# Alternative data mining/machine learning methods for the analytical evaluation of food quality and authenticity – A review

Ana M. JIMÉNEZ-CARVELO✉, Antonio GONZÁLEZ-CASADO, M. Gracia BAGUR-GONZÁLEZ, Luis CUADROS-RODRÍGUEZ

Department of Analytical Chemistry, Faculty of Science, University of Granada, C/ Fuentenueva s/n, E-18071, Granada, Spain

## Abstract

In recent years, the variety and volume of data acquired by modern analytical instruments in order to conduct a better authentication of food has dramatically increased. Several pattern recognition tools have been developed to deal with the large volume and complexity of available trial data. The most widely used methods are principal component analysis (PCA), partial least squares-discriminant analysis (PLS-DA), soft independent modelling by class analogy (SIMCA), k-nearest neighbours (kNN), parallel factor analysis (PARAFAC), and multivariate curve resolution-alternating least squares (MCR-ALS). Nevertheless, there are alternative data treatment methods, such as support vector machine (SVM), classification and regression tree (CART) and random forest (RF), that show a great potential and more advantages compared to conventional ones. In this paper, we explain the background of these methods and review and discuss the reported studies in which these three methods have been applied in the area of food quality and authenticity. In addition, we clarify the technical terminology used in this particular area of research.

## Keywords

Data mining; random forest; CART; decision tree; food analysis

✉ Corresponding author: telephone: +34958240797; fax: +34958243328; e-mail: amariajc@ugr.es

## 1. Introduction

The assurance of food authenticity is the main concern of many consumers and manufacturers of high-quality products as well as official bodies and authorities in response to the need to protect consumers by detecting potential food fraud. Food authenticity is necessarily linked to the compliance; hence an authentic foodstuff is a product which strictly complies with genetic identity, natural composition, geographical and typological origin, ingredients, production technology, implicit quality features and explicit claims stated in the label. Overall, the food fraud entails a deception about the origin, quality or quantity of a foodstuff aimed of making an illicit profit. Globalization and free trade agreements have fostered an increased exchange of and access to food around the world. However, this has also led to an increase in problems associated with food fraud. There are three kind of main food frauds: non-conformity, adulteration and contamination. Non-conformity occurs when a food product does not fulfil the features which are stated in the label; it is identified by counterfeiting or imitation. Adulteration involves a deliberate and non-stated alteration of the intrinsic composition of the original food product. At least, contamination involves an unintended or accidental presence of extrinsic substances.

The serious nature of food fraud depends on the kind of fraud carried out. For instance, a food adulteration might be the substitution of the original ingredients by ingredients cheaper [Spink, Hegarty, Fortin, Elliot, & Moyer, 2019], as in the case of the olive oil that could be adulterated with cheaper vegetable oils. In this case, the consumer is paying more for a food product of inferior quality, but it does not involve any health risk. However, there are other types of food frauds which might affect to the human health. For example, the use of contaminated commodities, ingredients or allergens. In this sense, it is important that the food chain and the possible food fraud are extremely controlled by official bodies [Manning, 2016]. On the other hand, it is also important to ensure the authenticity of the product in term of geographical origin in order to control the replacements of genuine food products [Huch, Pezzei, & Huck-Pezzei, 2016; Medina, Perestrelo, Silva, Pereira, & Câmara, 2019]. Analytics involves several activities such as the analytical determination of specific physico-chemical characteristics, the qualification/quantitation of adulterants and/or contaminants and residues, and the verification of quality-differentiated technical requirements.

In this context, multivariate data analysis and pattern recognition techniques are powerful tools to conduct quality control and food authentication [Zielinski et al., 2014; Bevilacqua at al., 2017; Brereton et al., 2017; Callao & Ruisánchez, 2018; Efenberger-Szmechk, Nowak, & Kregiel, 2018; Granato et al., 2018].

The main purpose of multivariate pattern recognition methods is to perform the most appropriate data treatment in order to model and characterize a set of objects or samples that exhibit a particular feature or behaviour. To this end, significant and non-evident information is extracted to establish relationships between the objects/samples of the set, or between the set of objects/samples and one or several characteristics, according to the similarity of their spectra, chromatograms, elementary analysis, images, and so on. These tools must also be able to classify new samples into a certain group and reliably predict the value of a specific property in a fast and objective way [Brereton, 2015].

Pattern recognition methods are divided into two main groups: unsupervised methods, whose main tools are principal component analysis (PCA) and hierarchical cluster analysis (HCA); and supervised methods of analysis, such as k-nearest neighbours (kNN) [Steinbach & Tan, 2009], partial least squares-discriminant analysis (PLS-DA) [Ballabio & Consonni, 2013], and soft independent modelling by class analogy (SIMCA) [Oliveri & Smilde, 2012], among others. Likewise, in machine learning field are known as unsupervised and supervised learning techniques [Kavakiotis et al., 2017]. Figure 1 shows a straightforward flowchart of conventional pattern recognition methods.

Figure 1

An exploratory analysis is generally used to scout the data structure and determine whether there are trends in the data set. Principal component analysis (PCA) is a valuable statistical tool whose goal is to maximize the information of the variance in the data and show it visually in as few components as possible. It is mainly used to provide information on natural groupings of objects/samples and to reduce the number of variables necessary to represent the system, providing a new set of latent variables known as 'principal components' [Bro & Smilde, 2014]. Nevertheless, sometimes PCA has been erroneously applied and is used in some studies as a classification method to develop and validate classification models [Hakki, 2014; Chung, Kim, Lee, & Kim, 2015; Sun, Lin, Li, Shen, & Luo, 2015]; this is a serious mistake that unfortunately still happens. Cluster analysis is based on the intrinsic similarity between groups of objects/samples. The results of the hierarchical clusters analysis are presented as a dendrogram where the objects/samples are distributed in a ramified tree where the data are organised in categories and subcategories (branches) and the nodes represents the clusters according to their similarity [Drab & Daszykowski, 2014].

Supervised methods of analysis are divided into two groups: (i) classification or qualification methods and (ii) calibration or quantitation methods. Multivariate classification/qualification methods have been defined as chemometric techniques designed to find mathematical models that can recognize which class each object/sample belongs to base on a particular data set; they involve the use of various chemometric algorithms with two main statistical backgrounds related to discrimination and class-modelling approaches [Marini, 2010; Belvilacqua et al., 2014].

There are many classification methods but the most common ones are kNN, PLS-DA and SIMCA. Multivariate calibration/quantitation methods are in fact multivariate regression methods aimed at determining the functional relationships between the analytical signal acquired from a set of samples and a characteristic feature of such samples such as their composition. The most widely used algorithm is partial least squares (PLS) regression [Mehmood & Ahmed, 2016]. It should be noted that, although the classification is intrinsically a qualitative process, the assignment of the objects or samples to a specific class can have a qualitative basis (as by kNN or SIMCA) or a quantitative basis (as by PLS-DA). Indeed, the PLS-DA method involves performing a multivariate regression and placing a numeric value to each object/sample first, and then classifying them into a specific class [Brereton & Lloyd, 2014]. In addition, there are other kinds the multivariate methods that are applied when working with second order data. That means that a matrix of data is obtained for each sample rather than a vector of data (first order data). In this case, the most common methods are parallel factor analysis (PARAFAC) and multivariate curve resolution – alternating least squares (MCR-ALS).

The development of a pattern recognition supervised model involves two stages. The first stage is to build the model using a set of objects or samples whose class or particular features are known (i.e., training set or calibration set). In this stage, an internal validation or cross-validation could be applied in order to assess the goodness of fit of the model from the samples/objects of the training set. However, cross-validation by its own designs purpose, never able to achieve all the necessary objectives of a right validation [Esbesden & Geladi, 2010]. The second stage is to evaluate and externally validate the performance of the model built in the previous stage; this is done using additional objects or samples (i.e., test set or validation set) that fulfil the same requirements but were not part of the original training set [Szymanska et al., 2015; Westad & Marini, 2015]. In these methods, it is assumed that there are enough reference objects/samples that act as analytical standards because the outcomes of interest (i.e., the qualitative class or the value of one or more quantitative features) are formerly known or have been accurately measured. There are not definitive rules on the minimum number of samples/objects which are necessary for model development as this depends on the particular problem; it would however be desirable to devote the 40-50% of the reference samples/objects for the validation set.

The assessment of the quality of the classification models is evaluated through several performance features. These are estimated using the contingency table which records the number of both correct and incorrect assignations for each class in which samples of the validation set are arranged [Cuadros Rodríguez, Pérez Castaño, & Ruiz Samblas, 2016; Ballabio, Grisoni, & Todeschini, 2018]. In the same way, specific figures of merit have been proposed to assess the multivariate calibration models [Olivieri et al., 2006; Oliieri, 2014]. However, the assessment of the multivariate models is not enough and the whole analytical method should also be properly validated [Van der Veer, Van Ruth, & Akkermans, 2011; Alewijn, Van der Voet, & Van Ruth, 2016].

As regards the effective use of classification methods, some authors argue that it is better to use class-modelling methods such as SIMCA to perform an adequate food authentication [Rodionova & Titova, 2016]. This is because class-modelling methods operate, in the training stage, by defining a well-delimited acceptance region that contains all the objects/samples of the target class; consequently, only new objects/samples located in the acceptance region are assigned as belonging to the target class.

In recent years, the applications of new pattern recognition algorithms are growing in the area of food, due to their advantages and potential to solve complex problems related to food authenticity. The most widely used ones are support vector machine (SVM), classification and regression tree (CART), and random forest (RF), which can be used in both classification and calibration models. Surprisingly, their application is still scarce in the area of food quality and authenticity, although they are widely used in other areas such as metabolomics. Some authors have even reported their advantages compared to conventional techniques. For example, it has been stated that [Gromski et al., 2015] "… compared to PLS-DA, SVM is not influenced by the distribution of the different sample classes but rather focuses on which side of the support vectors particular test samples fall on". Similarly, the advantages of the RF algorithm have been reported in the area of ecology [Cutler et al., 2007].

