Elsevier, vol 1, chpt 1.13, 2009, p 426.

[47] B. Walczak (Ed.), Wavelet in Chemistry, Elsevier, Amsterdam, NL, 2000.

[48] R.R. Coifman, Y. Meyer and M.V. Wickerhauser in: Y. Meyer and S. Roques (Eds.), Progress in Wavelet analysis and applications, Editions Frontieres, France, 1993.

[49] L. Birgé and P. Massart in: D. Polard (Ed.), Festschrift for L. Le Cam, Springer, Germany, 1997, pp 55-87.

[50] M. Misiti, Y. Misiti, G. Oppenheim and J.M. Poggi, Wavelet Toolbox™ 4 User's Guide, The MathWorks, Inc., Natick, MA, 2010.

[51] F. Savorani, G. Tomasi and S.B. Engelsen, icoshift: A versatile tool for the rapid alignment of 1D NMR spectra. J. Magn. Reson. 202 (2010) 190–202.

[52] S. Wold, E. Johansson and M. Cocchi in: H. Kubinyi (Ed.), 3D QSAR in Drug Design: Theory, Methods and Applications, ESCOM Science Publishers, Leiden, 1993.

[53] ABM Production Regulation, EC 583/2009, Regulation (EC) n° 583/2009 of 3 July 2009.

[54] P. Williams and K. Norris (eds.), Near-Infrared technology in the agricultural and food industries, 2nd edition. American Association of Cereal Chemists, Inc.Minnesota, USA, 2001.

[55] C.A. McGill, A. Nordon and D. Littlejohn, Comparison of in-line NIR, Raman and UV-visible spectrometries, and at-line NMR spectrometry for the monitoring of an esterification reaction. Analyst, 127 (2002) 287-292.

[56] B. Blanco and I. Villarroja, NIR spectroscopy: a rapid response analytical tool. Trends Anal. Chem., 21 (2002) 240-250.

[57] F. Cheli, D. Battaglia, L. Pinotti and A. Baldi, State of the Art in Feedstuff Analysis: A Technique-Oriented Perspective, J. Agric. Food Chem. 2012 60 (38) 9529–9542.

[58] A. Savitzky, M.J.E. Golay, Smoothing and differentiation of data by simplified least squares procedures, Anal. Chem., 36 (1964) 1627–1639.

[59] Barnes R.J., Dhanoa M.S., Lister S.J., Standard Normal Variate Transformation and De-trending of Near-Infrared Diffuse Reflectance Spectra, Applied Spectroscopy (1989) 43 772-777.

[60] Commission Regulation (EC) No 1565/2000 of 18 July 2000, Furfuryl and furan derivatives with and without additional side-chain substituents and heteroatoms from chemical group, 14 Adopted on 27 April 2005 The EFSA Journal, (2005), 215, 1-73.

[61] Commission Regulation (EC), question EFSA-Q-2003-236, Furfural and Furfural Diethyl-acetal, The EFSA Journal, (2004), 67, 1-27.

[62] M. Cocchi, R. Consonni, C. Durante, M. Grandi, S. Manzini, A. Marchetti and S. Sighinolfi, Changes in the chimica composition of reduced cooked musts during the heating process. J. Agric. Food Chem., 56 (2008) 6397-6407.

[63] M. Cocchi, P. Lambertini, D. Manzini, A. Marchetti and A. Ulrici, Determination of Carboxylic Acids in Vinegars and in Aceto Balsamico Tradizionale di Modena by HPLC and GC methods. J.Agric.Food Chem., 50 (2002) 5255-5261.

[64] C. Durante, M. Cocchi, M. Grandi, A. Marchetti and R. Bro, Application of N-PLS to gas chromatographic and sensory data of traditional balsamic vinegars of Modena. Chemometr. Intell. Lab., 83 (2006) 54-65.

