An overview of multivariate qualitative methods for food fraud detection

M. Pilar Callao, Itziar Ruisánchez*

Chemometrics, Qualimetric and Nanosensors Grup, Department of Analytical and Organic Chemistry, Rovira i Virgili University, Marcel·lí Domingo s/n, 43007 Tarragona, Spain

Abstract

Multivariate qualitative methods are an analytical strategy for addressing problems related to food fraud that cannot be solved with just one variable. Some examples are sample authentication since the required response is complex in nature and sample adulteration, when knowing the concentration of adulterant is not looked for. Establishing a multivariate qualitative method involves several steps: data collection, pre-treatment, exploration techniques, classification techniques, and method validation. When more than one data source is available, data fusion can be apply to improve the results of a single technique.

This review describes the state of the art of multivariate qualitative analysis for determining food fraud, and differentiates between authentication and adulteration. All the mentioned steps are discussed and, as example, recently published papers are commented.

Keywords: multivariate qualitative method, food authentication, food adulteration, classification, data fusion, validation

* Corresponding author: phone:+34 977558299; fax: +34 977558446; email: mariapilar.callao@urv.cat (M.P. Callao)

Highlights:

Multivariate qualitative methods can solve problems of food fraud.
Food authentication and food adulteration are discussed.
Classification techniques are the main tools for multivariate qualitative analysis.
Data fusion is a new perspective for improving classification results.
Research into multivariate qualitative method validation is still necessary.

1. Introduction

Qualitative methods are by no means new. Although they are not used in routine laboratory tasks as much as quantitative methods, they are currently on the rise and have been attracting increasingly greater interest, mainly for their screening potential.

Qualitative methods can be classified using several criteria but in all cases they are used in problems that require a binary response (yes/no). If response was achieved from multiple non-specific signals, a multivariate classification approach is required. These strategy is also referred as non-target analysis since the data set is used as a fingerprint of the sample.

According to the literature, multivariate qualitative methods are increasingly used in many fields (chemistry, process monitoring, etc.). Of course, multivariate classification is becoming increasingly important in food science too (Ballabio et al., 2009). In this paper, we focus more precisely on multivariate qualitative methods for problems of food fraud. In food fraud analysis, there are two main problems: a) authenticating the origin of a product in terms of geographical or botanical/animal provenance, or the manufacturing process, b) proving the absence of adulteration or the addition of a non-declared substance.

As far as product authentication is concerned, in many countries there are laws that require agricultural products to have information about their geographical origin on the labels. The EU has encouraged the use of labelling to identify products by introducing regulations, first in 1992 and more recently in 2006 (EU regulations 510/2006, 509/2009 and 1898/2006). Those regulations define the following geographical indications for food products: protected designation of origin (PDO), protected geographical indication (PGI) and traditional specialities guaranteed (TSG). The use of geographical indications implies market recognition and it is related to the price of the product. To solve the problem of

authentication, the response required is qualitative; that is, binary (yes / no; belongs / does not belong, etc.). However, a single signal often cannot solve the problem, so a multivariate approach is usually required.

The second problem, food adulteration, is attracting increasing attention because it is an emerging risk, given the complex and global nature of food supply chains. One of the major concerns about adulteration is that it may involve a health risk or economic benefit. Food adulteration problems can be solved in two ways: if the adulterant is known, a quantitative analysis is usually carried out but, if it is not, a qualitative analysis (it is or it is not adulterated) may be satisfactory.

A bibliographic search of the last five years shows how keywords such as “food authentication” or “food adulteration” and “classification” were increasingly found in scientific articles. They mainly refer to the use of classification techniques with a multivariate signal provided by different instrumental techniques. Recently, several reviews have been published on specific instrumental techniques that are used with a chemometric approach for food analysis (Bosque-Sendra et al., 2012, Domingo et al., 2014, Casale et al. 2014, Danezis et al. 2016), the use of chemometric techniques for specific food analysis (Camiña et al., 2012, Domingo et al., 2014, Esslinger et al., 2014, Haddi et al., 2015; Nascimento et al. 2017, Kamal et al., 2015), or the metabolomic analysis of food (Cubero-Leon, et al., 2014).

This overview focused on the development of multivariate qualitative methods for the detection of food fraud. Figure 1 schematically presents an overall protocol for this purpose. It should be noted that the analytical determinations that give rise to the data set are mainly instrumental measures that provide multiple data for each sample analysed (i.e. absorbance at different wavelengths), although they can also be independent measures from different techniques (i.e. pH, conductivity, etc.). The former are more common, because the experimental cost is very small.

The paper has been divided into sections that correspond to the different steps implemented in a multivariate qualitative method. Section 2 (exploratory analysis) and section 3 (classification techniques) are the steps that have been studied most, so the main characteristics of the different approaches will be commented. Section 4 (data fusion) is the step more recently introduced in multivariate qualitative analysis. Section 5 focuses on the

validation step. Recently some studies (Lopez, M.I. et al., 2015; Riedl, J. et al., 2015) deals with it, although further research is required to develop unified protocols. In each section, the chemometric techniques are briefly described, although for more in-depth explanations the reader is addressed to the basic bibliography.

2. Exploratory analysis

The exploratory or unsupervised analysis provide information about the relationship between samples, between variables and/or between samples and variables. Various tools can be used and their theoretical basis has been well explained in many scientific articles and recent books on chemometrics (Esbensen et al., 2009, Li Vigni et al., 2013).

Information about the relationship between samples reveals whether there are natural groups or trends in sample distribution that are consistent with prior knowledge about them. For example, if a strategy is established for detecting authentication and both authentic and non authentic samples are submitted to an unsupervised analysis, they should present a distribution that shows some tendencies. If there are not tendencies, the characterization of the samples must be not adequate and the experimentation carried out must be redefined. In addition, unsupervised techniques make it possible to detect the presence of possible outliers: i.e. samples distributed differently and separate from the main group. These samples should be rejected as they can have a negative impact on the use of supervised techniques.

The relationship between variables shows which of them give complementary information and which give similar or redundant information. On the other hand the relationship between samples and variables indicates which variables are important (and which are not) for distinguishing groups of samples. This type of information can be valuable to simplify the database or, in some cases, to reduce experimentation.

The most popular unsupervised exploratory technique is based on the well-known principal components analysis (PCA) (Esbensen et al., 2009, Li Vigni et al., 2013). PCA generates new variables as a linear combination of the original variables. These new variables retain maximum information from the original data matrix and are called principal components (PCs). The first PC is the one that retains most explained variance (more data information)

while the second PC explains the information that is not modelled by the first PCs, and so on. When it is used as exploratory technique, the information from the two or three first PC's are plot. So, sample and variable distribution are showed. Its main limitation is when the first PC´s do not contain enough information.

Other exploratory techniques are cluster analysis (CA) (Lee et al., 2009), in which samples (or variables) are linked to others according to their similarity. Groups considering similarity values are defined. The main limitation of this technique is that it does not show the overall relationship between all the samples but only between the ones that are close together. Neither does it give any information about the relationship between samples and variables. On the other hand, it uses all the information contained in the data and can be considered to complement the PCA representation.

As Table 1 shows, most authentication or adulteration studies use the PCA technique before applying a classification technique. Some studies also use cluster analysis techniques (Mir-Marqués et al., 2016, Azevedo, M.S. et al. 2017).

Some of the studies reviewed only present a PCA exploratory analysis, and interpret both the scores and the loading plot (Malheiro et al., 2013, Boggia et al., 2013, Üçüncüoğlu et al., 2013, Dahimi et al., 2014). For instance, PCA was used in the study of six fresh wild mushroom species for taxonomical and authentication purposes (Malheiro et al., 2013). The authors used the loading plot to identify the volatile secondary metabolites (11 volatile compounds out of forty-six) that characterize each mushroom species and which have highest power of discrimination. These compounds seem to play a crucial biomarker role in the characterization of the six wild species of mushrooms.

Similarly, a screening method was proposed to detect pomegranate juice adulteration by the addition of cheaper fruit juices (i.e., grape and apple juices) or by dilution (Boggia et al., 2013). PCA was performed as a preliminary data examination, and the score plots showed a satisfactory separation among the various juice categories. By analysing the loadings, the authors once again determined which variables were the most important for separating the various mixture compositions. In particular, PC1 points to dilution while both of the first two PCs point to the use of filler juice.

There are some works that only analyses the score plot. As an example, (Üçüncüoğlu et al., 2013) the authors talk about the best PCA model, when in fact, they did not build a model at all; they just used the PC1/PC2 score plot to check whether the test samples were close to the predefined sample classes (butter, adulterated butter and margarine). It should be pointed out that unsupervised pattern recognition methods, such as PCA and cluster analysis must not be confused with classification methods (supervised pattern recognition).

3. Classification techniques

The type of binary response required by qualitative analysis (yes/no, belongs/does not belong, etc.) can be obtained by applying a classification technique. These techniques require classes (or categories) to be defined. Each class consists of a set of samples with a common property (i.e authentic sample) and different from the other classes (i.e. non authentic sample). All samples, from different classes, must be characterized by the same variables and then a classification rule is set. The final goal is to individually assign a unknown sample characterized by the same variables to one (or none) of the predefined classes.