As stated above the supervised multivariate methods are split in two groups (i) qualification or classification methods and (ii) quantification methods. In turn, classification methods are conventionally divided in discriminant analysis methods and class modelling methods depending on how the model is built. Discriminant analysis, as PLS-DA, works by establishing the boundaries between the different classes defined by the training objects while the class modelling methods, as SIMCA, define successive enclosed space domain which contain the objects of each class. Nevertheless, the models generated by decision trees methods (DT), as CART or RF, do not establish the separation of data in different classes as way above but the samples are divided into subsets (or classes) based on the value of certain variables, and this process is repeated on each derived subset of samples [Ai et al., 2014]. Consequently, the classifications are based on a set of concatenated decisions, similar to artificial neural networks (ANN). Figure 2 (a) shows a straightforward flowchart of the most common data mining/ chemometrics methods used for the analytical evaluation of food quality and authenticity, and figure 2(b) shows schematically how these methods operate. Table 1 assembles some of the advantages and disadvantages of them.

Figure 2

Table 1

All the methods cited above perform the classification using a threshold that is automatically established by the typical software of treatment multivariate data as PLS_Toolbox (under Matlab) (Eigenvector Research, WA, USA), SOLO (Eigenvector Research, WA, USA), SIMCA (Umetrics, Sweden) Unscrambler, (CAMO, Norway), Pirouette (Infometrix, WA, USA) or perClass Toolbox (under Matlab) (perClass BV, The Netherlands), to list only the most known. However, practitioners can decide on the classification threshold to conduct a more reliable classification [Vitale, Marini, & Ruckebush, 2018]. Table 2 collects a summary of the most common data mining methods which can be applied with the different software of multivariate data analysis.

Table 2

This paper reviews and describes the use of these alternative data mining/machine learning methods (i.e., SVM, CART, and RF) in the area of food analysis. Examples are provided to demonstrate the potential of these techniques in this area of study.

## 2. Background

### 2.1 Some basic terms

The automation and computerization of analytical laboratories have resulted in numerous changes; one of them is the acquisition of a high volume of data, giving rise to a new scientific discipline known as 'big data science', which has had a strong impact on many scientific disciplines. In chemistry, the term 'big data' refers to large and complex data sets that contain useful and non-evident chemistry-related information that must be extracted using complex data analysis tools [Parastar & Tauler, 2018]. Nevertheless, having a large amount of data does not mean that adequate answers can be provided unless the right data processing tools are applied. Collecting data is not synonymous with possessing information; data must be treated and interpreted to convert them into useful information for the user or the analyst. This subject, the right use of big data, and how it could satisfy the ISO/IEC 17025 requirements in the accreditation of laboratories has been already described [Ghernaout, Aichouni, & Alghamin, 2018].

The nomenclature used to refer to this kind of tools depends on the area of study. Analytical chemistry is the area with the greatest variability of terms. Some authors use the terms 'pattern recognition methods' or 'multivariable analysis methods', but the most commonly-used term is 'chemometric tools' to refer to the methods applied to the treatment of chemistry-related data. At its inception, chemometrics were defined as *an approach to analytical and measurement science that uses mathematical, statistical and other methods of formal logic to determine (often by indirect means) the properties of substances that otherwise would be very difficult to measure directly* [Lavine, 2000]. Currently, the International Union of Pure and Applied Chemistry (IUPAC) considers chemometrics as *the science of relating measurements made on a chemical system or process to the state of the system via application of mathematical or statistical methods* [Hibbert, 2016].In the field of engineering, these types of techniques for the processing of signals or images are often referred to as 'computational intelligence' or 'artificial intelligence' tools. The IUPAC has defined artificial intelligence as *the capability of a machine to perform human-like intelligence functions such as learning, adapting, reasoning and self-correction. The main areas of application are currently in expert systems, computer vision, natural language processing, robotics, and speech synthesis and recognition* [Kingston & Kingston, 1994]. Other authors define this term as *the interaction of several kinds of disciplines, such as computer science, cybernetics, information theory, psychology, linguistics, and neurophysiology. Artificial intelligence is a branch of computer science involved in the research, design and application of intelligent computers* [Lu, Chen, & Zheng, 2012]. Artificial neural networks (ANN) are the most widely used algorithm in this area. They are based on a series of 'nodes' or 'artificial neurons' that are interconnected with each other in a network that attempts to simulate the network of neurons in the human brain [Hibbert, 2016]. ANN are not explained in this study due to their different applications, although it is also usually classified as an alternative data treatment method [Yu, Low, & Zhou, 2018; Ropoli, Panagou, & Nychas, 2016; Marini, 2009].

In the areas of health care and biology (e.g., medicine, pharmacy, biology and biotechnology) the term 'bioinformatics' is routinely used and defined as the *discipline encompassing the development and utilization of computational facilities to store, analyse, and interpret biological data* [Duffus, Nordberg, & Templeton, 2007].

## 2.2 Data mining *vs.* machine learning

'Data mining' is a general term that encompasses all these tools regardless of the area of study in which they are used. This term appeared in the 1960s, but only became consolidated in the 1980s with the concept of 'knowledge discovery in databases' (KDD) [Mikut & Resichl, 2011; Han, Kamber, & Pei, 2012]. The term 'machine learning' is also commonly used for the same purpose [Zheng, Fue, & Ying, 2014]. Both terms are often used interchangeably to refer to all these processing data techniques although, strictly speaking, some differences can be observed between them.

Data mining can be used for descriptive purposes (i.e., showing similarities between the elements of a data set), or predictive purposes (i.e., predicting specific features of new data based on models that have previously been built and validated). It is based on the collection, storage, and treatment of a large amount of data in order to make the best decisions about a particular problem. It is an interdisciplinary field with the overall objective of revealing relationships in data from whatever source or origin. To this end, complex data treatment tools are used to detect and identify hidden patterns, associations, and structures that are proper of the raw big data set, or to select and filter useful information from big databases [Mitra & Acharya, 2003]. The concept of machine learning is also known as *the techniques involved in dealing with vast data in the most intelligent fashion (by developing algorithms) to derive actionable insights. In these techniques, we expect the algorithms to learn by them without being explicitly programmed* [https://www.analyticsvidhya.com]. Consequently, data mining refers to the area in general and machine learning refers exclusively to the algorithms used and it is linked to pattern recognition.

The IEEE International Conference on Data Mining, held in Hong Kong 2006, identified the top 10 data mining algorithms which were among the most influential data mining algorithms in the research community: C4.5, k-Means, SVM, Apriori, EM, PageRank, AdaBoost, kNN, Naïve Bayes, and CART. A survey paper was published describing the basis of each one [Wu et al., 2006]. Two of these, SVM and CART, are considered in this review.

In addition, data mining methods have been classified in four machine learning categories: (i) information-based learning, (ii) similarity-based learning, (iii) probability-based learning, and (iv) error-based learning. In general, DT methods (e.g., CART and RF) fall into the category of information-based learning and SVM belongs to the category of error-based learning [Keller, Name, & D'arcy, 2015]. We consider this classification to be very appropriate, since the methods are sorted according to how they build the different regions for each object/sample class of the classification model.

## 2.3 Data mining in food analysis

In recent years, data mining has been used more frequently in the area of food analysis, leaving the concepts of pattern recognition techniques or methods and chemometric tools to refer to the algorithms used to process the data. Both data mining and chemometrics represent very similar concepts. In fact, the only difference is that chemometrics has been used in reference to the application of machine learning techniques in order to obtain information of material system from data of mainly chemical or physical-chemical nature, while data mining is extensively used in many other areas such as security, facial recognition, customised marketing, medical diagnosis, air navigation, etc.

Researchers have reviewed some of the data mining methods that are increasingly used in chemometrics, that is, exploratory data analysis, artificial neural networks, pattern recognition, and digital image processing [Mutihac & Mutihac, 2008; Kumar, Bansal, Sarma, & Rawal, 2014; Messai, Farman, Sarraj-Laabidi, Hammami-Semmar, & Semmar, 2016]. Data mining methods are widely used in the area of food quality to verify compliance with regulations and quality-differenced requirements in order to ensure the authenticity of food. Besides this, consumers increasingly demand more information and knowledge about foodstuffs from producers.

The food sector is highly competitive and global, so food producers seek to become consolidated in emerging domestic and international markets and to make a difference with their products. Product differentiation is key to take a leading position in the global market of the sector. For example, a good strategy to outcompete competitors is to take advantage of the difference in the chemical composition or organoleptic characteristics of food. As a result, a current trend in analytical chemistry is to develop quick and reliable analytical methods to authenticate food products. This has led to the development of more powerful analytical instruments and the use of new methodologies to obtain more and better information about the objects/samples of study. An example of this is the development of more advanced sensors that can monitor food with a high level of detail, collecting a large volume of data. Thus, alternative methods to conventional data processing techniques are required.

Traditionally, chemometrics has been used in the area of food analytical chemistry to refer to the use of well-known conventional methods such as PCA, kNN, SIMCA, PLS-DA, and algorithms applied to second-order data such as parallel factor analysis (PARAFAC) and multivariate curve resolution-alternating least squares (MCR-ALS) [Zielinski et al., 2014; Callao & Ruisánchez, 2018; Rodopi, Panagou, & Nychas, 2016; Dai, Sun, Xiong, Cheng, & Zeng, 2014]. Nevertheless, as explained in the Introduction section, it is becoming more frequent to use the most up-to-date data processing methods, in food analysis, since they exhibit advantages and greater power than the conventional methods previously cited.

Figure 3 shows a plot of the trend in publications on food chemistry that have applied data mining methods in recent years. As can be seen, the SVM method is the most widely used method; however, in the last years the RF algorithm, which was used scarcely in food chemistry, has become more widespread and in 2018 the number of papers applying RF has tripled to the ones using SVM. In addition, the figure 4 shows the increase in the use of the 'data mining' term in the papers published in the area of food analytical chemistry in recent years.