[65] M. Cocchi, R. Bro, C. Durante, D. Manzini, A. Marchetti, F. Saccani, S. Sighinolfi and A. Ulrici, Analysis of sensory data of Aceto Balsamico Tradizionale di Modena (ABTM) of different ageing by application of PARAFAC models. Food Qual. Prefer., 17 (2006) 419-428.

[66] National Rule from 'Giunta Camerale', Doc. n° 162, 02/12/2002, List of Master Testers of ABTM.

[67] J. Pawliszyn, Solid Phase Microextraction: theory and practice. Wiley-VCH, New York, 1997.

[68] S. Wold, Cross-Validatory Estimation of the Number of Components in Factor and principal Components Models. Technometrics, 20 (1978) 397-405.

[69] A. Smilde, R. Bro and P. Geladi, Multi-way analysis, Application in the chemical sciences. Wiley, John Wiley & Sons, Ltd., England, 2004, pp171-173.

[70] A. Agiomyrgianaki, P.V. Petrakis and P. Dais, Influence of harvest year, cultivar and geographical origin on Greek extra virgin olive oils composition: A study by NMR spectroscopy and biometric analysis, Food chemistry 135 (4) (2012) 2561-2568.

[71] M. Bevilacqua, R. Bucci, A.D. Magri, and F. Marini, Tracing the origin of extra virgin olive oils by infrared spectroscopy and chemometrics: A case study, Anal. Chim. Acta, 717 (2012) 39-51.

[72]M. Casale, C. Casolino, P. Oliveri and M. Forina, The potential of coupling information using three analytical techniques for identifying the geographical origin of Liguria extra virgin olive oil, Food Chem. 118 (2010) 163-170.

[73] EC 123/1997, Regulation (EC) n° 123/97 of 23 January 1997.

Tables

Table 1. Summary of all the variables used for the selection of the representative farms for the two investigated areas.

Area A	**Area B**
Productivity variables 2 farm areas (numerical) 11 grape varieties (binary code) *Spatial coordinates* 2 geographical coordinates (numerical) *Geo-morphological variables* 1 parameter defining 6 classes (binary code) *Pedological variables* 6 parameters defining 15 classes (binary code) slope (numerical) texture (numerical) O_2 availability (numerical) alkalinity (numerical) Ca content (numerical) *Lithological variables* 1 parameter defining 4 classes (binary code)	*Productivity variables* 2 farm areas (numerical) 11 grape varieties (binary code) *Spatial coordinates* 2 geographical coordinates (numerical) *Geo-morphological variables* 1 parameter defining 13 classes (binary code) *Pedological variables* 1 parameter defining 16 classes (binary code)

Table 2. Scheme of available concentrated must samples, listed on the basis of their geographical origin and variety.

	Concentrated musts			
Geographic origin	**white**	**red**	**extra-red**	**rosé**
Emilia Romagna	17	15		14
Apulia	3	5		1
Argentina	3			
Spain		7	2	
Total	**67**			

Table 3. SIMCA model sensitivity and specificity values for both training and test sets for Emilia-Romagna class, based on NIR signals.

Class	**LVs**	**Explained variance (%)**		**SENSITIVITY (%)**	**SPECIFICITY (%)**
Emilia Romagna	10	99.76	Training set	100	--
			Test set	53	62

Table 4. SIMCA model sensitivity and specificity values for both training and test sets for Emilia-Romagna class after WPTER variable selection, based on NIR signals

Class	LVs	Explained variance (%)		SENSITIVITY (%)	SPECIFICITY (%)
Emilia Romagna			Training set	100	--
			Test set	71	72

Table 5. PLS-DA model sensitivity and specificity values for both training and test sets for Emilia-Romagna class, based on NIR signals.

Class	LVs	Explained variance (%)		SENSITIVITY (%)	SPECIFICITY (%)
Emilia Romagna	10	99.52	Training set	93	93
			Test set	60	71

Table 6. SIMCA model sensitivity and specificity values for both training and test sets for Emilia-Romagna class, based on MIR signals.