The classification techniques can be divided into two main blocks. One block is discriminant analysis (also referred to as 'hard modelling'), which aims to divide data space up into separate regions, each of which corresponds to one class. The other main block focuses on class-modelling analysis (also known as 'soft modelling'), which models each class independently (Marini, 2010).

The main discriminant techniques are: linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA); k nearest neighbours (KNN) and partial least squares discriminant analysis (PLS-DA). The main modelling techniques are: soft independent modelling of class analogy (SIMCA) and unequal dispersed classes (UNEQ). The theoretical basis of these techniques has been well explained in many scientific articles and recent books on chemometrics (Bevilacqua et al., 2013). Other techniques with more restricted applications are: support vector machines (SVM) (Majcher et al., 2015, Mu et al., 2016,), density-based methods (potential functions) (Oliveri et al., 2014) and artificial neural networks (Mu et al., 2016). In addition, some applications use techniques similar to the ones mentioned above (with the same theoretical basis) but with a slight variation.

Discriminant techniques required at least to classes and classify unknown samples in the closest class (i.e. they are assigned to the class whose characteristics are most similar). Because all samples are assigned to a class, it is more difficult to detect outliers. Modelling techniques classify samples into just one class, in more than one or in none, so the result is sometimes ambiguous or inconclusive. In particular, class-modelling techniques make it possible to model only one class. One-class modelling is thus more useful when samples of only one class are available because it is impossible to cover all other areas (this may be the case of an authentication problem in which it is easy to characterize the "authentic" class, but the chance of samples being "non authentic" is almost impossible to cover all).

Table 1 summarizes a series of studies that focus on classification techniques. They have been chosen in an attempt to be representative of the subject of this review and cover both authentication and adulteration applications.

In food authentication problems, most of the studies revised deal with multi-category classification strategy. The class of interest and other classes that can potentially lead to fraud are defined. As examples, three classes (Benicarló, Valencia and Murcia) were defined to authenticate artichokes from a protected designation of origin (Mir-Marqués et al., 2016). To differentiate rice varieties, eight classes were defined six types of rice and two types of wild rice (Zhu et al., 2015). To identify the geographical origin of sea cucumber, seven classes were stablished corresponding to seven areas in northern China (Zhang, X. Et al., 2017).

Other studies use a two-category classification strategy (Bevilacqua et al., 2012, Chiesa, L. et al. 2016), which determines whether a sample is authentic – i.e., it comes from a particular brand, it was produced in a specific place-origin or with specific raw materials, it complies to what declared in the label, etc. – or not. By way of example, to authenticate samples of extra virgin olive oil from the PDO area of Sabina, one class was defined with samples from Sabina and another class with samples from other origins (other areas of Italy or Mediterranean countries) (Bevilacqua et al., 2012). Few cases have been found in which one-class-model is used in authentication problems (Oliveri, P. et al., 2014, Zhang, L. et al; 2015).

In food adulteration problems, there are two main approaches. The two-class approach is implemented when the adulterant is known. Therefore, one class is defined for the adulterated samples and another for the unadulterated samples. Examples of the two-class strategy are the evaluation of contamination and degradation in infant formula (Inoue et al., 2015) and the discrimination between authentic beefburgers and beefburgers adulterated with offal (Zhao et al., 2014). Various papers have used this strategy (López et al., 2014a, Xu et al., 2013a, Di Anibal et al., 2015).

The one-class approach is implemented when the adulterant is not known and only the unadulterated class is defined. Although the one-class approach is not new, its application has recently been increase. It has been employed to detect melamine adulteration in milk (Chen et al., 2017), and distinguish a range of adulterants in kudzu starch, including four cheaper plant starches – namely, sweet potato, potato, maize and cassava starches – and a commonly used illegal whitening agent, talcum powder (Xu et al., 2015). One study analyses and compares the two strategies on the adulteration of hazelnut paste (López et al., 2014b).

In some situations, it is known that more than one adulterant can be found in a sample. In these cases, a multi-class strategy is followed, in which, as well as the unadulterated class, there are as many other classes as adulterants. As example of this strategy, a class modelling approach is implemented to detect five common adulterants in raw milk. So, six classes are defined –unadulterated, hydrogen peroxide, sodium citrate, sodium carbonate, formaldehyde and starch – (De Souza Godim, 2017a).

Although the main objective in some papers (Fadzlillah et al., 2013, Zhao et al., 2015, Mu et al., 2016, Santos et al., 2016) was qualitative in nature (i.e. to determine if a sample was adulterated or not) once the adulteration detection system had been developed, a multivariate regression method was also developed to determine the concentration of the adulterant.

In recent works (Georgouli et al. 2017 and Amiry et al. 2017) the number of classes is established according to the adulterant concentration. This approach involves addressing a quantitative problem with tools of the qualitative multivariate analysis.

Focusing on the instrumental techniques used, the most common ones are spectroscopic. Within the field of spectroscopy, one of the most widely used in the food industry is infrared

spectroscopy in its different regions (NIR, MIR, FTIR). Their advantages are that can analyse samples with little or no preparation, it is easy to use, it collects data quickly and it can be used as a fingerprint technique. Other spectroscopic techniques that are used quite often are ultra-violet (UV-Vis) (Sen et al., 2016, Boggia et al., 2013), fluorescence (Di Anibal et al., 2015, Mir-Marqués et al., 2016, Mu et al., 2016), Raman (Üçüncüoğlu et al., 2013, Zhao et al., 2015) and nuclear magnetic resonance (NMR) (Santos et al., 2016). To a lesser extent, element techniques such as inductively coupled plasma atomic emission spectrometry (ICP) (Ortea et al., 2015, Mir-Marqués et al., 2016), parameters such as the colour index (Sen et al., 2016) and isotope-ratio mass spectrometry (IRMS) (Ortea et al., 2015) also appear in the referenced bibliography. More recently, chromatographic techniques – mainly gas chromatography (Malheiro, R. et al., 2013) with or without mass spectrometric detection – have been applied. Taking into account that nowadays many laboratories have a variety of analytical equipment, and they can obtain the instrumental signal quickly and easily, most of the studies (table 1) analysed more than one instrumental technique for a specific problem when spectroscopy data were used.

As can be seen in table 1, the most common chemometric approaches use SIMCA and PLS-DA classification techniques or some variation. SIMCA is a modelling classification technique in which each class is modelled independently from all others, in such a way that it can be applied to any strategy (from one-class to multi-class). In addition, information about the modelling power and discriminating power of variables can be obtained. On the other hand, PLS-DA is a discrimination technique based on the PLS regression technique adapted to a supervised classification task. Therefore, more than one class has to be defined (two-class or multi-class) and samples are always assigned to one class. Recently, a variation of the technique – one-class partial least squares (OCPLS) – has been developed for the one-class approach, although very few papers can be found. By way of example, OCPLS was used to detect adulterations in whole milk powder (Xu et al., 2013b, Chen et al. 2017) and in starch (Xu et al., 2015).

The choice of the most appropriate classification technique depends on many factors (class criteria definition, homogeneous sample distribution, number of input variables, number of samples, etc.). Therefore, it is common practice to apply more than one classification technique and evaluate their goodness for the problem under study. It should also be borne in mind that once the problem has been properly defined and samples characterized by

variables (samples are analysed), applying more than one classification technique has a minimum experimental cost.

A wide variety of samples and analytes have been studied. Essentially, food authenticity involves conforming to the description provided by the producer or processor. So, any food (processed/natural) is susceptible to fraud in terms of their label specifications (geographic origin, PDO, etc.). In food adulteration problems, the options are more numerous because of the wide variety of food types and ingredients (compositional change by adding/subtracting, sample dilution, etc.). In most cases, food is adulterated for economic reasons, and the adulterant can be known in advance, i.e. spices are adulterated with forbidden Sudan dyes (Di Anibal et al., 2015).

4. Data fusion

At times, some problems can only be solved by using extra instrumental techniques that provide complementary information. Data fusion, is an approach to obtain a single result from more than one source. There are three types of data fusion: low-, mid- and high-level data fusion. The basis of each one are described in literature (Borràs et al., 2015, Marquez et al, 2016).

Table 2 summarizes a series of studies about data fusion strategies in various food and quality control processes. Most of the applications addressed authentication problems and, to a lesser extent, adulteration.

Initially, most of the applications involve fusing data blocks from two complementary techniques, but recently the fusion of data from three (Alamprese et al., 2013, Ulloa et al., 2013, Erich et al., 2015, Forina el at., 2015, Borràs et al. 2016), four (Erich et al. 2015) and even five techniques (Biancolillo et al., 2014) has been described. In most cases, at least one of the fused techniques was spectroscopic, mainly IR vibrational spectroscopy (MIR, NIR) and, to a lesser extent Raman, NMR, UV-Vis and fluorescence. Also in most cases, fusion was done with physical–chemical parameters (Pizarro et al., 2013, Nunes et al., 2016), or indexes (Ottavian et al., 2014, Chen et al., 2014) or without spectroscopic techniques, i.e. sensors (Chen et al., 2014, Haddi et al., 2014), electronic-tongue (Ulloa et al., 2013, Teye et al. 2015), isotope ratios (Monakhova et al., 2014, Erich et al., 2015), liquid chromatography (Bajoub et al. 2017, Obisesan et al. 2017), etc

All the papers reviewed compare the results obtained with the data fusion strategy with the ones obtained independently for each data block, and in almost all cases data fusion was superior. Just one case, the data fusion did not sufficiently improve the results obtained by a single technique (HS-MS) to classify one out of the six pre-defined classes (Borràs et al., 2016).