Figure 3

Figure 4

## 3. Alternative methods

In most cases, the reported studies apply conventional multivariate pattern recognition methods with the main purpose of analysing the similarity between signals for food identification, classifying food according to various criteria (e.g., botanical or animal species, geographical origin), detecting adulterations and other non-conformities, and predicting properties related to food quality, such as antioxidant capacity [Jiang, Zheng, & Lu, 2015; Cuberon-Leon, Peñalver, & Maquet, 2016; Popescu et al., 2015; Pisano, Silva, & Olivieri, 2015] and stability. Several authors have reviewed the published studies about the use of these methods [Bosque-Sendra, Cuadros-Rodríguez, Ruiz-Samblás, & de la Mata, 2012; Berrueta, Alonso-Salces, & Héberger, 2007; Olivieri & Downey, 2012; Khakimov, Gürderniz, & Engelsen, 2015; Olivieri, 2012; Ortiz & Sarabia, 2007].

Recently, a comprehensive and valuable review has provided an overview of all the stages of the analysis of large analytical chemical datasets [Szymanska, 2018]. Nevertheless, it only considered conventional data processing methods and SVM, leaving out new data mining methods such as CART, RF, and others. Similarly, other recent reviews have focused on traditional chemometric methods and do not include any reference to these alternative methods [Callao & Ruisánchez, 2018, Cocchi, 2017]. This demonstrates that the inclusion of data mining methods in the area of analytical chemistry and specifically in food chemistry is relatively recent.

Considering only SVM, CART, and RF methods, the first method is the most widely used in food analytical chemistry and has been explored in a greater number of studies than the others (see Figure 2) [Mutihac & Mutihac, 2008; Brereton & Lloyd, 2010; Luts et al., 2010]. The goal of SVM is to find the best hyperplane in space that differentiates between the classes of the objects/samples by applying a maximization method. The aim is to maximize the 'margin', which is based on the sum of the distances from the hyperplane to the closest samples, that is, those correctly classified into their corresponding classes; SVM penalizes the number of misclassified samples. SVM algorithm uses a set of mathematical functions that are defined as the kernel. The kernel functions transform the original data into the required format. If the hyperplane is built in the original space the SVM model applies a linear kernel (it works similarly to the PLS-DA algorithm); if it is built in a different space (e.g., a higher dimension space), the SVM model is non-linear and alternative kernel functions must be used as the radial basis function (RBF) [Xu, Zomer & Brereton, 2006].

The main advantage of SVM over PLS-DA is that it creates a separation between the regions of the different classes when these are not sufficiently evident. Nevertheless, it is easier and faster to conduct a classification using PLS-DA than using SVM, since PLS-DA only performs a regression by partial least squares on the original data whereas SVM takes into account the transformation of the data in a higher dimension space. SVM is used for different purposes in the area of food analytical chemistry: (i) to classify food according to its geographic origin, (ii) to conduct a sensory evaluation, (iii) to detect adulterations, (iv) to quantify compounds, and (v) to conduct quality control.

DT is one of the most popular classification machine learning methods and is also widely used in the selection of features to determine food quality [Debska & Guzowska-Swider, 2011]. DT are sequential models which logically combine a sequence of simple comparisons between a numeric value of an input variable against a threshold value or a nominal attribute against a set of possible values [Kotsiantis, 2013]. DT divides the variable space into rectangular regions and predict the label associated with an particular instance by traveling from a root node of a tree to a leaf, where each label corresponds to one class. Each of these results creates additional nodes that branch out into other possibilities. This creates a structure that resembles a tree. There are three types of nodes: probability nodes, decision nodes, and terminal nodes [Witten & Frank, 2005]. DT can be translated into a set of rules by creating a separate rule for each path from the root to a leaf in the tree. Thus, the classification of a new sample begins in the root node of the tree and follows the branch that is appropriate to its outcome. The most known DT method is CART which is a single tree that shows many branches where the data set is split according to the selected decision, and the procedure is repeated as often as necessary. Figuratively, CART implies building a tree by growing and pruning it [Kucheryavskly, 2018].

DT models can be combined into ensembles by using boosting or bagging for yielding better predictive results than any of their constituent models when used separately [Kotsiantis, 2013; Kucheryavskly, 2018]. Boosting and bagging imply a sequential improving of a single tree by using random subsets from the whole dataset to build a set of small trees. These merged DT are called 'ensemble methods' and sometimes 'decision forest'. The main idea behind this is to combine several individual classifiers to obtain a classifier that outperforms every one of them [Rokach, 2010; Ruiz-Samblás, Cadenas, Pelta, & Cuadros-Rodríguez, 2014]. The main difference between both ensemble methods is the iteration. Boosting works iteratively weighting the individual instances in each run and learning successive models from the miss-classified examples while bagging generates independent models, each one from a different data-subset. One of the most known methods for bagging the trees is RF. Figure 5 shows the differences between the ensemble processes of boosting, bagging and random forest [Yang, Hwa-Yang, Zhou & Zomaya, 2010].

Figure 5

RF involves a set of stochastically different trees, each built from its own bootstrap samples [Mitchell, 2014], i.e., a combination of decision trees that is built using different sets of randomly selected input) variables [Gromski et al., 2015; Granitto, Gasperi, Biasioli, Trainotti, & Furlanello, 2007; Kucheryavskly, 2018]. Figure 6 graphically shows the operation of the RF method [Mitchell, 2014]: six decision trees forming a (very small) Random Forest for classification; trees A, B and E assign to the red class, however trees C and D assign to green class and tree F assigns to yellow class, so that the Random Forest will classify the object as red by a majority. An additional advantage of RF, compared to other classification methods such as PLS-DA or SVM, is the ability to directly discriminate in a single process between a set of samples/objects into a number of class higher than two (i.e. a multiclass classification problem).

Figure 6

## 4. Applications in food authentication

Table 3 reviews the most recent papers (i.e., published since 2010) in which SVM, DT, CART, and RF are applied in food analytical chemistry, among other more conventional chemometric methods. As mentioned in the Introduction section, RF has been scarcely used in food analytical chemistry. Yet, in recent years several papers have shown its potential in this area. Moreover, new software has been developed to apply RF using spectroscopic techniques in food chemistry [Smith, Baker, & Palmer, 2018].

Table 3

One of the most important aspects in order to assure the reliability of a pattern recognition supervised model is the validation step. This is fundamental for the assessment of the quality of the classification/quantification rate obtained in the multivariate models. The samples/objects used in this stage, which constitute the validation set or the test set, should be other than those used in the training stage. Thus the recommended is to use an external validation set. Nevertheless, sometimes the total number of samples of study is very limited and it is not possible to generate an external validation, and therefore it is carried out an internal cross-validation. Consequently, the quality performance features of the different multivariate models are calculated from the results obtained in cross-validation step. In this sense, the papers collected in table 1 there are 34 which applying internal cross-validation and 41 external validation.

Next, a comprehensive description on the gathered papers is carried out. For this, two blocks have been considered: support vector machine methods and decision trees methods. In addition, the meaning of all the abbreviations or acronyms is stated at the foot of the table3.

### 4.1 Support vector machine methods

#### 4.1.1 Fruits and juices

The studies carried out for the fruits and juices authentication are focused on the determination of some compounds as pesticides and additives, on the detection of adulterations and on the classification according to the geographical origin [Fan, Lai, Rasco, & Huang, 2015; Khanmohammadi et al., 2014; Naderi-Boldaji et al., 2015; Hong & Wong, 2014]. For example Guo et al. [Guo, Ni, & Kokot, 2016] developed different models for the quality control of jujube (Z. jujube Mill.) applying LDA, LS-SVM and BPANN to build classification models in order to classify the samples from four geographical regions. Moreover, PLS, LS-SVM and BPANN were used to quantify the content of total sugars, total phenols, and total acids, and the total antioxidant activity and concluded that the LS-SVM prediction models produced best results compared to the conventional chemometrics.

### 4.1.2 Honey and sugar

Honey is a sweet substance produced by bees from the nectar of flowers. Most studies, in which SVM is used, are based on the classification of commercial honeys from different geographical regions applying mainly spectroscopic and chromatographic analytical techniques. In this regard, El Alami et al. [El Alami et al., 2018] developed a model to discriminate between honeys from France and Morocco, classified correctly all the samples. On the other hand, Wei et al. [Wei, Wang & Wang, 2010] honey samples from different regions from China. In addition, the floral origin of the honey samples was predicted using rheometric features. Concerning the authentication of sugar, Ramírez-Morales et al. [Ramírez-Morales, Rivero, Fernández-Blanco, & Pazos, 2016], applying NIR spectroscopy and support vector regression (SVR) to perform the quality control of ºBrix and sucrose parameters of sugar industry.

### 4.1.3 Liquors and spirit beverages

Most of the papers are focused on performing a quality control of the different beverages in order to detect adulterations with other fake drinks [Pérez-Caballero et al., 2017; Andrade, Ballabio, Gómez-Carracero, & Pérez Caballero, 2017; Contreras et al., 2010; Ceballos-Magaña et al., 2012]. It should be highlighted the study published by Cheng et al. [Cheng, Fan & Yan, 2013] in which the authors tested two ways of data reduction using PCA and PLS prior to the application of SVM to classify different kinds of liquors from China. Overall, the authors concluded that the reduction of data by PLS was the best.

### 4.1.4 Meat

Regarding the authentication of meat, the reported works are focused on the analysis of volatile compounds collected using GC-MS or electronic nose to carry out a quality control in order to differentiate between fresh and refrigerated meat [Papadopolou, Panagou, Mohareb, & Nychas, 2013; Arredondo et al., 2014; Moharabeb, Papadopolou, Panagou, & Nychas, 2016].

### 4.1.5 Milk and dairy products

Majcher et al. [Majcher, Kaczmarek, Klenporf-Pawlik, Pikul, & Jelén, 2015] developed a rapid method for the authentication of cheeses protected under a 'Denomination of Origin' using SPME-MS as analytical measuring technique. The classification methods used were LDA, SIMCA and SVM. The highest classification accuracy (97.9%) for the test set was obtained using SVM.