Class	LVs	Explained variance (%)		SENSITIVITY (%)	SPECIFICITY (%)
Emilia Romagna	8	99.49	Training set	100	--
			Test set	70	72

Table 7. PLS-DA model sensitivity and specificity values for both training and test sets for Emilia-Romagna class, based on MIR signals

Class	LVs	Explained variance (%)		SENSITIVITY (%)	SPECIFICITY (%)
Emilia Romagna	4	93.63	Training set	90	75
			Test set	94	89

Table 8. Summary of all sensitivity and specificity results obtained by the application of classification models on MIR and NIR spectra of the Emilia Romagna class. Values within brackets are relative to test set.

One class model				
Emilia-Romagna	**MIR data**	**SIMCA**	**PLSDA**	
	SENSITIVITY %	100 (71)	90 (94)	
	SPECIFICITY %	70	75 (89)	
	NIR data	**SIMCA**	**PLSDA**	**WPTER-SIMCA**
	SENSITIVITY %	100 (53)	93 (60)	100 (71)
	SPECIFICITY %	62	92 (71)	72

Table 9. Summary of available samples for each monitored cooking process. Samples are arranged in time order.

Producer	**A**				**B**	**C**			
Cooking process	**C1**	**C2**	**C3**	**C4**	**C5**	**C6**	**C7**	**C8**	**C9**
Time interval of sampling (hours of heating)	time 0 (same grape juice)				time 0	time 0 (same grape juice)		time 0 (same grape juice)	
	2	2	2	2	1	2	-	2	2
	4	4	4	4	2	4	4	4	4
	6	6	6	6	3	6	6	6	6
	8	8	8	8	4	1st refilling (sampling from small pan, labelled R1)		refilling (sampling from two small pans, labelled R1 and R2)	
	during the night no staff was at work in order to collect samples				5				
					6	6,5	6,5		
	20	20	20	20	7	8	8		
	-	refilling			8	8,75	8,75		
	22	-	22	-	9	2nd refilling (sampling from small pan, labelled R2)		6,75	6,75
	23	23	23	23	10			8	-
	24	-	24	-	11	9	9	10	10
	26	26	26	26	12	10	10	12	12
	28	-	28	-	13	12	12	14	14
	30	30	30	30	14	13,5	13,5	during the night no staff was at work in order to collect samples	
	32	-	32	-	15	3rd refilling (sampling from small pan, labelled R3)			
	35	35	35	35	16				
	38	-	38	-	18	14	14		
	41	41	41	41	19	night shift			
	43	43	43	43	20	22	22	22	22
N°of samples	*17*	*12*	*17*	*12*	*20*	*16*	*15*	*12*	*11*

Table 10. Capacities of wooden casks and starting years of each series

Series of Casks	Casks Capacity (litre)						Starting year
	#6	#5	#4	#3	#2	#1	
A	50	40	32	30	27	18	1975
B	50	35	35	30	25	20	1985
C	60	50	40	30	25	20	1990
D	30	24	20	16	13	10	1976
E	40	32	25	22	20	14	1978
F	70	50	40	30	20	20	1993

Table 11. Training and test set subdivision for each class data set

CLASS	Training set	Test set
Liguria	12	9
Apulia	12	10
Greece	8	4
Spain	5	5
Tunisia	4	3
TOTAL	**41**	**31**

Table 12. The best SIMCA models performance for both the classification steps

CLASS	N° of PCs	Explained Variance	RMSECV	Sensibility (training)	Sensibility (test)	Specificity
One class model						
Liguria	3	96.81%	3.93 106	100	100	100
Not-Liguria	12	98.12%	5.32 106	83	68	43
Three classes model						
Liguria	3	96.81%	3.93 106	100	100	100
Apulia	3	88.27%	5.24 106	100	100	83
Foreign	7	95.60%	3.96 106	100	83	91

Figure legends

Figure01: Plots of G-efficiency values as function of the different number of samples used in each layer, for (a) **A** and (b) **B** matrices. The stars (*) indicate the number of samples chosen for each layer.