Most of them focus on mid-level data fusion and compare it to low-level data fusion. The comparison shows that mid-level data fusion generally gave better ability of classification than low-level data fusion. Just one paper (Nunes et al., 2016) reports better results with low-level data fusion. It should be taken into account that it is not feasible to implement low-level data fusion when there are a very high number of variables if the chosen classification technique don't allow to deal with. For instance, the high number of input variables prevented the LDA classification method from being used (Pizarro et al., 2013). Authors that implemented LDA with spectroscopic data, usually worked with the scores of the PCA (Pizarro et al., 2013, Erich et al., 2015, Forina et al., 2015). In this context, it should be pointed out that only low-level data fusion was implemented in two cases in which each data block had few variables. Five variables from a data block of tin oxide-based Taguchi Gas Sensors were fused with six variables from a data block of potentiometric sensors (Haddi et al., 2014). The amount of twelve rare earth elements were fused with the amount of fifteen trace elements for yellow split pea authentication (Drivelos et al., 2014).

Dealing with very high dimensionality data makes it mandatory to select or reduce variables, so mid-level data fusion is the one to be chosen. To select the variables, there is quite a variety of methodologies, ranging from very simple ones such as the Fisher criterion (Ni et al., 2012, Alamprese et al., 2013) analysis of the variance (ANOVA) (Monakhova et al., 2014, Erich et al., 2015) and basic statistics (Márquez et al., 2016) to more complex ones like stepwise decorrelation (Forina et al. 2015), wavelet transform (Wenjuan et al. 2017) and interval PLS (Wenjuan et al. 2017, Obisesan et al., 2017). To reduce, or compress, variables quite simple methodologies based on index calculations (Ulloa, et al, 2013, Chen et al., 2014, Ottavian et al., 2014), and scores of the principal component decomposition (PCA) (Pizarro et al., 2013, Ulloa et al., 2013, Silvestri et al., 2014, Teye et al., 2015, Borràs et al. 2016, Obisesan et al 2017) or of PLS decomposition (Biancolillo et al., 2014, Spiteri et al., 2016, Nunes et al., 2016, Borràs et al. 2016, Bajoub et al. 2017) can be used alongside more complex ones such as the clustering of latent variables (CLV) (Monakhova et al., 2014,

Erich et al., 2015), PARAFAC loadings and the peak areas of MCR resolved components (Silvestri et al., 2014).

Two studies (Márquez et al., 2016, Obisesan et al., 2017) compare the results of mid- and high-level data fusion. One feature of high–level fusion is that classification models do not have to be developed with exactly the same samples. This gives additional flexibility to high-level fusion.

Of the three levels of data fusion, low- and mid-level are the most commonly used. The choice between low- and mid-level, is mainly dependent on the number of variables to be fused. The main drawback of low-level data fusion is that the increase in information obtained by adding one or more blocks of data to describe the sample may not compensate for amount of irrelevant or spurious variance brought by the addition of the same blocks. When the number of variables is high, mid-level is the recommended one. The comparison shows that, mid-level data fusion generally has better classification abilities than high-level data fusion and high-level fusion is better than low-level.

5. Multivariate qualitative method validation

Nowadays, the validation protocols for qualitative methods are poorly developed. The main reference is the Commission Decision CD/657/EC, 2002. From it, efforts are being made to standardize guidelines and terminology (López et al., 2015). Figure 2 shows a proposal of the steps to be followed in the validation of multivariate qualitative methods. In addition, the performance parameters are indicated considering whether the model is for quantifiable or categorical sample property.

To validate a method, a series of samples which are known to belong (or not) to the pre-defined class/es are used. When it is possible, the data set is divided into training and test set considering that the division has to be representative in each class. Among several possibilities, randomly (Silvestri et al., 2014, Oliveri et al., 2014, Teye et al., 2015, Erich et al. 2015, Borràs et al. 2016, Obisesan et al., 2017), Kennard-stone algorithm (Ottavian et al., 2014, Nunes et al 2016, Wenjuan et al. 2017) or duplex (Biancolillo, et al., 2014, Silvestri et al., 2014), are the most implemented. It has to be emphasized that the number of objects used to build a classification model is often critical, since few objects cannot represent all the factors involved in class variability.

386 An alternative, is to use the whole data set as the training set using the cross-validation
387 strategy. Cross-validation can be carried out through several strategies: contiguous blocks,
388 leave-one-out, random subsets, cancelation groups, venetian blinds, among others.

389 The output obtained when a sample is predicted is: belongs / does not belong to the class
390 considered. Therefore, in comparison to its authentic membership, the result could be: true
391 positive, false positive, true negative, false negative and inconclusive (not assigned or
392 assigned to more than one class) (López et al., 2015, De Souza Godim, 2017a).

393 Generally speaking, almost all the referenced papers validate (performance parameter
394 estimation) in terms of assignation ability (or error), which give the ratio of properly (or
395 wrongly) assigned samples for each class. (López et al., 2015).

396 When the classification problem is to differentiate or discriminate among two or more
397 categories, ability – or error – are calculated for each category and considering the whole
398 data set without the categories (global ability). When they are calculated from the training
399 set (classification abilities) could be too optimistic and sometimes seriously misleading since
400 they are autopredictive. When they are calculated from the test set (prediction abilities) are
401 more reliable for assessing the model quality.

402 In some of the reviewed papers, multivariate performance parameters (either global or for a
403 category) were also expressed as sensitivity and specificity values (Monakhova et al., 2014,
404 Nunes et al. 2016, Borràs et al. 2016, Bajoub 2017). The sensitivity of a model is the
405 percentage of the objects of a class accepted by the class model. The specificity is the
406 percentage of the objects of the categories different from the modelled one rejected by the
407 class model.

408 It should be pointed out that sensitivity and specificity are closely related to ability values. If
409 the one-class approach is used, these parameters are the same, but when at least two
410 classes are modelled, these parameters are related but they are not strictly the same if some
411 samples are classified to none of the categories, or to more than one category (inconclusive
412 assignations) (Lopez et al., 2015).

413 Some authors (López et al., 2014, Drivelos et al., 2014, Perez-Castaño et al., 2015) also
414 present other related performance parameters – for example, Youden's index, likelihood
415 ratio, efficiency, discriminant power, etc. – as a way of characterizing a qualitative
416 multivariate model. In these three references, two categories are modelled (i.e. A and B)

417 with the classification results being positive (assigned to class A) and negative (assigned to
418 class B). Therefore, these parameters were calculated from the contingency table obtained
419 from the classification results.

420 Finally, how other performance parameters, such as CCα (decision limit), CCβ (detection
421 capability), unreliability region, etc., which may be of interest in adulteration problems, are
422 estimated is still not well established for multivariate qualitative methods. CCα is the
423 concentration limit at which the qualitative method detects the contaminant (it is present)
424 with a α error of stating that the contaminant is present when in fact it is not (false non-
425 compliant decision or false positive result). CCβ is the concentration limit at which the
426 qualitative method detects the contaminant (it is present) with a β error of stating that the
427 contaminant is not present when in fact it is (false compliant decision or false negative
428 result). The unreliability region is defined by the two limits CCα and CCβ. To estimate these
429 parameters in multivariate methods, some authors (López et al., 2014, De Souza et al.,
430 2017b) propose the use of probability of detection (POD) curves, well known in univariate
431 qualitative methods.

432

433 **6. Conclusions**

434 Multivariate qualitative methods are a good option for addressing problems of food fraud
435 that cannot be solved with just one variable, either because the required response is
436 complex in nature or because no single signal acts as an unambiguous marker. For food
437 authentication, they are the only option and for food adulteration they are recommended
438 when the adulterant is not known.

439 The steps for conducting a multivariate qualitative analysis are well established and
440 documented in the literature, although research is still being carried out in an attempt to seek
441 improvements, either by experimenting with new data sources or developing new
442 algorithms.

443 The authentication and assessing non adulteration of foodstuff will benefit from advances in
444 data fusion and the synergic information obtained from more than one technique. Since
445 laboratories nowadays have a variety of analytical equipment, any data fusion strategy is a
446 feasible way of dealing with qualitative analysis. Combining information from different

447 instrumental sources can improve the results but, depending on the problem and on the
448 maximum permitted error, the improvement has to be carefully evaluated in term of cost-
449 benefit ratio. However, although spectroscopic measurements (most used) are economical,
450 measuring by more than one technique represents an additional cost.

451 The validation stage still needs to be developed further and, in our opinion, this is where
452 research efforts ought to lie. Validation involves establishing a set of measurable attributes
453 (performance parameters) that define the method´s quality. Quantitative methods have been
454 the subject of numerous studies, which have resulted in the production of international
455 guidelines. By contrast, there is still no consensus about the validation protocol and the
456 terminology used for qualitative methods. Such basic performance parameters as sensitivity
457 and specificity are already being used but others like robustness, stability, detection limits
458 and the unreliability region still require a great deal of work to be done.