### 4.1.6 Plant products

This subsection collects the studies related to the authentication of pepper, tea, cocoa, coffee and rice using spectrometric (UV-Vis, FTIR, NIR, Raman and ICP), voltammetric and chromatographic (HPLC) analytical techniques [Li, Sun, Pu, & Jayas, 2017; Liu et al., 2014; Zheng et al., 2009; [Gonçalvez et al., 2016]; Wood, Allaway, Boult, & Scott, 2010; Teye & Huang, 2015; Barbosa et al., 2014; Bona et al., 2017; Barbosa et al., 2016; Maione, Lemos Batista, Campiglia, Barbosa, & Barbosa, 2016b; Kyu et al., 2017; Feng, Zhang, Cong, & Zhu, 2013]. In these works SVM was applied for the purpose of performing the quality control and authenticity evaluation of the foodstuffs.

One of the most significant works was carried out by Teye et al. [Teye, Huang, Han, & Botchway, 2014], in which FDA, kNN and SVM classification models to distinguish between cocoa bean samples were developed. The results revealed that SVM was better than kNN and FDA since 100% of the samples were correctly classified.

The tea authenticity studies are focussed on distinguishing the botanical or geographical origin. Among the studies is outstanding the one published by Liu et al. [Liu et al., 2014] that applied the voltammetry as analytical technique from which the whole analytical signal was used to build the SVM classification model.

### 4.1.7 Vegetable oils

It is noteworthy that for the authentication of vegetable oils is mostly applied the chromatographic techniques in contrast to the rest of food, in which is more common the use of the spectroscopic techniques. Moreover, high performance liquid chromatography (HPLC) coupled to different detection system is more usual than gas chromatography (GC) for the authentication of the olive oil.

All the reported studies are focused on; (i) the detection of adulterations; (ii) the verification of the geographical origin; and (iii) the discrimination of different kinds of edible oils [Dong, Zhang, Zhang, & Wang, 2013; Devos, Downey, & Duponchel, 2014; Sayago, González-Domínguez, Beltrán, & Fernández-Recamales, 2018; Ordukaya & Karlik, 2017; Jiménez-Carvelo, Pérez-Castaño, González-Casado, & Cuadros-Rodríguez, 2017a; Jiménez-Carvelo, González-Casado, Pérez-Castaño, & Cuadros-Rodríguez, 2017b]. Furthermore, the quantification of olive oil in blends with other vegetable oils and the classification according to the cultivar are reported [Dong, Zhang, Zhang, & Wang, 2012; Jiménez-Carvelo, Osorio, Koidis, González-Casado, & Cuadros-Rodríguez, 2017c; Jiménez-Carvelo, González-Casado, & Cuadros-Rodríguez, 2017d; Jiménez-Carvelo, Cruz, Olivieri, González-Casado, & Cuadros-Rodríguez, 2019].

#### 4.1.8 Wine

Another relevant aspect of food authenticity is the varietal authentication. This is often the case of the studies published about the analysis of the wine, since the chemical composition it is greatly influenced by the kind of grape, besides of the agronomic conditions.

An attractive strategy to take a prominent position over the competitors is to take advantage of the difference in chemical composition, bearing a recognised quality-differentiated food seal as the 'Protected Designation Origin' (PDO) or 'Protected Geographical Indication' (PGI). Thus, it is important to develop rapid methods to authenticate such protected foodstuffs. In this sense, Costa et al. [Costa, García Llobodanin, Alves Castro, & Barbosa, 2018] reported a study based on the analysis of different parameters of the wine and ~~the~~ then SVM was applied to discriminate wine from Brazil of wine from Uruguay; the classification model achieved an accuracy rate of 79.97%. Martelo-Vidal et al. [Martelo-Vidal, & Vazquez] developed LDA, SIMCA and SVM classification models based on the polyphenolic profile to differentiate between wines from Spanish PDO 'Rias Baixas' and 'Ribeira Sacra'.

#### 4.1.9 Others

In this category are included the studies carried out to ensure the quality and authenticity of tofu and vinegar.

Xu et al. [Xu et al., 2012] applied FTIR spectroscopy to analyse the shelf-life of the tofu. The samples were measured with different age (from 29 to 161 days). In the subsequent statistical analysis different pre-processing methods were tested before to the development of the PLS and SVM multivariate models. The models were evaluated and SVM was the best option, since it was obtained the lowest root mean squared error of prediction (RMSEP). Although the difference with PLS was such a small that the authors concluded that PLS should be used since it is low complexity.

Bao et al. [Bao et al., 2014] developed an analytical method for the quality control of the ºbrix and pH of the white vinegar. What is remarkable about this study was the application of PLS prior to the use of LS-SVM in order to select the latent variables (LVs), which were used as the inputs of the LS-SVM to develop the calibration model. The predictive capability of the models was evaluated estimating the correlation coefficient (r), the root mean square error of calibration and prediction (RMSEC & RMSEP), and the residual predictive deviation (RPD).

### 4.2 Decision tree methods: CART and RF

#### 4.2.1 Fruits and juices

Organic foods are appreciated by customers increasing their sales in the last years. The published scientific paper devoted to authentication of fruits and juices are focused on the differentiation of ecologic products from non-ecologic ones and the detection of additives.

Maione et al. [Maione et al., 2016a] developed an analytical method using ICP-MS to discriminate organic from conventional grape juice. In addition, they carried out the comparison between different data mining methods (SVM, CART and MLP) and the results obtained showed that all the methods provided good results for the intended purpose.

#### 4.2.2 Honey

The quality of honey varies depending on the floral and geographical origin. The conventional methods are based on the measure of several physicochemical parameters what are time-consuming and involve a high consumption of solvents. For this reason, Popek et al. [Popek, Halagarda, & Jursa, 2017] and Chuddzinks et al. [Chudzinksa & baralkiewicz, 2011] proposed different analytical methods combined with CART in order to reduce the time and the complexity of the analysis.

#### 4.2.3 Spirit beverages

The study carried out by Martínez-Jarquín et al. [Matínez-Jarquín, Moreno-Pedraza, Cázarez-García, & Winkler, 2017] was based on the application of the mass spectrometry along with PCA and RF to discriminate agave tequilas from traditionally processed mezcal. The most noticeable of this work was the application of RF to select the number of the variables, which were used in the new PCA model. Surprisingly PCA was applied as a classification method when it is an unsupervised pattern recognition method which only should be used to explore the variability of the samples in the dataset and/or to screening the inherent sample grouping when the dimensionality of the data is reduced. However RF, which is in itself is a multivariate classification method, is not applied to this end.

#### 4.2.4 Milk and dairy products

Fabris et al. [Fabris et al., 2010] developed different multivariate classification methods with the data acquired using PTR-TOF-MS in order to perform the quality control of Trentingrana cheese, when it is produced with milk stores in different conditions. Four binary classification models were built using PDA, DPLS, SVM and RF in order to select the best data mining method. Finally, they concluded that all the methods provided similar performance.

#### 4.2.5 Rice

Rice is a staple food in many developing and least developed countries. There are a lot of countries producers of rice, being Chine the world's largest producer; therefore the differentiation in the global market is important for the producers. In this sense the reported studies are focused on classifying the rice according to the geographical origin [Kyu et al, 2017; Mahdavi, Farimani, Fathi, & Chassempour, 2015; Weng et al., 2018].

Maione et al. [Maione, Lemos Batista, Campiglia, Barbosa, & Barbosa, 2016b] developed different classification methods to discriminate rice samples according to their geographical origin. For this purpose, the authors applied SVM, RF and MLP methods. They evaluated these multivariate classification methods using the following performance metrics: accuracy, sensitivity, specificity, and area under the receiving operating curve (AUC); in all cases, SVM and RF yielded better results than MLP.

#### 4.2.6 Tea

RF was applied to classify the tea according to the botanical and geographical origin and to discriminate between different varieties of tea [Gonçalvez et al., 2016; Zheng et al., 2009; Wang, Huang, Fan, & Lu, 2015].

Ni et al. [Ni, et al., 2018] built several classification models to distinguish between green tea from different regions in China. They compared the results using LDA, PLS-DA, and DT. The best results were obtained when DT was applied.

4.2.7 Vegetable oils

Edible vegetable oils are globally a kind of important food, a lot of them are present in several diets of the different regions of the world, such as olive oil in Mediterranean diet, which is characteristic of Spain, Italy and Greece or seeds oils in Asian. The process of obtaining of some high-price vegetable oils is expensive and consequently these are subject to possible adulterations in order to reduce the production cost. For this reason, ensuring the authenticity of the vegetable oils is currently required in order to detect such adulterations.

In this regard the published papers are focused on the detection of adulterations, the discrimination of edible oils according to the botanical and geographical origin [Zhang et al., 2014; Hu et al., 2014; Ruiz-Samblás, Cadenas, Pelta & Cuadros-Rodríguez, 2014; Nasibov, Kantarci, Vahaplar, & Kinay, 2016; Sayago, González-Domínguez, Beltrán, & Fernández-Recamales, 2018; Jiménez-Carvelo, Cruz, Olivieri, González-Casado, & Cuadros-Rodríguez, 2019]. Although there is one study which stands out since the RF method was used in an unconventional way, Ai et al. [Ai et al., 2014] analysed the fatty acid composition of six different kinds of vegetable oils (tea, olive, rapeseed, corn, sunflower and sesame oil) and they applied RF as unsupervised technique to carry out a cluster analysis in order to test if there were natural grouping of the oils samples.

4.2.8 Wine

Within the wine industry the authenticity evaluation in terms of geographical and brand origin influence in the choice of the consumers. Such is the case of Loannou-Papayianni et al. [Loannou-Papayianni, Kokkinfta, & Theocharis, 2011] who developed an analytical method using FTIR and CART to authenticate Cypriot traditional wine and to differentiate it from its competitors. Gómez-Meire et al. [Gómez-Meire, Falqué, Díaz & Fdez-Riverola, 2014] appled GC-MS combined with RF and MLP to ensure and to classify wine elaborated in Galicia (a region of the Nord of Spain); they concluded that the application of machine learning methods allows ensuring the authenticity of different white wines elaborated from several grape varieties and origins.