Figure02: Scores plot, PC1 vs. PC2 vs. PC3, of **A** (in plane) matrix. Gray circles represent the 70 sites, selected by means of D-optimal Onion Design.

Figure03: Scores plot, PC1 vs. PC2 vs. PC3, of **B** (hill) matrix. Gray circles represent the 30 sites, selected by means of D-optimal Onion Design.

Figure04: Scores plot, PC1 vs. PC2, for the collected diffractograms. Symbols indicate different producers A, B, C and D. Labels identify the sampling points (numbers from 1 to 3 or from 1 to 5) and depths (letters from a to e).

Figure05: PC1 (black) and PC2 (gray) loadings plot vs. 2θ.

Figure06: Concentrated musts NIR spectra **(a)** before and **(b)** after preprocessing by second derivate tool.

Figure07: (a) Raw and (b) preprocessed ATM-MIR spectra.

Figure08: PC1 scores vs. samples. Samples are labelled according to their heating time. Refilling samples used for C6÷C9 processes are labelled as R1, R2 and R3 according to Table 9.

Figure09: PC1 loadings vs. variables measured for the samples of the different heating processes.

Figure10: Graphical representation of data array analyzed with PARAFAC analysis.

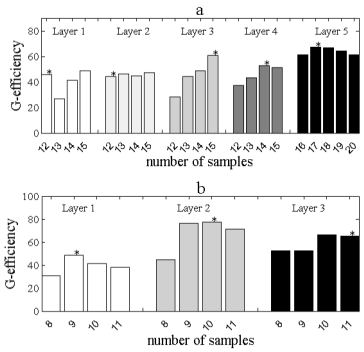
Figure11: F1 vs. F2 loadings plot for (a) the first mode (cooked juice samples), labels correspond to the number of heating hours, (b) the second mode (variables) and (c) the third mode (heating processes), different processes are labelled from C1 to C9.

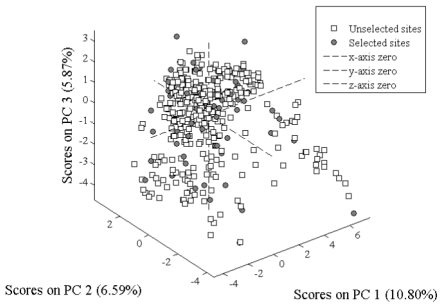
Figure12: Root mean squares errors leave one producer out (RMSEP-LOP) values, obtained separately for each producer as function of (a) visual, (b) aroma and (c) texture attributes.

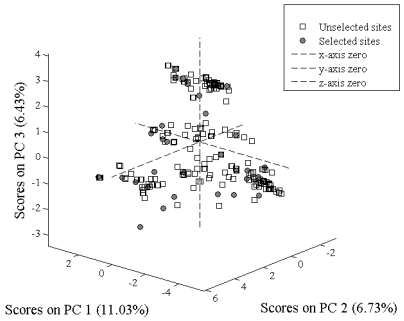
Figure13: PLS regression coefficients relative to the average of the *aroma* attributes.

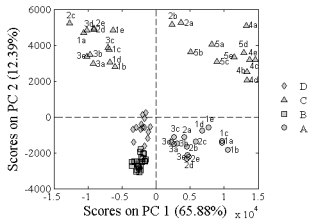
Figure14: Leverage values for (a) first and (b) third mode, ha and hc, respectively. Residual sum of squares (SSres) for (c) first and (d) third mode.

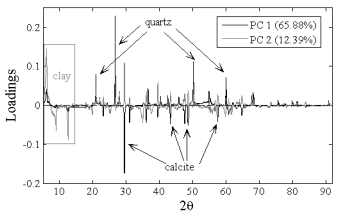
Figure15: GC profiles corresponding to the volatile fraction of Liguria (Northern of Italy), Apulia (Southern of Italy), Greece, Spain and Tunisia classes, obtained by separately averaging all the signals belonging to the each class.