459

460 **7. References**

461 Alamprese, C., Casale, M., Sinelli, N., Lanteri, S., & Casiraghi, E. (2013). Detection of
462 minced beef adulteration with turkey meat by UV-vis, NIR and MIR spectroscopy. *LWT -*
463 *Food Science and Technology*, *53*, 225-232.
464
465 Amiry,S, Esmaiili, M., & Alizadeh, M. (2017) Classification of adulterated honeys by
466 multivariate analysis. *Food Chemistry, 224*, 390-397.
467
468 Azevedo, M. S., Seraglio S.K.T., Rocha, G., Balderas C.B., Piovezan M., Gonzaga, L. V.,
469 Falkenberg, D.B., Fett, R., de Oliveira, M.A.L., & Costa, A.C.O. (2017). Free amino acid
470 determination by GC-MS combined with a chemometric approach for geographical
471 classification of bracatinga honeydew honey (Mimosa scabrella Bentham). *Food Control,*
472 *78*, 383-392.
473
474 Bajoub, A., Medina-Rodríguez, S., Gómez-Romero, M., Ajal, E.A., Bagur-González, M.G.,
475 Fernández-Gutiérrez, A., & Carrasco-Pancorbo, A. (2017). Assessing the varietal origin of
476 extra-virgin olive oil using liquid chromatography fingerprints of phenolic compound, data
477 fusion and chemometrics. *Food Chemistry, 215*, 245–255.

478

Ballabio, D., & Todeschini, R. (2009). Infrared Spectroscopy for Food Quality Analysis and control. In Da-Wen Sun (Eds.), Multivariate Classification for Qualitative Analysis. (Chapter 4, pp. 83-102). Amsterdam: Elsevier.

Binett, G., Del Coco, L., Ragone, R, Zelasco, S., Perri, E., Montemurro, C., Valentini, R., Naso, D., Fanizzi, F.P.b, Schena, F.P. (2017) Cultivar classification of Apulian olive oils: Use of artificial neural networks for comparing NMR, NIR and merceological data. *Food Chemistry, 219*, 131-138.

Bevilacqua, M., Bucci, B., Magrì, A. D., Magrì, A. L., & Marini, F. (2012). Tracing the origin of extra virgin olive oils by infrared spectroscopy and chemometrics: A case study. *Analytica Chimica Acta, 717*, 39-51.

Bevilacqua, M., Bucci, R., Magrì, A. D., Magrì, A. L., Nescatelli, R., & Marini, F. (2013). Data Handling in Science and Technology Volume *28*, 1st Edition. In F. Marini, (Eds.), *Classification and Class-Modelling* (pp. 175-233). Amsterdam: Elsevier.

Biancolillo, A., Bucci, R., Magrì, A. L., Magrì, A.D., & Marini, F. (2014). Data-fusion for multiplatform characterization of an italian craft beer aimed at its authentication. *Analytica Chimica Acta, 820*, 23-31.

Boggia, R., Casolino, M.C., Hysenaj, V., Oliveri, P., & Zunin, P. (2013). A screening method based on UV–Visible spectroscopy and multivariate analysis to assess addition of filler juices and water to pomegranate juices. *Food Chemistry, 140*, 735-741.

Bona, E., Marquetti, I., Link J. V., Makimori, G.Y.F., Arca V.C., Lemes, A.L.G., Ferreira J.M.G., Scholz, M.B., Valderrama, P., & Poppi, R.J. (2017). Support vector machines in tandem with infrared spectroscopy for geographical classification of green arabica coffee. LWT_*Food Science and Technology, part B, 76,* 330-336.

Borràs, E., Ferré, J., Boqué, R., Mestres, M., Aceña, L., & Busto, O. (2015). Data fusion methodologies for food and beverage authentication and quality assessment –A review. *Analytica Chimica Acta, 891*, 1-14.

511 Borràs, E., Ferré, J., Boqué, R., Mestres, M., Aceña, L., Calvo, A., & Busto, O. (2016). Olive
512 oil sensory defects classification with data fusion of instrumental techniques and multivariate
513 analysis (PLS-DA). *Food Chemistry, 203,* 314–322.
514
515 Bosque-Sendra, J.M., Cuadros-Rodriguez, L., Ruiz-Samblas, C., & De la Mata, A.P. (2012).
516 Combining chromatography and chemometrics for the characterization and authentication
517 of fats and oils from triacylglycerol compositional data—A review. *Analytica Chimica Acta,*
518 *724,* 1-11.
519
520 Camiña, J.M., Pellerano, R.G., & Marchevsky, E.J. (2012). Geographical and Botanical
521 Classification of Honeys and Apicultural Products by Chemometric Methods. A Review.
522 *Current Analytical Chemistry*, *8,* 408-425.
523
524 Casale, M., & Simonetti, R. (2014). Review: Near infrared spectroscopy for analyzing olive
525 oils. *J. Near Infrared Spectrosc. 22,* 59–80.
526
527 Chen, Q., Sun, C., Ouyang, Q., Liu, A., Li, H., & Zhao, J. (2014). Classification of vinegar
528 with different marked ages using olfactory sensors and gustatory sensors. *Analytical*
529 *Methods, 6,* 9783-90.
530
531 Chen, H., Ten, C., Lin, Z., & Wu, T. (2017). Detection of melamine adulteration in milk by
532 near-infrared spectroscopy and one-class partial least squares. *Spectrochimica Acta Part*
533 *A: Molecular and Biomolecular Spectroscopy, 173*, 832–836.
534
535 Chiesa, L., Panseri, S., Bonacci, S. , Procopio, A., Zecconi, A., Arioli, F. Cuevas, F.J. &
536 Moreno-Rojas, J.M. (2016). Authentication of Italian PDO lard using NIR spectroscopy,
537 volatile profile and fatty acid composition combined with chemometrics. *Food Chemistry,*
538 *212,* 296-304.
539
540 Commission Regulation (EC) No 509/2009 of 16 June 2009 establishing the standard import
541 values for determining the entry price of certain fruit and vegetables.
542

Commission Regulation (EC) No 1898/2006 of 14 December 2006 laying down detailed rules of implementation of Council Regulation (EC) No 510/2006 on the protection of geographical indications and designations of origin for agricultural products and foodstuffs.

Council Regulation (EC) No 510/2006 of 20 March 2006 on the protection of geographical indications and designations of origin for agricultural products and foodstuffs.

Cubero-Leon, E. Peñalver R., & Maquet, A. (2014). Review on metabolomics for food authentication. *Food Research International, 60,* 95–107.

Dahimi, O., Rahim, A.A., Abdulkarim, S.M., Hassan, M.S., Hashari, S.B., Mashitoh, A.S., & Saadi, S. (2014). Multivariate statistical analysis treatment of DSC thermal properties for animal fat adulteration. *Food Chemistry, 158,* 132-138.

Danezis, G.P., Tsagkaris, A.S., Camin, F., Brusic, V.R., & Georgiou, C.A. (2016). Food authentication: Techniques, trends & emerging approaches ,*Trends in Analytical Chemistry-TrAC,* 85, 123-132.

Di Anibal, C.V., Rodríguez, M.S., & Albertengo, L. (2015). Synchronous fluorescence and multivariate classification analysis as a screening tool for determining Sudan I dye in culinary spices. *Food Control, 56,* 18-23.

Domingo, E., Tirelli, A.A., Nunes, C.A., Guerreiro, M.C., & Pinto, S.M. (2014). Melamine detection in milk using vibrational spectroscopy and chemometrics analysis: A review. *Food Research International, 60,* 131-139.

Drivelos, S.A., Higgins, K., Kalivas, J.H., Haroutounian, S.A., & Georgiou, C.A. (2014). Data fusion for food authentication. Combining rare earth elements and trace metals to discriminate “Fava Santorinis” from other yellow split peas using chemometric tools. *Food Chemistry, 165,* 316–322.

EC, 2002. Commission Decision 2002/657/EC of 12 August 2002 implementing Council Directive 96/23/EC concerning the performance of analytical methods and the interpretation of results. Official Journal of European Communities, L221, 17.08.2002, 8-36.

577

578 Erich, S., Schill, S., Annweiler, E., Waiblinger, H.U., Kuballa, T., Lachenmeier, D.W., &
579 Monakhova, Y.B. (2015). Combined chemometric analysis of 1H NMR, 13C NMR and stable
580 isotope data to differentiate organic and conventional milk. *Food Chemistry, 188*, 1-7.
581

582 Esbensen K.H., & Geladi P. (2009). Comprehensive chemometrics: chemical and
583 biochemical data analysis. In S. Brown, R. Tauler, & B. Walczak, (Eds.), *Principal*
584 *component analysis: concept, geometrical interpretation, Mathematical background,*
585 *algorithms, history, practice.* (vol. 2, pp. 211–27, Chapter 2.13). Amsterdam: Elsevier Ltd.
586

587 Esslinger, S., Riedl, J., & Fauhl-Hassek, C. (2014). Potential and limitations of non-targeted
588 fingerprinting for authentication of food in official control. *Food Research International, 60,*
589 189-204.
590

591 Fadzlillah, N. A., Rohman, A., Ismail, A., Mustafa, S., & Khatib, A. (2013). Application of
592 FTIR-ATR spectroscopy coupled with multivariate analysis for rapid estimation of butter
593 adulteration. *Journal of Oleo Science, 62*, 555-62.