## 4. Final remarks

SVM, CART, and RF are an alternative group of pattern recognition methods that are yielding promising results in the area food quality and authenticity. Considering only these three methods, SVM is by far the most widely used and, in most cases, it is stated that SVM has an improved performance compared to other better-known conventional methods such as PLS-DA.

In addition, CART and RF are alternative pattern recognition methods that are currently used in the area of food. In other related areas such as metabolomics, some authors have already highlighted the advantages of these machine learning methods as compared to conventional techniques. However, there are still very few reported studies in which CART and RF are used in studies on food analytical chemistry even though their value has been widely proven and they have yielded outstanding results.

649 **References**

Ai, F.F., Bin, J., Zhang, Z.M., Huang, J.H., Wang, J.B., Liang, Y.Z., & Yu, L. (2014). Application of random forests to select premium quality vegetable oils by their fatty acid composition. *Food Chemistry*, *143*, 472-478. https://doi.org/10.1016/j.foodchem.2013.08.013

Alewijn, M., Van der Voet, H., & Van Ruth, S. (2016). Validation of multivariate classification methods using analytical fingerprints – Concept and case study on organic feed for laying hens. *Journal of Food Composition and Analysis, 51*,15–23. https://doi.org/10.1016/j.jfca.2016.06.003

Andrade, J.M., Ballabio, D., Gómez-Carracedo, M.P., & Pérez-Caballero, G. (2017). Nonlinear classification of commercial Mexican tequilas. *Journal of Chemometrics, 31: e2939,* 1-14. https://doi.org/10.1002/cem.2939

Arredondo, T., Oñate, E., Santander, R., Tomic, G., Silva, JR., Sanchez, E., & Acevedo, CA. (2014). Application of neural networks and meta-learners to recognize beef from OTM cattle by using volatile organic compounds. *Food* and *Bioprocess Technology*, *7*, 3217-3225. https://doi.org/10.1007/s11947-014-1289-7

Ballabio, D., Grisoni, F., & Todeschini, R. (2018). Multivariate comparison of classification performance measures. *Chemometrics and Intelligent Laboratory Systems*, *174*, 33-44. https://doi.org/10.1016/j.chemolab.2017.12.004

Ballabio, D., & Consonni, V. (2013).Classification tools in chemistry. Part 1: linear models. PLS-DA. *Analytical Methods*, *5*, 3790-3798. https://doi.org/10.1039/c3ay40582f

Bao, Y., Liu, F., Kong, W., Sun, D.W., He, Y., & Qiu, Z. (2014). Measurement of soluble solid contents and pH of white vinegars using VIS/NIR spectroscopy and least squares support vector machine. *Food and Bioprocess Technology, 7,* 54-61. https://doi.org/ 10.1007/s11947-013-1065-0

Barbosa, R.M., de Paula, E.S, Paulelli, A.C., Moore, A.F., Oliveira Souza, J.M., Lemos Batista, B., Campiglia, AD., & Barbosa, Jr, F. (2016). Recognition of organic rice samples based on trace elements and support vector machine. *Journal of Food Composition and Analysis*, *45*, 95-100. https://doi.org/10.1016/j.jfca.2015.09.010

Barbosa, R.M., Lemos Batista, B., Varrique, R.M., Coelho, V.A., Compiglia, A.D., & Barbosa, Jr, F. (2014). The use of advanced chemometric techniques and trace element levels for controlling the authenticity of organic coffee. *Food Research International, 61,* 246-251. https://doi.org/10.1016/j.foodres.2013.07.060

Barbosa, RM., Batista, B.L., Bariao, CV., Varrique, RM., Coelho, VA., Campiglia, AD., & Barbosa, Jr, F. (2015). A simple and practical control of the authenticity of organic sugarcane samples based on the use of machine-learning algorithms and trace elements determination by inductively coupled plasma mass spectrometry. *Food Chemistry, 184*, 154-159. https://doi.org/10.1016/j.foodchem.2015.02.146

Belvilacqua, M., Nescatelli, R., Bucci, R., Magrì, A.D., Magrì, A.L., & Marini, F. (2014). Chemometric Classification Techniques as a Tool for Solving Problems in Analytical Chemistry. *Journal of AOAC International*, *97*, 19-28. https://doi.org/10.5740/jaoacint.SGEBevilacqua

Berrueta, L.A., Alonso-Salces, R.M., & Héberger, K. (2007). Supervised pattern recognition in food analysis. Journal of Chromatography A, 1158,196-214. https://doi.org/10.1016/j.chroma.2007.05.024

Bevilacqua, M., Bro, R., Marini, F., Rinnan, A., Rasmussen, M., & Skov, T. (2017). Recent chemometrics advances for foodomics. *Trends in Analytical Chemistry, 96*, 42-51. https://doi.org/10.1016/j.trac.2017.08.011

Bona, E., Marquetti, I., Varaschim Link, J., Figueiredo Makimori, G.Y., de Costa Arca, V., Guimaraes Lemes, A.L., Garcia Ferreira, J.M., dos Santos 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, 76*, 330-336. https://doi.org/10.1016/j.lwt.2016.04.048

Bosque Sendra, J.M., Cuadros Rodríguez, L., Ruiz Samblás, C., & De la Mata, P. (2012). Combining chromatography and chemometrics for the characterization and authentication of fats and oils from triacyglycerol compositional data- A review. Analytica Chimica Acta, 724, 1-11.

https://doi.org/10.1016/j.aca.2012.02.041

Brereton, R., & Lloyd, G.R. (2010). Support vector machine for classification and regression. Analyst, 135, 230-267. https://doi.org/10.1039/b918972f

Brereton, R.G., Jansen, J., Lopes, J., Marini, F., Pomerantsev, A., Rodionova, O., Roger, JM., Walczack, B., & Tauler, R. (2017). Chemometrics in analytical chemistry–part I: history experimental design and data analysis tools. *Analitycal and Bioanalytical Chemistry, 409*, 5891-5899. https://doi.org/10.1007/s00216-017-0517-1

Brereton, R.G., & Lloyd, G.R. (2014). Partial least squares discriminant analysis: taking the magic away. *Journal of Chemometrics, 28*, 213–225. https://doi.org/10.1002/cem.2609

Brereton, R.G. (2015). Pattern recognition in chemometrics. *Chemometrics and Intelligent Laboratory Systems*, *149*, 90-96.https://doi.org/10.1016/j.chemolab.2015.06.012

Bro, R., & Smilde, A. (2014). Principal component analysis. *Analytical Methods, 6*, 2812–2831. https://doi.org/10.1039/c3ay41907j

Callao, M.P, & Ruisánchez, I. (2018). An overview of multivariate qualitative methods for fraud detection. *Food Control, 86*, 283-293. https://doi.org/10.1016/j.foodcont.2017.11.034

Ceballos-Magaña, S.G., Jurado, J.M., Muñiz-Valencia, R., Alcázar, A., de Pablos, F., & Martín, M.J. (2012). Geographical authentication of tequila according to its mineral content by means of support vector machines. *Food Analytical Methods, 5,* 260-265. https://doi.org/10.1007/s12161-011-9233-1

Cheng, P., Fan, W., & Yan, Xu. (2013). Quality grade discrimination of Chinese strong aroma type liquors using mass spectrometry and multivariate analysis. *Food Research International, 54,* 1753-1760. https://doi.org/10.1016/j.foodres.2013.09.002

Chudzinska, M., & Baralkiewicz, D. (2011). Application of ICP-MS method of determination of 15 elements in honey with chemometric approach for the verification of their authenticity. *Food and Chemical Toxicology*, *49*,2741-2749. https://doi.org/10.1016/j.fct.2011.08.014

Chung, I.M., Kim, J.K., Lee, J.K., & Kim, SH. (2015). Discrimination of geographical origin of rice (Oryza sativa L.) by multielement analysis using inductively coupled plasma atomic emission spectrocopy and multivariate analysis. *Journal of Cereal Science, 65*, 252–259. https://doi.org/10.1016/j.jcs.2015.08.001

Cocchi M. (2017). Chemometrics for food quality control and authentication. In R.A. Meyers (Ed.), *Encyclopedia of Analytical Chemistry*, John Wiley & Sons, Online © 2006-2017. https://doi.org/10.1002/9780470027318.a9579

Contreras, U., Barbosa-García, O., Pichardo-Molina, J.L., Ramos-Ortíz, G., Maldonado, J.L., Meneses-Nava, M.A., Ornelas-Soto, N.E, & López-de-Alba, P.L. (2010). Screening method for identification of adulterated and fake tequilas by using UV-Vis spesctroscopy and chemometrics. *Food Research International, 43,* 2356-2362. https://doi.org/ 10.1016/j.foodres.2010.09.001

Costa, N.L., García Llobodanin, L.A., Alves Castro, I., & Barbosa, R. (2018). Geographical classification of Tannat wines based on support vector machines and feature selection. *Beverages, 4,97,* 1-15. https://doi.org/ 10.3390/beverages4040097

Cuadros Rodríguez, L., Pérez Castaño, E., & Ruiz Samblas, C. (2016). Quality performance metrics in multivariate classification methods for qualitative analysis. *Trends in Analytical Chemistry,80*, 612–624. https://doi.org/10.1016/j.trac.2016.04.021

Cubero-Leon, E., Peñalver, R., & Maquet, A. (2016). Review on metabolomics for food authentication. Food Research International, 60, 95-107. https://doi.org/10.1016/j.foodres.2013.11.041

Cutler, D.R., Edwars, T.C., Beard, K.H., Cutler, A., Hess K.T., Gibson, J., & Lawler, J.J. (2007). Random forest for classification in ecology. Ecology, 88, 2783-2792. https://doi.org/10.1890/07-0539.1

da Costa, N.L., Castro, I.A., & Barbosa, R. (2016). Classification of cabernet sauvignon from two different countries in South America by chemical compounds and support vector machine. *Applied Artificial Intelligence, 30*,679-689. http://dx.doi.org/10.1080/08839514.2016.1214416

Dai, Q., Sun, D.W., Xiong, Z., Cheng, J.H., & Zeng, X.A. (2014). Recent advances in data mining techniques and their applications in hyperspectral image processing for the food industry. Comprehensive Reviews in Food Science and Food Safety, 13, 891-905. https://doi.org/10.1111/1541-4337.12088

Dankowska, A., & Kowalewski, W. (2018). Comparasion of different classifications methods for analyzing fluorescence spectra to characterize type and freshness of olive oils. *European Food Research and Technology.* https://doi.org/10.1007/s00217-018-3196-z.