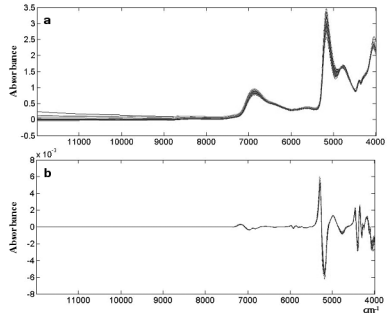

a
Layer 1
Layer 2
Layer 3
Layer 4
Layer 5
G-efficiency
number of samples
b
Layer 1
Layer 2
Layer 3
G-efficiency
number of samples

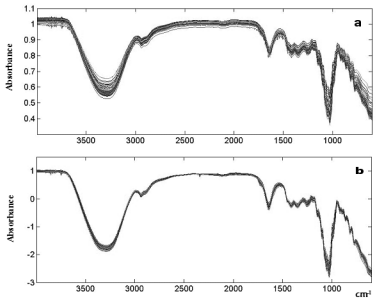

Unselected sites
Selected sites
x-axis zero
y-axis zero
z-axis zero
Scores on PC 3 (5.87%)
Scores on PC 2 (6.59%)
Scores on PC 1 (10.80%)

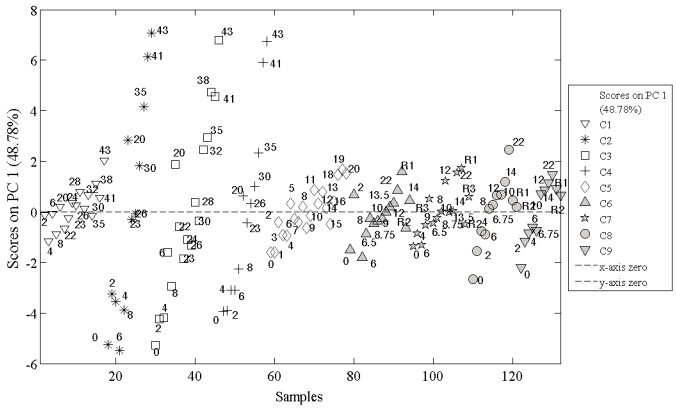

Unselected sites
Selected sites
x-axis zero
y-axis zero
z-axis zero
Scores on PC 3 (6.43%)
Scores on PC 1 (11.03%)
Scores on PC 2 (6.73%)

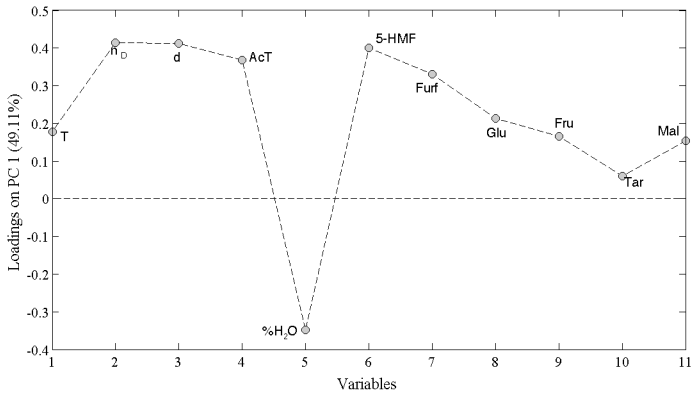

Scores on PC 2 (12.39%)
Scores on PC 1 (65.88%)
x 10^4
6000
4000
2000
0
-2000
-4000
-1.5
-1
-0.5
0
0.5
1
1.5
D
C
B
A