594 Forina, M., Oliveri, P., Bagnasco, L., Simonetti, R., Casolino, M.C., NizziGrifi, F., & Casale,
595 M. (2015). Artificial nose, NIR and UV–visible spectroscopy for the characterisation of the
596 PDO Chianti Classico olive oil. *Talanta, 144,* 1070–1078
597

598 Georgouli, K:, Martinez del Rincon, J. & Koidis, A. (2017) Continuous statistical modelling
599 for rapid detection of adulteration of extra virgin olive oil using mid infrared and Raman
600 spectroscopic data *Food Chemistry, 217,* 735-742.
601

602 Godim, C.S., Junqueira, R.G., Souza, S.V.C., Ruisánchez, I., & Callao, M.P. (2017a).
603 Detection of several common adulterants in raw milk by MID-Infrared spectroscopy and one-
604 class and multiclass multivariate strategies. *Food Chemistry, 230,* 68–75.
605

606 Godim, C.S., Junqueira, R.G., Souza, S.V.C., Callao, M.P., & Ruisánchez, I. (2017b).
607 Determining performance parameters in qualitative multivariate methods using probability
608 of detection (POD) curves. Case study: two common milk adulterants. *Talanta*, *168,* 23–30.
609

610 Haddi, Z., Mabrouk, S., Bougrinia, M., Tahri, M., Sghaier, K., Barhoumi, H., El Bari, N.,
611 Maaref, A., Jaffrezic-Renault, N., & Bouchikhi, B. (2014). E-Nose and e-Tongue combination
612 for improved recognition of fruit juice samples. *Food Chemistry, 150*, 246-253.
613
614 Inoue, K., Tanada, C., Sakamoto, T., Tsutsui, H., Akiba, T., Zhe, J., Todoroki, K., Yamano,
615 Y., & Toyo'oka, T. (2015). Metabolomics approach of infant formula for the evaluation of
616 contamination and degradation using hydrophilic interaction liquid chromatography coupled
617 with mass spectrometry. *Food Chemistry, 181,* 318-324.
618
619 Jiménez-Carvelo, A.M., Pérez-Castaño, E., González-Casado, A., Cuadros-Rodríguez, L.
620 (2017) One input-class and two input-class classifications for differentiating olive oil from
621 other edible vegetable oils by use of the normal-phase liquid chromatography fingerprint of
622 the methyl-transesterified fraction *Food Chemistry, 221,* 1784-1791.
623
624 Kalogiouri, N.P., Alygizakis N.A., Aalizadeh, R. & Thomaidis, N.S. (2016) Olive oil
625 authenticity studies by target and nontarget LC–QTOF-MS combined with advanced
626 chemometric techniques, *Analytical and Bioanalytical Chemistry 408*, 7955–7970.
627
628 Kamal, M., & Karoui, R. (2015). Analytical methods coupled with chemometric tools for
629 determining the authenticity and detecting the adulteration of dairy products: A review.
630 *Trends in food Science and Technology, 46*, 27-48.
631
632 Lee, I., & Yang, J. (2009). Comprehensive chemometrics: chemical and biochemical data
633 analysis In: S. Brown, R. Tauler, & B. Walczak, (Eds.), *Common clustering algorithms.* (vol.
634 2, pp. 211–27, Chapter 2.17). Amsterdam: Elsevier Ltd.
635
636 Li Vigni, M., Durante, C., & Cocchi, M. (2013). Data Handling in Science and Technology:
637 Chemometrics in Food Chemistry. In F. Marini, (Eds.), *Exploratory Data Analysis* (volume
638 28, pp. 55-126). Amsterdam: Elsevier B.V.
639
640 López, M.I., Colomer, N., Ruisánchez, I., & Callao, M.P. (2014a). Validation of multivariate
641 screening methodology. Case study: Detection of food fraud. *Analytica Chimica Acta, 827,*
642 28-33.
643

López, M.I., Trullols, E., Callao, M.P., & Ruisánchez, I. (2014b). Multivariate screening in food adulteration: Untargeted versus targeted modelling. *Food Chemistry, 147,* 177-181.

López, M.I., Callao, M.P., & Ruisánchez, I. (2015). A tutorial on the validation of qualitative methods: From the univariate to the multivariate approach. *Analytica Chimica Acta, 891,* 62-72.

Majcher, M. A., Kaczmarek, A., Klensporf-Pawlik, D., Pikul, J., & Jeleń, H. H. (2015). SPME-MS-Based electronic nose as a tool for determination of authenticity of PDO cheese, oscypek. *Food Analytical Methods, 8,* 2211-2217.

Malheiro, R., Pinho, P.G., Soares, S., Ferreira, A.C.S., & Baptista, P. (2013). Volatile biomarkers for wild mushrooms species discrimination. *Food Research International, 54,* 186-194.

Maia, M., & Nunes, F.M. (2013). Authentication of beeswax (Apis mellifera) by high-temperature gas chromatography and chemometric analysis, *Food Chemistry, 136,* 961-968.

Marini F. (2010). Classification methods in chemometrics. *Current Analytical Chemistry, 6*, 72-79.

Márquez, C., López, M.I., Ruisánchez, I., & Callao, M.P. (2016). FT-Raman and NIR spectroscopy data fusion strategy for multivariate qualitative analysis of food fraud. *Talanta, 161,* 80-86.

Mir-Marqués, A., Elvira-Sáez, C., Cervera, M.L., Garrigues, S., & De la Guardia, M. (2016). Authentication of protected designation of origin artichokes by spectroscopy methods. *Food Control, 59*, 74-81.

Monakhova, Y. B., Godelmann, R., Hermann, A., Kuballa, T., Cannet, C., Schäfer, H., Spraul, M., & Rutledge, D. N. (2014). Synergistic effect of the simultaneous chemometric analysis of ^{1}H NMR spectroscopic and stable isotope (SNIF-NMR, ^{18}O, ^{13}C) data: Application to wine analysis. *Analytica Chimica Acta, 833*, 29-39.

Mu, T., Chen, S., Zhang, Y., Chen, h., Guo, P., & Meng, F. (2016). Portable Detection and Quantification of Olive Oil Adulteration by 473-nm Laser-Induced Fluorescence. *Food Analytical Methods, 9,* 275-279.

Nascimento C.F., Santos P. M, Pereira-Filho E.R., & Rocha F.R.P. (2017). Recent advances on determination of milk adulterants. *Food Chemistry 221,* 1232-1244.

Ni, Y., Mei, M., & Kokot, S. (2012). One- and two-dimensional GC–MS and HPLC–DAD fingerprints of complex substances: A comparison of classification performance of similar, complex Rhizoma Curcumae samples with the aid of chemometrics. *Analytica Chimica Acta, 712,* 37-44.

Nunes, K.M., Andrade, M.V.O., Santos Filho, A.M.P., Lasmar, M.C., & Sena, M.M. (2016). Detection and characterisation of frauds in bovine meat in natura by non-meat ingredient additions using data fusion of chemical parameters and ATR-FTIR spectroscopy. *Food Chemistry 205,* 14-22.

Oliveri, P., López, M.I., Casolino, M. C., Ruisánchez, I., Callao, M.P., Medini, L., & Lanteri, S. (2014). Partial least squares density modeling (PLS-DM) – A new class-modeling strategy applied to the authentication of olives in brine by near-infrared spectroscopy. *Analytica Chimica Acta, 851,* 30-36.

Obisesan, K.A., Jiménez-Carvelo, A.M., Cuadros-Rodriguez, L., Ruisánchez, I., & Callao, M.P. (2017). HPLC-UV and HPLC-CAD chromatographic data fusion for the authentication of the geographical origin of palm oil. *Talanta, 170, 413-418.*

Ortea, I., & Gallardo, J.M. (2015). Investigation of production method, geographical origin and species authentication in commercially relevant shrimps using stable isotope ratio and/or multi-element analyses combined with chemometrics: An exploratory analysis. *Food Chemistry, 170*, 145-153.

Ottavian, M., Fasolato, L., Serva, L., Facco, P., & Barolo, M. (2014). Data Fusion for Food Authentication: Fresh/Frozen–Thawed Discrimination in West African Goatfish (Pseudupeneus prayensis) Fillets. *Food Bioprocess Technology, 7*, 1025-36.

Perez-Castaño, E., Ruiz-Samblás, C., Medina-Rodríguez, S., Quirós-Rodríguez, V., Jiménez-Carvelo, A. M., Valverde-Som, L., González-Casado, A., & Cuadros-Rodríguez, L. (2015). Comparison of different analytical classification scenarios: application for the geographical origin of edible palm oil by sterolic (NP) HPLC fingerprinting. *Analytical Methods, 7,* 4192-4201.

Pizarro, C., Rodríguez-Tecedor, S., Pérez-del-Notario, N., Esteban-Díez, I., & González-Sáiz, J. M. (2013). Classification of Spanish extra virgin olive oils by data fusion of visible spectroscopic fingerprints and chemical descriptors. *Food Chemistry, 138,* 915-922.