Debska, B. & Guzowska-Swider, B. (2011). Decision trees in selection of featured determined food quality. Analytica Chimica Acta, 705, 261-271. https://doi.org/10.1016/j.aca.2011.06.030

Devos, O., Downey, G., & Duponchel, L. (2014). Simultaneous data pre-processing and SMV classification model selection based on a parallel genetic algorithm applied to spectroscopic data of olive oils. *Food Chemistry, 148*, 124-130. https://doi.org/10.1016/j.foodchem.2013.10.020

Dong, W., Zhang, Y., Zhang, B., & Wang, X. (2012). Quantitative analysis of adulteration of extra virgin olive oil using Raman spectroscopy improved by bayesin framework least squares support vector machines. *Analytical Methods, 4,* 2772-2777. https://doi.org/10.1039/C2AY25431J

Dong, W., Zhang, Y., Zhang, B., & Wang, X. (2013). Rapid prediction of fatty acid composition of vegetable oil by Raman spectroscopy coupled with least saures support vector machines. *Journal of Raman Spectroscopy, 44,* 1739-1745. https://doi.org/10.1002/jrs.4386

Drab, K., & Daszykowski, M. (2014). Clustering in Analytical Chemistry. *Journal of AOAC International*, 97, 29–38. https://doi.org/10.1016/0898-5529(89)90042-0

Duffus, J.H., Nordberg, M., & Templenton, D.M. (2007). Glossary of terms used in toxicology (IUPAC Recommendations 2007). *Pure and Applied Chemistry*, 79, 1153-1344. https://doi.org/10.1351/pac200779071153

Efenberger-Szmechtyk, M., Nowak, A., & Kregiel, D. (2018). Implementation of chemometrics in quality evaluation of food and beverages. *Critical Reviews in Food Science and Nutrition, 58,* 1747-1766. https://doi.org/10.1080/10408398.2016.1276883

El Alami El Hassani, N., Tahri, K., Llobet, E., Bouchikhi, B., Errachid, A., Zine, N., & El Bari, N. (2018). Emerging approach for analytical characterization and geographical classification of Moroccan and French honeys by means of a voltammetric electronic tongue. *Food Chemistry, 243*, 36-42. https://doi.org/10.1016/j.foodchem.2017.09.067

Esbensen, K.H, & Geladi, P. (2010). Principles of proper validation: use and abuse of re-sampling for validation. *Journal of Chemometrics, 24*,168–187. https://doi.org/10.1002/cem.1310

Fabris, A., Biasioli, F., Granito, P.M., Aprea, E., Cappelin, L., Schuhfried, E., Soukoulis, C., Märk, T.D., Gasperi, F., & Endrizzi, I. (2010). PTR-TOF-MS and data mining methods for rapid characterisation of agro-industrial samples: influence of milk storage conditions on the volatile compounds profile of Trentingrana cheese. *Journal of Mass Spectrometry, 45*, 1065-1074. https://doi.org/10.1002/jms.1797

Fan, Y., Lai, K., Rasco, B.A., Huang, Y. (2015). Determination of carbaryl pesticide in Fuji apples using surface-enhanced Raman spectrocopy coupled with multivariate analysis. *LWT - Food Science and Technology, 60*, 352-357. https://doi.org/10.1016/j.lwt.2014.08.011

Feng, X., Zhang, Q., Cong, P., & Zhu, Z. (2013). Preliminary study on classification of rice and detection of paraffin in the adulterated samples by Raman spectroscopy combined with multivariate analysis. *Talanta,115*, 548-555. https://doi.org/10.1016/j.talanta.2013.05.072

Ghernaout, D., Aichouni, M., & Alghamin, A. (2018). Overlapping ISO/IEC 17025:2017 into big data: a review and perpectives. International Journal of Science and Qualitative Analysis, 4, 83-92. https://doi.org/10.11648/j.ijsqa.20180403.14

Gómez-Meire, S., Campos, C., Falqué, E., Díaz, F., Fdez-Riverola, F. (2014). Assuring the authenticity of northwest Spain white varieties using machine learning techniques. *Food Research International, 60,* 230-240. http://dx.doi.org/10.1016/j.foodres.2013.09.032

Glossary of common machine learning, statistics and data science terms. https://www.analyticsvidhya.com/glossary-of-common-statistics-and-machine-learning-terms/ (accessed on 2019, February, 25).

Gonçalvez Dias Diniz, P.H., Ferreira Barbosa, M., Tavares de Melo Milanez, K.D., Pistonesi, M.F., & Ugulino de Araújo, M.C. (2016). Using UV-Vis spectroscopy for simultaneous geographical and varietal classification of tea infusions simulating a home-made tea cup. *Food Chemistry, 192*, 374-379. https://doi.org/10.1016/j.foodchem.2015.07.022

Granato, D., Putnik, P., Bursac Kovacevic, D., Sousa Santos, J., Calado V., Silva Rocha, R., Gomes Da Cruz, A., Jarvis, B., Rodionova, O.Y., & Pomerantsev, A. (2018). Trends in Chemometrics: food authentication microbiology, and effects of processing. *Comprehensive Reviews in Food Science and Food Safety, 17,* 663-677. http://dx.doi.org/ 10.1111/1541-4337.12341

Granitto, P.M., Gasperi, F., Biasioli, F., Trainotti, E., & Furlanello, C. (2007). Modern data mining tools in descriptive sensory analysis: a case study with a random forest approach. Food Quality and Preference,18, 681-689. https://doi.org/10.1016/j.foodqual.2006.11.001

Gromski, P.S., Muhamadali, H., Ellis, D.I., Xu, Y., Correa, E., Turner, M.L., & Goodacre, R. (2015). A tutorial review: metabolomics and partial least squares-discriminant analysis – a marriage of convenience or a shotgun wedding. *Analytica Chimica Acta, 879,*10-23. https://doi.org/10.1016/j.aca.2015.02.012

Guo, Y., Ni, Y., &Kokot, S. (2016). Evaluation of chemical components and properties of the jujube fruit using near infrared spectroscopy and chemometrics. *Spectrochimica Acta, Part A: Molecular and Biomolecular Spectroscopy,153*, 79-86. https://doi.org/10.1016/j.saa.2015.08.006

Hakki Boyaci, I., Tümay Temiz, H., Selin Uysal, R., Murat Velioglu, H., Jafarzedh Yadegari, R., & Mahmoudi Rishkan, M. (2014). A novel method for discrimination of beef and horsemeat using Raman spectroscopy. *Food Chemistry,148*,37–41. https://doi.org/10.1016/j.foodchem.2013.10.006

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: concepts and techniques*. (3rd ed.). Waltham: Morgan Kauffmann Publishers.

Hibbert, D.B. (2016). Vocabulary of concepts and terms in chemometrics (IUPAC Recommendations 2016). *Pure and Applied Chemistry, 88*, 407-443. https://doi.org/10.1515/pac-2015-0605

Hong, X., & Wang, J. (2014). Detection of adulteration in cherry tomato juices based on electronic nose and tongue: comparison of different data fusion approaches. *Journal* of *Food Engineering*, *126*, 89-97. https://doi.org/10.1016/j.jfoodeng.2013.11.008

Hu, Wei., Zhang, L., Li, P., Wang, X., Zhang, Q., Xu, B., Sun, X., Ma, F., & Ding, X. (2014). Characterization of volatile compounds in four vegetable oils by headspace two-dimensional comprehensive chromatography time-of-flight mass spectrometry. *Talanta,129,* 629-635. https://doi.org/10.1016/j.talanta.2014.06.010

Huch, C.W., Pezzei, C.K., & Huck-Pezzei, V.A.C. (2016). An industry pespective of food fraud. *Current opinion in Food Science, 10,* 32-37. http://dx.doi.org/10.1016/j.cofs.2016.07.004

Jiang, L., Zheng, H., & Lu, H. (2015). Applications of UV spectrometry and chemometric models for detecting olive oil-vegetable oil blends adulteration. *Journal of Food Science and Technology*, *52*, 479-485. https://doi.org/10.1007/s13197-013-1003-1

Jiménez Carvelo, A.M, González Casado, A., & Cuadros Rodríguez, L. (2017a). A new analytical method for quantification of olive oil and palm oil in blends with other vegetable edible oils based on the chrtomatographic fingerprints from the methyl-transesterified fraction. *Talanta, 164*, 540-547. https://doi.org/10.1016/j.talanta.2016.12.024

Jiménez Carvelo, A.M., González Casado, A., Pérez Castaño, E., & Cuadros Rodríguez, L. (2017b). Fast-HPLC fingerprinting to discriminate olive oil from other edible vegetable oils by multivariate classification methods. *Journal of AOAC International*, *100*, 345-350. https://doi.org/10.5740/jaoacint.16-0411

Jiménez Carvelo, A.M., Osorio, M.T., Koidis, A., González Casado, A., & Cuadros Rodríguez, L. (2017c). Chemometric classification and quantification of olive oil in blends with any vegetable oils using FTIR-ATR and Raman. *LWT – Food Science and Technology*, *86*, 174-184. https://doi.org/10.1016/j.lwt.2017.07.050