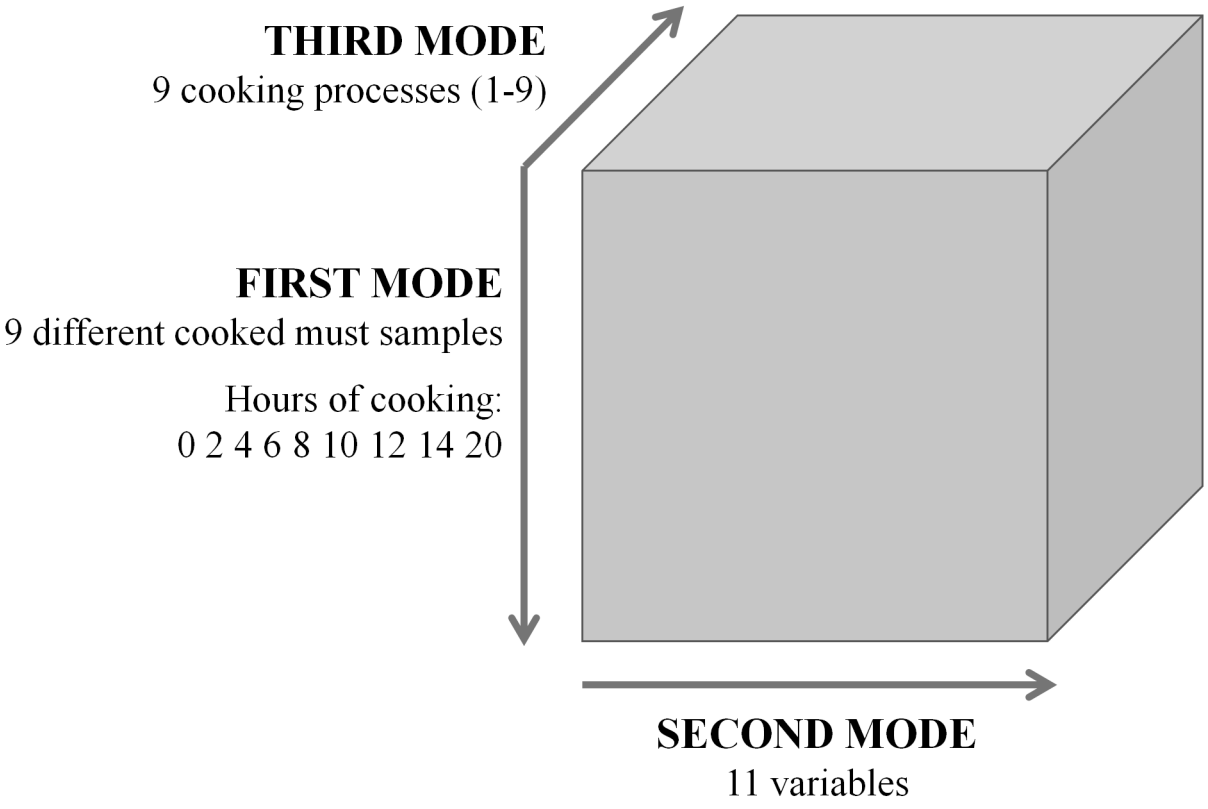

PC 1 (65.88%)
PC 2 (12.39%)
quartz
clay
calcite
Loadings
2θ
0.2
0.1
0
-0.1
-0.2
10
20
30
40
50
60
70
80
90