Riedl, J., Esslinger, S., & Fauhl-Hassek, C. (2015). Review of validation and reporting of non-targeted fingerprinting approaches for food authentication. *Analytica Chimica Acta, 885,* 17-32.

Santos, P. M., Pereira-Filho, E.R., & Colnago, L.A. (2016). Detection and quantification of milk adulteration using time domain nuclear magnetic resonance (TD-NMR). *Microchemical Journal, 124,* 15-19.

Sen, I., & Tokatli, F., (2016). Differentiation of wines with the use of combined data of UV–visible spectra and color characteristics. *Journal of Food Composition and Analysis, 45,* 101-107.

Serrano-Lourido, D., Saurina, J., Hernández-Cassou, S., & Checa, A. (2012). Classification and characterisation of Spanish red wines according to their appellation of origin based on chromatographic profiles and chemometric data analysis. *Food Chemistry, 135,* 1425-1431.

Silvestri, M., Elia, A., Bertelli, D., Salvatore, E., Durante, C., Li Vigni, M., Marchetti, A., & Cocchi, M. (2014). A mid level data fusion strategy for the Varietal Classification of Lambrusco PDO wines. *Chemometrics and Intelligent Laboratory Systems, 137*, 181-189.

Spiteri, M., Dubin, E., Cotton, J., Poirel, M., Corman, B., Jamin, E., Lees, M., & Rutledge, D. (2016). Data fusion between high resolution 1H-NMR and mass spectrometry: a synergetic

743 approach to honey botanical origin characterization. *Analytical and Bioanalytical Chemistry,*
744 *408,* 4389-4401.
745
746 Teye, E., Huang, X., Takrama, J., & Haiyang, G. (2015). Integrating NIR and e-tongue
747 together with chemometric analysis for accurate classification of cocoa bean varieties.
748 *Journal of Food Process Engineering, 37*, 560–566.
749
750 Üçüncüoğlu, D., İlaslan, K., Boyacı, I.H., & Özay, D.S. (2013). Rapid detection of fat
751 adulteration in bakery products using Raman and near-infrared spectroscopies. *European*
752 *Food Research and Technology, 237,* 703-710.
753
754 Ulloa, P.A., Guerra, R., Cavaco, A.M., da Costa, A.M.R., Figueira, A.C., & Brigas, A.F.
755 (2013). Determination of the botanical origin of honey by sensor fusion of impedance e-
756 tongue and optical spectroscopy. *Computers and Electronics in Agriculture, 94,* 1-11.
757
758 Wenjuan, S., Xin, Z., Zhuoyong Z., & Ruohua, Z., (2017). Data fusion of near-infrared and
759 mid-infrared spectra for identification of rhubarb. *Spectrochimica Acta Part A: Molecular and*
760 *Biomolecular Spectroscopy, 171,* 72–79.
761
762 Xu, L., Shi, P.T., Ye, Z.H., Yan, S.M., & Yu, X.P. (2013a). Rapid analysis of adulterations in
763 Chinese lotus root powder (LRP) by near-infrared (NIR) spectroscopy coupled with
764 chemometric class modeling techniques. *Food Chemistry, 141*, 2434-9.
765
766 Xu, L., Yan, S.M., Cai, C.B., & Yu, X.P. (2013b). One-class partial least squares (OCPLS)
767 classifier. *Chemometrics and Intelligent Laboratory Systems, 126,* 1-5.
768
769 Xu, L., Shi, W., Cai, C.B., Zhong, W., & Tu, K. (2015). Rapid and nondestructive detection
770 of multiple adulterants in kudzu starch by near infrared (NIR) spectroscopy and
771 chemometrics. *LWT - Food Science and Technology, 61,* 590-595.
772
773 Zhang, X., Liu, Y., Li, Y & Zhao, X. (2017). Identification of the geographical origins of sea
774 cucumber (Apostichopus japonicus) in northern China by using stable isotope ratios and
775 fatty acidprofil·les. *Food Chemistry, 218*, 269-276.
776

Zhang, L., Li, P., Sun, X., Mao, J., Ma, F., Ding, X., & Zhang, Q. (2015). One-class classification based authentication of peanut oils by fatty acid profil·les. *RSC Advances, 103,* 85046-51.

Zhao, M., Downey, G., & O'Donnell, C.P. (2014). Detection of adulteration in fresh and frozen beefburger products by beef offal using mid-infrared ATR spectroscopy and multivariate data analysis. *Meat Science, 96*, 1003-11.

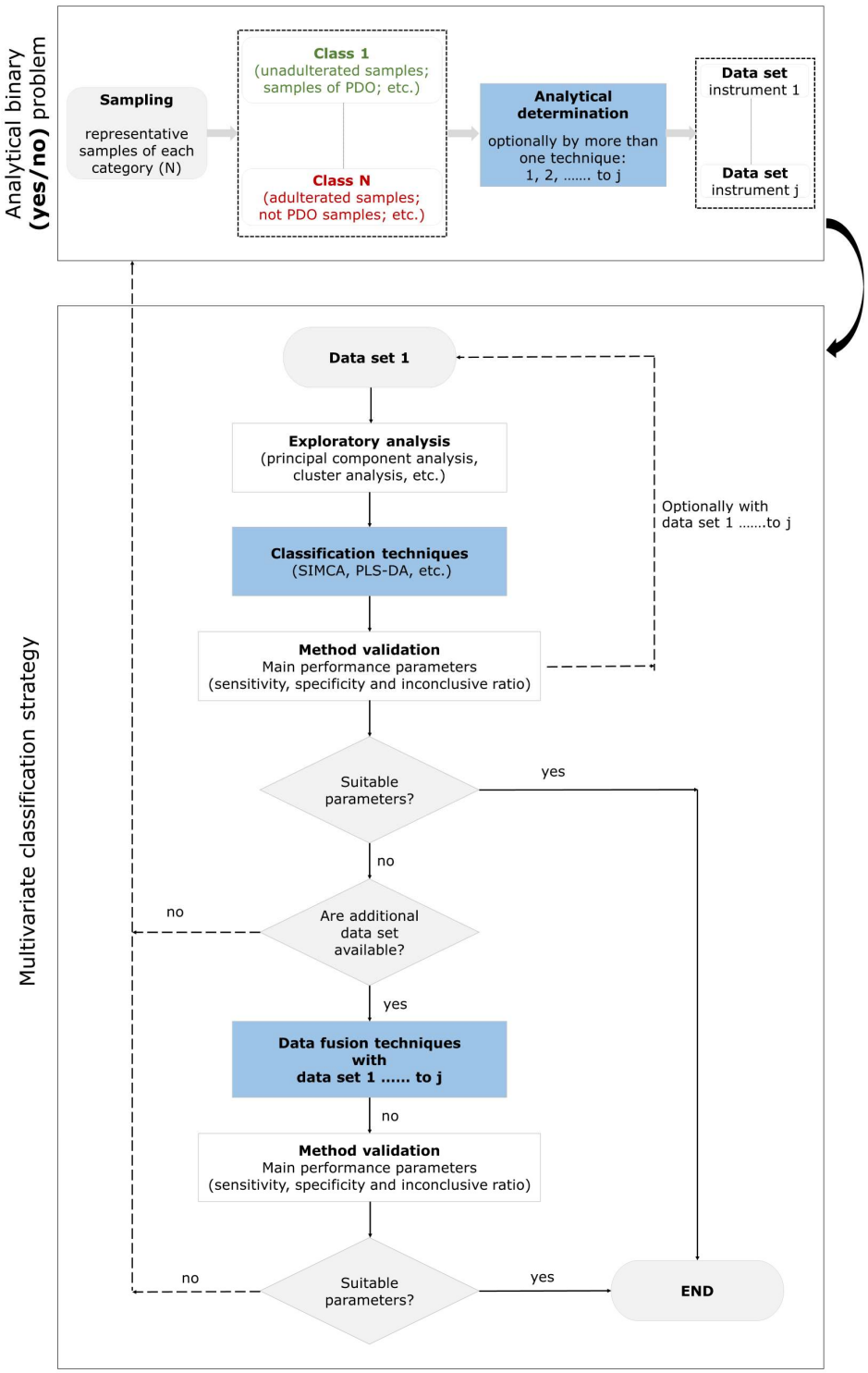
Zhao, M., Downey, G., & O'Donnell, C.P. (2015). Dispersive raman spectroscopy and multivariate data analysis to detect offal adulteration of thawed beefburgers. *Journal of Agricultural and Food Chemistry, 63,* 1433-1441.

Zhu, D., & Nyström, L. (2015). Differentiation of rice varieties using small bioactive lipids as markers. *European Journal of Lipid Science and Technology, 117,* 1578-1588.