Jiménez Carvelo, A.M, Pérez Castaño, E., González Casado, A., & Cuadros Rodríguez, L. (2017d). One input-class and two input-class classifications for differentiating olive oil from other edible vegetable oils by use of the normal phase liquid chrtomatography fingerprint of the methyl-transesterified fraction. *Food Chemistry, 221,* 1784-1791. https://doi.org/10.1016/j.foodchem.2016.10.103

Jiménez-Carvelo A.M., Cruz, C.M, Olivieri, A.C., González-Casado, A.,& Cuadros-Rodríguez, L. (2019). Classification of olive oils according to their cultivars based on second-order data using LC-DAD. *Talanta, 195*, 69-76. https://doi.org/10.1016/j.talanta.2018.11.033

Ji-yong, S., Xiao-bo, Z., Xiao-wei, H., Jie-wen, Z., Yanxiao, L., Limin, H., & Jianchun, Z. (2013). Rapid detecting total acid content and classifiying different types of vinegar based on near infrared spectroscopy and least-squares support vector machine. *Food Chemistry, 138,* 192-199. http://dx.doi.org/10.1016/j.foodchem.2012.10.060

Jurado, J. M., Alcázar, A., Palacios-Morillo, A., & de Pablos. (2012). Classification of Spanish DO white wines according to their elemental profile by means of support vector machine. *Food Chemistry, 135*, 898-903. http://dx.doi.org/10.1016/j.foodchem.2012.06.017

Kavakiotis, I., Tsave, O., Salifoglou , A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine learning and data mining methos in diabetes research. C*omputational and Structural Biotechnology Journal, 15*, 104-116. https://doi.org/10.1016/j.csbj.2016.12.005

Kelleher, J.D., Namee, B.M., & D'arcy, A. (2015). *Fundamentals of machine learning for predictive data analysis. Algorithms, worked examples, and case studies*. Cambridge: Massachusetts Institute of Technology.

Kemal Aloglu, A., de B Harrington, P., Sahin, S., Demir, C., & Gunes, M.E. (2017). Chemical profiling of floral and chestnut honey using high-performance liquid chromatography-ultraviolet detection. *Journal of food composition and analysis*, *62*, 205-210. https://doi.org/10.1016/j.jfca.2017.06.002

Khakimov, B., Gürdeniz, G., & Engelsen, S.B. (2015). Trends in the application of chemometrics to foodomics studies. *Acta Alimentaria, 44*,4-31. https://doi.org/10.1556/AAlim.44.2015.1.1

Khanmohammadi, M., Karami, F., Mir-Marqués, A., Garmarudi, A.B., Garrigues, S., & de la Guardia, M. (2014). Classification of persimmon fruit origin by near infrared spectrometry and least squares-support vector machines. *Journal of Food Engineering*,*142*,17–22. https://doi.org/10.1016/j.jfoodeng.2014.06.003

Kingston, H.M., & Kingston, M.L. (1994). Nomenclature in laboratory robotics and automation (IUPAC Recommendations 1994). *Pure and Applied Chemistry*, *66,*609-630. http://dx.doi.org/10.1155/S1463924694000040

Kotsiantis, S.B. (2013). Decision trees: a recent overview. *Artificial Intelligence Review, 39*, 261-283. https://doi.org/10.1007/s10462-011-9272-4

Kucheryavskly, S. (2018). Analysis of NIR spectroscopic data using decision trees and their ensembles. *Journal of Analysis and Testing, 2,* 274-289. https://doi.org/10.1007/s41664-018-0078-0

Kumar, N., Bansal, A., Sarma, G.S., & Rawal, R.K. (2014). Chemometrics tools in analytical chemistry: An overview. *Talanta, 123*, 186-199. https://doi.org/10.1016/j.talanta.2014.02.003

Kyu Lim, D., Phuoc Long, N., Mo, C., Dong, Z., Cui, L., Kim, G., & Kwon Won, S. (2017). Combination of mass spectrometry-based targeted lipidomics and supervised machine learning algorithms in detecting adulterated admixtures of white rice. *Food Research International,100,*814-821. https://doi.org/10.1016/j.foodres.2017.08.006

Lavine, B.K. (2000). Chemometrics. *Analytical Chemistry. 72*, 91–97. https://doi.org/10.1021/a1000016x

Li, J.L., Sun, D.W., Pu, H., & Jayas, D.S. (2017). Determination of trace thiophanate-methyl and its metabolite carbendazim with teratogenic risk in red bell peper (Capsicum annuum L.) by surface-enhanced Raman imaging technique. *Food Chemistry, 218,* 543-552. https://doi.org/10.1016/j.foodchem.2016.09.051

Li, Y., Zhang, J.Y., & Wang, Y.Z. (2018). FT-MIR and NIR spectral data fusion: a synergetic strategy for the geographical traceability of Panax notoginseng. *Analytical and Bioanalytical Chemistry, 410*, 91-103. https://doi.org/10.1007/s00216-017-0692-0

Liu, N., Liang, Y., Bin, J., Zhang, Z., Huasang, J., Shu, R., & Yang, K. (2014). Classification of green and black teas by PCA and SVM analysis of cyclic voltammetric signals from metallic oxide-modified electrode. *Food Analytical Methods, 7*, 472-480. https://doi.org/10.1007/s12161-013-9649-x

Liu, J., Pan, T.J., & Zhang, Z.Y. (2018). Incremental support vector machine combined with ultraviolet-visible spectroscopy for rapid discriminant analysis of red wine. *Journal of Spectroscopy, ID 4230681,* 1-5. https://doi.org/10.1155/2018/4230681

Liu, M., Wang, M., Wang, J., & Li, D. (2013). Comparison of random forest, support vector machine and back propagation neural network for electronic tongue data classification: Application to the recognition of orange beverage and Chinese vinegar. *Sensors and Actuators B: Chemical, 177,* 970-980. http://dx.doi.org/10.1016/j.snb.2012.11.071

Liu, S., Whitty, M. (2015). Automatic grape bruch detection in vineyards with an SVM classifier. Journal of applied logic, 13, 643-653. http://dx.doi.org/10.1016/j.jal.2015.06.001

Loannou-Papayianni, E., Kokkinofta, R.I., & Theocharis, C.R. (2011). Authenticity of Cypriot sweet wine Commandaria using FT-IR and chemometrics. *Journal of Food Science, 76 (3),* C420-C427. https://doi.org/10.1111/j.1750-3841.2011.02048.x

Lu, P., Chen, S., & Zheng, Y. (2012). Artificial intelligence in civil engineering. *Mathematical Problems in Engineering*, 1-22. https://doi.org/10.1155/2012/145974

Luts, J., Ojeda, F., de Plas, R.V., de Moor, B., Van Huffel, S. ,& Suykens, J.A.K. (2010). A tutorial on support vector machine-based methods for classifications problemas in chemometrics. *Analytica Chimica Acta, 665*, 129-145. https://doi.org/10.1016/j.aca.2010.03.030

Mahdavi, V., Farimani, M.M., Fathi, F., & Chassempour, A. (2015). A targeted metabolomics approach toward undestanding metabolic variations in rice under pesticide stress. *Analytical Biochemistry, 478,* 65-72. https://doi.org/10.1016/j.ab.2015.02.021

Maione, C., de Paula, E.S., Gallimberti, M., Batista, B.L,, Campiglia, A,D,, Barbosa, Jr F., & Barbosa, R.M. (2016a). Comparative study of data mining techniques for the authentication of organic grape juice based on ICP-MS analysis. *Expert Systems With Applications*, *49*, 60-73. https://doi.org/10.1016/j.eswa.2015.11.024

Maione, C., Lemos Batista, B., Campiglia, A.D., Barbosa Jr, F., & Barbosa, R.M. (2016b). Classification of geographic origin of rice by data mining and inductively coupled plasma mass spectrometry. *Computers and Electronics in Agriculture*, *121*,101-107. https://doi.org/10.1016/j.compag.2015.11.009

Majcher, M.A., Kaczmarek, A., Pawlik, D., Pikul, J., & Jelén, 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. https://doi.org/10.1007/s12161-015-0114-x

Maning, L. (2016). Food Fraud: policy and food chain. *Current Opinion in Food Science, 10,* 16-21. http://dx.doi.org/10.1016/j.cofs.2016.07.001

Marini, F. (2010). Classification methods in chemometrics. *Current Analytical Chemistry, 6*,72-79. https://doi.org/10.2174/157341110790069592

Marini, F. (2009). Artificial neural networks in foodstuff analyses: trends and perspectives. A review. *Analytica Chimica Acta, 635*, 121–131. https://doi.org/10.1016/j.aca.2009.01.009

Martelo-Vidal, M.J., Dominguez-Agis, F., & Vázquez, M. (2013). Ultraviolet/visible/near-infrared spectral analysis and chemometric tools for the discrimination of wines between subzones inside a controlled designation of origin: a case study of Rías Baixas. *Australian* Journal *of Grape and Wine Research*,19, 62-67. https://doi.org/10.1111/ajgw.12003

Martelo-Vidal, M.J., & Vázquez, M. (2016). Polyphenolic profile of red wines for the discrimination of controlled designation of origin. *Food Analytical Methods, 9*, 332-341. https://doi.org/10.1007/s12161-015-0193-8

Martínez-Jarquín, S., Moreno-Pedraza, A., Cázarez-García, D., & Winkler, R. (2017). Automated chemical fingerprinting of Mexican spirits derived from Agave (tequila and mezcal) using direct-injection electrospray ionisation (DIESI) and low-temperature plasma (LTP) mass spectrometry. *Analytical methods, 9,* 5023-5028. https://doi.org/ 10.1039/c7ay00793k

Medina, S., Perestrelo, R., Silva, P., Pereira, J.A.M, & Câmara, J.S. (2019). Current trends and recent advances on food authenticity technologies and chemometric approaches. *Trends in Food Science & Technology, 85,* 163-176. https://doi.org/10.1016/j.tifs.2019.01.017

Mehmood, T., & Ahmed, B. (2016). The diversity in the applications of partial least squares: an overview. *Journal of Chemometrics, 30*, 4–17. https://doi.org/10.1002/cem.2762

Messai, H., Farman, M., Sarraj-Laabidi, A., Hammami-Semmar, A., & Semmar, N. (2016). Chemometrics methods for specificity, authenticity and traceability analysis of olive oils: principles, classifications and applications. *Foods., 5(77),* 1-35. https://doi.org/10.3390/foods5040077

Mikut, R., & Resichl, M. (2011). Data mining tools. *WIREs Data Mining and Knowledge*,*1*, 431-443. https://doi.org/10.1002/widm.24

Mitchell, J.B.O. (2014). Machine learning methods in chemoinformatics. *WIREs Computational Molecular Science, 4*,468-480. https://doi.org/10.1002/wcms.1183

Mitra, S., & Acharya, T. (2003). *Data mining: multimedia, soft computing and bioinformatics*. New Jersey: John Wiley & Sons.