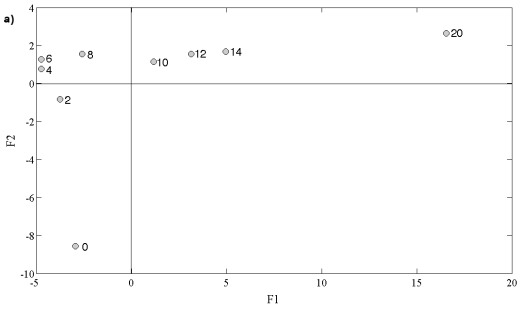

a
b
Absorbance
Absorbance
x 10^-3
11000
10000
9000
8000
7000
6000
5000
4000
cm-1

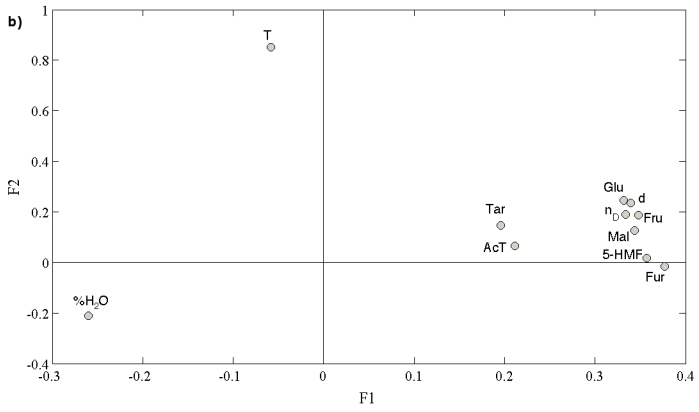

a
b
Absorbance
Absorbance
1.1
1
0.9
0.8
0.7
0.6
0.5
0.4
1
0
-1
-2
-3
3500
3000
2500
2000
1500
1000
cm-1

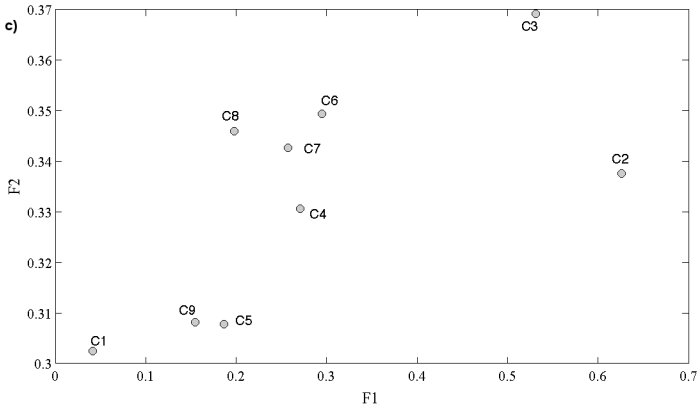

Scores on PC 1 (48.78%)
Samples
Scores on PC 1 (48.78%)
C1
C2
C3
C4
C5
C6
C7
C8
C9
x-axis zero
y-axis zero

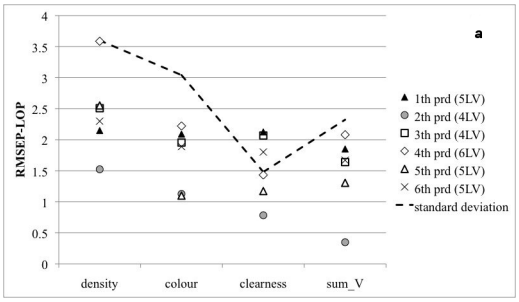

0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
1
2
3
4
5
6
7
8
9
10
11
Loadings on PC 1 (49.11%)
Variables
T
n_D
d
AcT
%H_2O
5-HMF
Furf
Glu
Fru
Tar
Mal

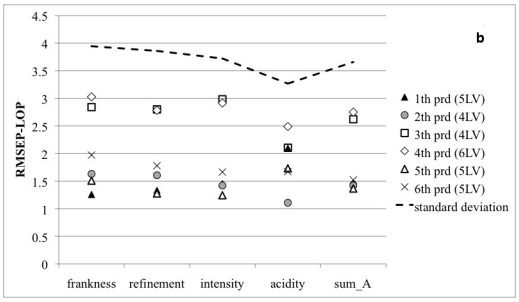

THIRD MODE
9 cooking processes (1-9)
FIRST MODE
9 different cooked must samples
Hours of cooking:
0 2 4 6 8 10 12 14 20
SECOND MODE
11 variables