791 Figure Captions

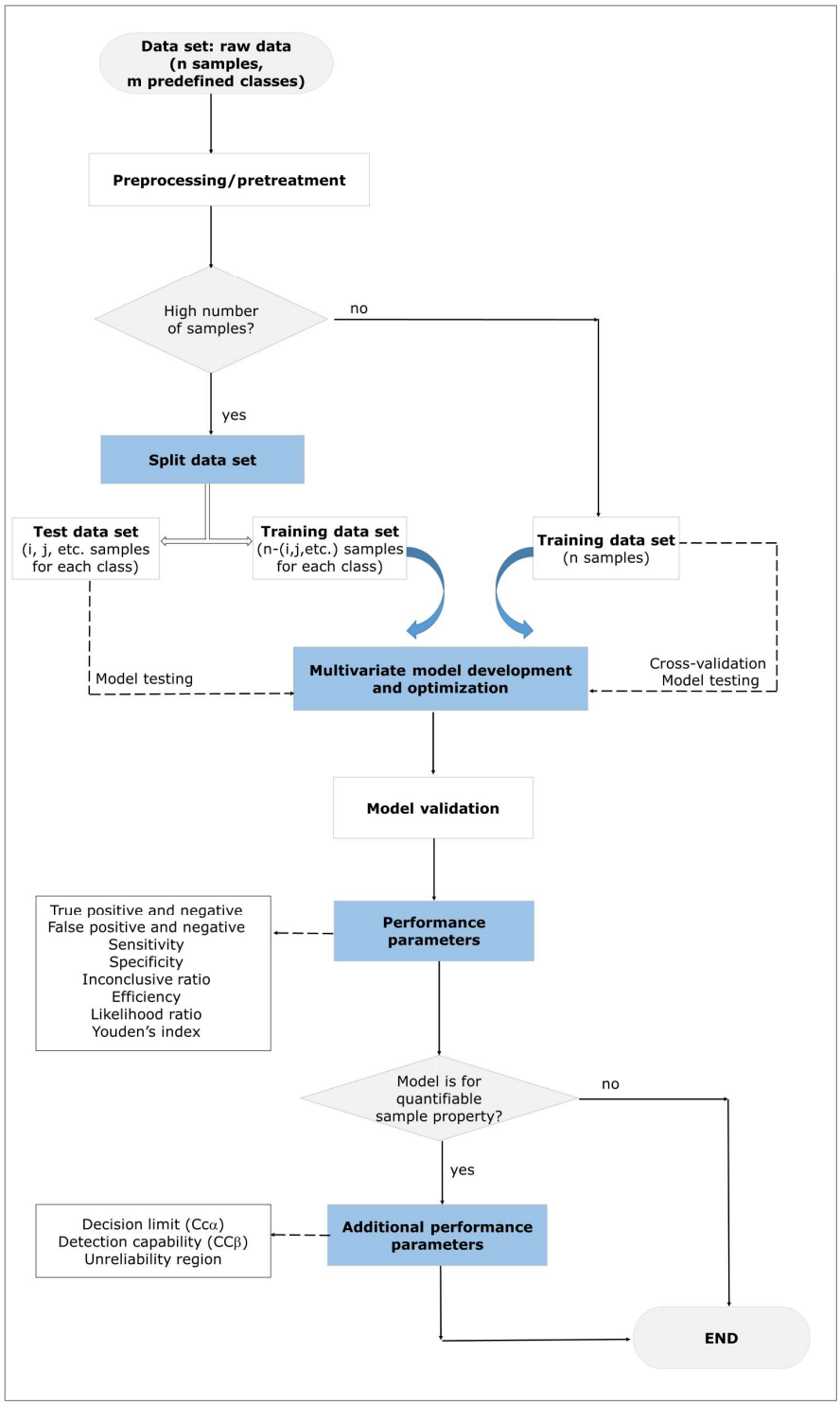
792

793 Fig. 1. Schematic overview of the whole process for multivariate qualitative method
794 development and validation.


Analytical binary (yes/no) problem
Sampling
representative samples of each category (N)
Class 1
(unadulterated samples; samples of PDO; etc.)
Class N
(adulterated samples; not PDO samples; etc.)
Analytical determination
optionally by more than one technique: 1, 2, ... to j
Data set instrument 1
Data set instrument j
Multivariate classification strategy
Data set 1
Exploratory analysis
(principal component analysis, cluster analysis, etc.)
Classification techniques
(SIMCA, PLS-DA, etc.)
Method validation
Main performance parameters
(sensitivity, specificity and inconclusive ratio)
Optionally with data set 1 .....to j
Suitable parameters?
yes
no
Are additional data set available?
no
yes
Data fusion techniques with data set 1 ..... to j
no
Method validation
Main performance parameters
(sensitivity, specificity and inconclusive ratio)
Suitable parameters?
no
yes
END

795

796 Fig. 2. Validation scheme of multivariate qualitative models and the performance
797 parameters.


Data set: raw data
(n samples,
m predefined classes)
Preprocessing/pretreatment
High number
of samples?
no
yes
Split data set
Test data set
(i, j, etc. samples
for each class)
Training data set
(n-(i,j,etc.) samples
for each class)
Training data set
(n samples)
Model testing
Multivariate model development
and optimization
Cross-validation
Model testing
Model validation
True positive and negative
False positive and negative
Sensitivity
Specificity
Inconclusive ratio
Efficiency
Likelihood ratio
Youden's index
Performance
parameters
Model is for
quantifiable
sample property?
no
yes
Decision limit (Ccα)
Detection capability (CCβ)
Unreliability region
Additional performance
parameters
END

798

Table 1. Applications of multivariate exploratory and classification techniques in food fraud

Authentication

author/year	goal	sample	classes	instrumental Technique	chemometric technique
Azevedo, M.S. et al., 2017	geographical origin	honey	…….	GC-MS	PCA, CA
Bevilacqua, M. et al. 2012	PDO	olive oil	2	MIR,NIR	PLS-DA,SIMCA
Binetti, G. et al., 2017	varieties	olive oil	4	NIR, H-NMR, MEO par.	PCA, ANN
Bona, E. et al., 2017	geographical origin	coffee	4	NIR, FTIR	PCA, SVM
Chiesa, L. et al. 2016	PDO	lard	2	NIR/GC	PLS-DA
Dahimini, O. et al.; 2014	pig lard, beef tallow and chicken fat	fats	……	DSC	PCA
Jimenez-Carvelo et al., 2017	type of vegetal oil	oil	1,2	HPLC	PCA, SVM, SIMCA
Kalogiouri, N. P. et al., 2016	sample quality	oil	2	LC-MS	PLS-DA
Majcher, M. et al, 2015	PDO	cheese	4	SPME-MS	PCA,LDA,SIMCA,SVM
Malheiro, R. et al.; 2013	botanical species	mushroom	……	GC-MS	PCA
Mir-Marqués, A. et al. ; 2016	PDO	artichokes	3	ICP-OES, NIR , XRF	PCA,CA,PLS-DA
Oliveri, P. et al., 2014	varieties	olives in brine	**2**	NIR	PLS-DM
Ortea, I. et al, 2015	geographical , production method and biological	shrimps	9, 2 and 7	IR-MS, ICP-MS	PCA,KNN,DA
Sen, I. et al.; 2016	vintage year and variety	wines	4 and 3	UV-VIS, physical parameters	PCA, OPLS-DA, PLS-DA

Serrano-Lourido, J. et al.2012	geographical origin	wines	3	HPLC	PCA, PLS-DA
Zhu, D et al, 2015	varieties	rice	8	UPLC-HR-Q-TOF-MS	PCA,OPLS-DA
Zhang, X. et al., 2017	geographical origin	sea cucumber	7	IRMS, GC	PCA, DA

Adulteration

author/year	**goal (adulterants)**	**sample**	**classes**	Instrumental Technique	**Chemometric thecnique**
Amiry, S. et al., 2017	direct and invert sugar syrup	honey	6	DSC, refactometry, VIS,..	PCA/LDA
Boggia, R. et al. 2013	type of fruit	juices	……..	UV-VIS	PCA
Chen,H. et al., 2017	melamine	milk	1	NIR	PCA, OCPLS
Di Anibal, C. et al.; 2015	Sudan I	spices	2	SF	PLS-DA
De Souza, C. et al; 2017	Formaldehyde, Hydrogen peroxide, Sodium carbonate, Sodium citrate, Starch	milk	6	MIR	PCA, SIMCA
Fadzlillah, N. et al., 2013	mutton fat	Butter	2	FTIR	DA
Georguli, K. et al 2017	hazelnut oil	virgin olive oil	10 and 4	Raman, FTIR	LDA, CLPP
Lopez, M.I et al., 2014	almond paste and chickpea flour	hazelnut paste	2 and 2	NIR	PCA,SIMCA
Maia,M. et al, 2013	unspecific	beewax	2	GC-MS	CA, PCA, LDA
Mu, T. et al.; 2016	worst vegetable oils	extra virgin olive oil	3	LIF	PCA, SVM, ANN
Santos, P. et al. 2016	water, whey, urea, hydrogen peroxide, synthetic urine and synthetic milk	milk	2	H-NMR	PCA,SIMCA,KNN
Üçüncüoğlu, D. et al; 2013	margarine	bakery products	……	NIR, Raman	PCA
Xu L. et al.; 2013	cassava, sweet potato, potato and maize starches	lotus root powder	2	NIR	SIMCA,PLSCM
Xu, L. et al. 2013	edible and industrial gelatine powder and soy protein powder	whole milk powder	1	NIR	OCPLS

Xu, L. et al. 2016	unspecific adulterants	kudzu starch	1	NIR	OCPLS
Zhao M, et al., 2014	Beef Offal	beef burgers	2	MIR	SIMCA,PLS-DA
Zhao, M. et al; 2015	Beef Offal	beef burgers	2	DRS	PCA,PLS-DA,SIMCA

Abbreviations:

Chemometric techniques: ANN, Artificial neural networks; CA, Cluster analysis; CBT, Classification binary trees; CDA, Canonical discriminant analysis; CLPP: Continuous locality preserving projections; CT, Classification
Tree; DA, Discriminant analysis; KNN, K-nearest neighbour; LDA, Linear discriminant analysis, OPLS, Orthogonal partial least squares; OCPLS, One-class partial least squares; PLSCM, Partial least squares class model;
PLS-DM, Partial least squares density modelling; PLS-DA, Partial least square discriminant analysis; PCA, Principal component analysis; SIMCA, Soft independent modelling of class analogy; SVM, Support vector machines.