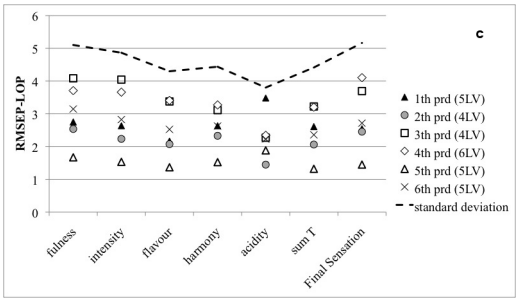

a)
20
6
8
10
12
14
4
2
0
F2
F1
4
2
0
-2
-4
-6
-8
-10
-5
0
5
10
15
20

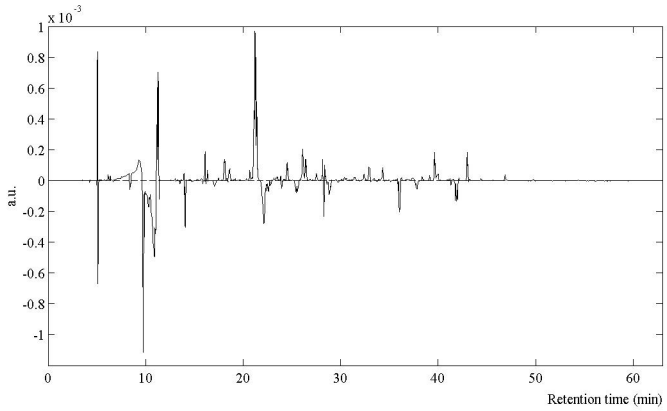

b)
T
Tar
AcT
Glu
d
n_D
Fru
Mal
5-HMF
Fur
%H_2O
F2
F1
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.3
-0.2
-0.1
0
0.1
0.2
0.3
0.4

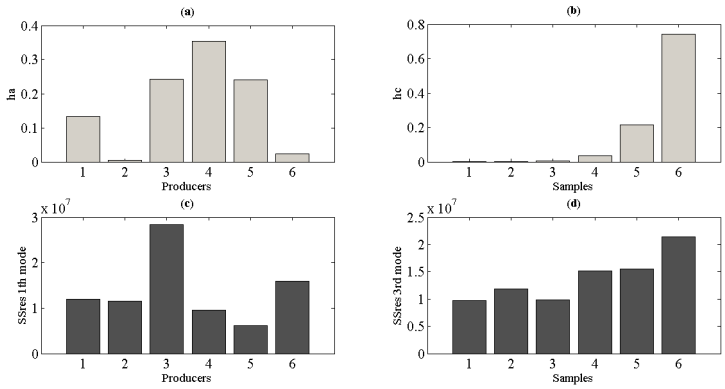

c)
C3
C6
C8
C7
C2
C4
C9
C5
C1
F2
F1
0.37
0.36
0.35
0.34
0.33
0.32
0.31
0.3
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7

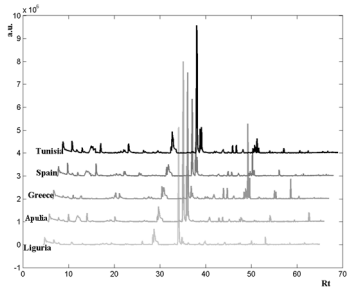

a
RMSEP-LOP
4
3.5
3
2.5
2
1.5
1
0.5
0
density
colour
clearness
sum_V
1th prd (5LV)
2th prd (4LV)
3th prd (4LV)
4th prd (6LV)
5th prd (5LV)
6th prd (5LV)
standard deviation


b
RMSEP-LOP
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
frankness
refinement
intensity
acidity
sum_A
1th prd (5LV)
2th prd (4LV)
3th prd (4LV)
4th prd (6LV)
5th prd (5LV)
6th prd (5LV)
standard deviation


c
RMSEP-LOP
6
5
4
3
2
1
0
fulness
intensity
flavour
harmony
acidity
sum T
Final Sensation
1th prd (5LV)
2th prd (4LV)
3th prd (4LV)
4th prd (6LV)
5th prd (5LV)
6th prd (5LV)
standard deviation


x 10^-3
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
a.u.
0
10
20
30
40
50
60
Retention time (min)


(a)
ha
Producers
(b)
hc
Samples
(c)
x 10^7
SSres 1th mode
Producers
(d)
x 10^7
SSres 3rd mode
Samples


x 10^7
a.u.
Tunisia
Spain
Greece
Apulia
Liguria
Rt