Instrumental techniques: DSC, Differential scanning calorimetry; DRS, Dispersive Raman spectroscopy; FTIR, Fourier transform Infrared; GC–MS, Gas chromatography mass spectrometry; H -NMR, Hydrogen magnetic nuclear resonance; HPLC, High performance liquid chromatography; ICP-MS Inductively coupled plasma mass spectrometry; ICP-OES, Inductively coupled plasma optical emission spectrometry.
IRMS, Isotope-ratio mass spectrometry; LIF, Laser induced fluorescence; MEO_par, merceological parameters; MIR, Mid-infrared spectroscopy; NIR, Near-infrared spectroscopy; P-NMR, Phosphor magnetic nuclear resonance; SF, Synchronous fluorescence;
SPM-MS, Solid phase micro extraction-mass spectrometry; UV-VIS, Ultraviolet and visible spectroscopy; UPLC-MS, Ultra performance liquid chromatography mass spectrometry; XRF, X-ray fluorescence.

Others: PDO, Protected Designation of Origin.

Table 2. Examples of data fusion strategies in food authentication/adulteration problems.

author/year	sample	category/analyte	instrumental technique	chemometric technique	fusion level (variable selection)	raw variables
Alamprese C. et al., 2013	minced beef	unadulterated, 4 adulteration levels	UV-Vis, NIR, MIR	PCA, LDA, PLS	mid (FC)	290, 1090, 825
Bajoub A. et al., 2017	olive oil	5 geographical origin	HPLC-DAD, HPLC-FLD	PCA, PLS-DA, SIMCA, KNN	low, mid (PLS-DA scores)	not specified
Biancolillo A. et al., 2014	beer	2 quality (high, low)	TG, MIR, NIR, UV, Vis	SIMCA, PLS-DA	low, mid (PLS-DA scores)	817, 1650, 3112, 165, 441
Borràs E. et al., 2016	olive oil	6 classes	HS-MS, MIR, UV-Vis	PLS-DA	Low, mid (PLS-DA scores)	301, 594, 701
Chen Q. et al., 2014	vinegar	4 ages (year)	ISEs, RGB	PCA, LDA	mid (C-index)	20, 3
Drivelos S.A. et al., 2014	yellow split pea	2 classes	ICP-MS (rare earth, trace elements)	OPA, MD, PLS-DA, KNN	low	12, 15
Erich S. et al., 2015	milk	2 classes	H-NMR, C-NMR, GC-FID, IRMS	PCA, LDA, FDA, PLS-DA	mid (M-ANOVA, CLV)	not specified, 3
Forina M. et al., 2015	olive oil	5 geographical origin	HS-MS, NIR, UV-visible	PCA, LDA, QDA-UNEQ	mid (STEP-DA)	20, 1500, 810
Haddi Z. et al., 2014	fruit juice	11 flavours	TGS, ISEs	PCA, CA , ARTMAP-NN	low	5, 6
Márquez C. et al., 2016	hazelnut	unadulterated, 2 adulterants	FT-Raman, NIR	SIMCA	high, mid (xdiff)	1510, 2166
Monakhova Y.B. et al., 2014	wine	grape variety, geographical origin, vintage year	H-NMR, IRMS	PCA, LDA, PLS-DA, FDA, ICA, MBPLS-DA	low, mid (M-ANOVA, CLV)	869, 5
Ni Y. et al., 2012	rhizome curcuma	3 types	GC-MS , HPLC-DAD	PCA, LDA, BP-ANN, LS-SVM	mid (FC)	27, 16
Nunes K.M. et al., 2016	bovine meat	unadulterated, 4 adulteration levels	ATR-FTIR, Phy-Chem	PCA, PLS-DA	low, mid (VIPscores)	1803, 5
Obisesan K.A. et al., 2017	palm oil	3 origin	HPLC-DAD, HPLC-CAD	PCA, PLS-DA	high, mid (PCA, iPLS)	3436, 1609

Ottavian M. et al., 2014	Goatfish	2 classes (fresh, frozen)	NIR, RGB	PCA, PLS-DA	mid (C-index)	401, 3
Pizarro C. et al., 2013	olive oil	3 geographical origin	UV, Phy-Chem	PCA, LDA, PLS-DA	low, mid (PCA)	206, 5
Silvestri M. et al., 2014	wine	3 varieties	H-NMR, EEM, HPLC-DAD	PCA, PLS-DA, NPLS-DA	mid (PCA, one-PARAFAC, MCR-area)	not specified
Spiteri M. et al., 2016	honey	5 monofloral origins	H-NMR, LC-HRMS-O MS, LC-HRMS-TOF MS	PCA, PLS-DA	low, mid (PCA, VIPscores)	29380, 58843, 1729
Teye E. et al., 2015	cocoa bean	5 varieties	NIR, ISEs	PCA, SVM	mid (PCA)	1557, 7
Ulloa P.A. et al., 2013	honey	4 commercial brands (botanical origin)	UV–Vis, NIR, e-tongue	PCA, CA (KNN), MPCA	low, mid (PCA, RII-Index)	201, 3348, 252
Wenjuan, S. et al., 2017	rhubarb	2 classes (official/unofficial)	NIR, MIR	PCA, PLS-DA, SIMCA, SVM, ANN	low, mid (WT, iPLS)	700, 700

Abbreviations:

Instrumental techniques: ATR-FTIR, attenuated total reflectance Fourier transform infrared spectroscopy; CSA, colorimetric sensor arrays; C-NMR, carbon nuclear magnetic resonance spectroscopy; EEM, emission-excitation fluorescence spectroscopy; e-nose, non-selective chemical sensors; e-tongue, impedance electronic tongue; FT-Raman, fourier transform raman spectroscopy; GC, gas chromatography; GC-FID gas chromatography with FID detector; HPLC, high-performance liquid chromatography; H-NMR, proton nuclear magnetic resonance spectroscopy; HPLC–DAD, HPLC–diode array detector; HPLC-FLD, HPLC fluorescence detector; HS-MS, head-space mass spectrometry; ICP-MS, inductively coupled plasma-mass spectrometry; IRMS, isotope ratio MS; ISEs, potentiometric chemical sensors or electronic tongue; LC-HRMS, liquid chromatography high resolution mass spectrometry; MIR, mid infrared spectroscopy; MS, mas spectroscopy or e-nose; NIR, near infrared spectroscopy; O MS, orbitrap mas spectroscopy detector; Phy-Chem, physico-chemical parameters; RGB, digital RGB image; SNIF-NMR, site-specific natural isotope fractionation – nuclear magnetic resonance; TG, thermogravimetry; TGS, gas sensor; TOF MS, time of flight MS; UV, ultraviolet spectroscopy; Vis, visible spectroscopy.

Chemometrics techniques: ANN, artificial neural network; ARTAMAP NN, fuzzy ARTMAP neural network; BP-ANN, back propagation-artificial neural networks; CA, cluster analysis; CDA, canonical discriminant analysis; FDA, factorial discriminant analysis; ICA, independent components analysis; KNN, K-Nearest Neighbours; LDA, linear discriminant analysis; LS-SVM, least squares-support vector machine; MD, Mahalanobis distance; MBPLS-DA, multi-block extension of PLS-DA; MPCA, multi-way PCA; NPLS-DA, multilinear PLS-DA; PARAFAC, OPA, orthogonal projection analysis; parallel factor analysis; PCA, principal component analysis; PLS-DA, partial least squares discriminant analysis; SIMCA, soft Independent modelling of class analogy; SVM, support vector machine; UNEQ-QDA, unequal-quadratic discriminant analysis.

Variable selection: ANOVA, one-way variance analysis; CLV, clustering of latent variables; C-index, color-index; D-index, distance-index; FC, fisher weight criterion; iPLS, interval partial least squares; M-ANOVA, multiway analysis of variance; MCR-area, peaks areas of multivariate curve resolution; one-PARAFAC, mode one PARAFAC loadings; PCA, principal components scores; RII-index, ratio of inter-distance to intra-distance in the score space, SWD, stepwise decorrelation; STEP-LDA Stepwise-Linear Discriminant Analysis; SWS, stepwise selection; VIPscores, weighted sums of squares of the PLS weights; WT, wavelet transform.

839

840

841