Mohareb, F., Papadopoulou, O.S., Panagou, E., & Nychas, G.J.E. (2016). Ensemble-based support vector machine classifiers as an efficient tool for quality assessment of beef fillets from electronic nose data. *Analytical Methods, 8*, 3711–3721. https://doi.org/10.1039/C6AY00147E

Mutihac, L.,& Mutihac, R. (2008). Mining in chemometrics. *Analytica Chimica Acta*, *6012*, 1-18. https://doi.org/10.1016/j.aca.2008.02.025

Naderi-Boldaji, M., Mishra, P., Ahmadpour-Samani, M., Ghasemi-Varnamkhasti, M., Ghanbarian D., Izadi, Z. (2018). Potential of two dielectric spectrosocpy techniques and chemometric analyses for detection of adulteracion in grape syrup. *Measurement, 127,* 518-524. https://doi.org/10.1016/j.measurement.2018.06.015

Nasibov, E., Kantarci Savas, S., Vahaplar, A., & Kinay, A.O. (2016). A survey on geographic classification of virgin olive oil with using T-operators in fuzzy decision tree approach. *Chemometrics and* Intelligent *Laboratory Systems*, *155*, 86-96. https://doi.org/10.1016/j.chemolab.2016.04.004

Ni, K., Wang, J., Zhang, Q., Yi, X., Ma, L., Shi, Y., & Ruan, J. (2018). Multi-element composition and isotopic signatures for the geographical origin discrimination of green tea in China: a case study of Xihu Longjing. *Journal of food* composition *and analysis*, *67*,104-109. https://doi.org/10.1016/j.jfca.2018.01.005

Oliveri, P., & Downey, G. (2012). Multivariate class modelling for the verification of food-authenticity claims. *Trends in Analytical Chemistry*, *35*,74-86. https://doi.org/10.1016/j.trac.2012.02.005

Olivieri, A.C., Faber, N.M., Ferré, J., Boqué, R., Kalivas, J.H., & Mark, H. (2006). Uncertainty estimation and figures of merits: for multiway calibration. *Pure and Applied Chemistry, 78*, 633–661. https://doi.org/10.1016/B978-0-444-63527-3.00013-8

Olivieri, A.C. (2014). Analytical figures of merits: from univariate to multiway calibration. *Chemical Reviews,114,* 5358–5378. https://doi.org/10.1021/cr400455s

Olivieri, AC. (2012). Recent advances in analytical calibration with multi-way data. *Analytical Methods, 4*,1876-1886. https://doi.org/10.1039/C2AY25064K

Ordukaya, E.,& Karlik, B. (2017). Quality control of olive oils using machine learning and electronic nose. *Journal of Food Quality*, *ID 9272404*, 1-7. https://doi.org/10.1155/2017/9272404

Ortiz, M.C, & Sarabia, L. (2007). Quantitative determination in chromatographic analysis based on n-way calibration strategies. *Journal of Chromatography A, 1158*, 94-110. https://doi.org/10.1016/j.chroma.2007.04.047

Papadopoulou, O.S., Panagou, E.Z., Mohareb, F.R., & Nychas, G.J.E. (2013). Sensory and microbiological quality assessment of beef fillets using a portable electronic nose in tandem with support vector machine analysis. *Food Research International*, 50, 241-249. https://doi.org/10.1016/j.foodres.2012.10.020

Parastar, H., & Tauler, R. (2018). Big (bio)chemical data mining using chemometric methods: a need for chemists. *Angewante Chemistry International Edition*, http://dx.doi.org/10.1002/anie.201801134 (in press). https://doi.org/10.1002/anie.201801134

Pérez-Caballero, G., Andrade, J.M., Olmos, P., Molina, Y., Jiménez, I., Durán, J.J., Fernádez-Lozano, C., & Miguel-Cruz. (2017). Authentication of tequilas using pattern recognition and supervised classification. *Trends in Analytical Chemistry, 94,* 117-129. http://dx.doi.org/10.1016/j.trac.2017.07.008

Pisano, P.L., Silva, M.F., & Olivieri, A.C. (2015). Anthocyanins as markers for the classification of Argentinean wines according to botanical and geographical origin. Chemometric modeling of liquid chromatography-mass spectrometry data. *Food Chemistry, 175*, 174-180. https://doi.org/10.1016/j.foodchem.2014.11.124

Popek, S., Halagarda, M., & Jursa, K. (2017). A new model to identify botanical origin of Polish honeys based on the physicochemical parameters and chemometric analysis. *LWT - Food Science and Technology, 77*, 482-487. https://doi.org/10.1016/j.lwt.2016.12.003

Popescu, R., Costinel, D., Dinca, O.R., Marinescu, A., Stefanescu, I., & Ionete, R.E. (2015). Discrimination of vegetable oils using NMR spectroscopy and chemometrics. *Food Control, 48*,84-90. https://doi.org/10.1016/j.foodcont.2014.04.046

Qiu, S., & Wang, J. (2017). The prediction of food additives in the fruit juice based on electronic nose with chemometrics. *Food Chemistry, 230*, 208-214. https://doi.org/10.1016/j.foodchem.2017.03.011

Ramírez-Morales, I., Rivero, D., Fernández-Blanco, E., & Pazos, A. (2016). Optimization of NIR calibration models for multiple processes in the sugar industry. *Chemometrics and Intelligent Laboratory Systems*, *159*,45-57. https://doi.org/10.1016/j.chemolab.2016.10.003

Ríos-Reina, R., Elcoroaristizabal, S., Ocaña-González, J.A., García-González, D., Amigo, J.M., & Callejón, R.M. (2017). Characterization and authentication of Spanish PDO wine vinegars using multidimensional fluorescence and chemometrics. *Food Chemistry, 230,* 108-116. https://doi.org/10.1016/j.foodchem.2017.02.118

Rodionova, O.Y., Titova, A.V., & Pomerantesev, A.L. (2016). Discriminant analysis in an innaproriate method of authentication. *Trends in Analytical Chemistry*, *78*, 17-22. https://doi.org/10.1016/j.trac.2016.01.010

Rokach, L. (2010). Ensemble-based classifiers. *Artificial Intelligence Review, 33*, 1-39. https://doi.org/10.1007/s10462-009-9124-7

Ropodi, A.I., Panagou, E.Z., & Nychas, G.J.E. (2016). Data mining derived from food analyses using non-invasive/non destructive analytical techniques; determination of food authenticity, quality & safety in tandem with computer science techniques. *Trends in Food Science* & *Technology*, *50*,11-25. https://doi.org/10.1016/j.tifs.2016.01.011

Ruiz-Samblás, C., Cadenas, J.M., Pelta, D.A., & Cuadros-Rodríguez, L. (2014). Application of data mining methods for classification and prediction of olive oil blends with other vegetable oils. *Analytical and Bionalytical Chemistry, 406,* 2591-2601. https://doi.org/10.1007/s00216-014-7677-z

Sayago, A., González Dominguez, R., Beltran, R., & Fernández Recamales, A. (2018). Combination of complementary data mining methods for geographical characterization of extra virgin olive oils based on mineral composition. *Food Chemistry, 261*,42-50. https://doi.org/10.1016/j.foodchem.2018.04.019

Smith, B.R., Baker, M.J., & Palmer, D.S. (2018). PRFFECT: a versatile tool for spectroscopists. *Chemometrics and* Intelligent *Laboratory Systems*, *172*, 33-42. https://doi.org/10.1016/j.chemolab.2017.10.024

Spink, J., Hegarty, P.V., Fortin, N.D., Elliot, C.T., & Moyer, D.C. (2019). The application of public policy theory to the emerging food fraud risk: next steps. *Trends in Food Science & Technology, 85,* 116.128. https://doi.org/10.1016/j.tifs.2019.01.002

Springer, A.E. (2019). Wine authentication: a fingerprinting multiclass strategy to classify red varietals through profund chemometric analysis of volatiles. *European Food Research and Technology, 245,* 179-190. https://doi.org/10.1007/s00217-018-3151-z

Stanimirova, I., Üstün, B., Cajka, T., Riddelova, K., Hajslova, J., Buydens, L.M.C., & Walczak, B. (2010). Tracing the geographical origin of honeys based on volatile compounds profiles assessment using pattern recognition techniques. *Food Chemistry, 118*, 171-176. https://doi.org/10.1016/j.foodchem.2009.04.079

Steinbach, M.,& Tan, P.N. (2009). kNN: k-Nearest neighbors in X. Wu, V. Kumar (Eds.), *The Top Ten Algorithms in Data Mining* (pp. 151-161). Boca Raton: Chapman & Hall / CRC Press.

Sun, X., Lin, W., L,i X., Shen, Q., & Luo, H. (2015). Detection and quantification of extra virgin olive oil adulteration with edible oils by FT-IR spectroscopy and chemometrics. *Analytical Methods, 7*, 3939–3945. https://doi.org/10.1039/C5AY00472A

Szymanska, E., Gerretzen, J., Engel, J., Geurts, B., Blanchet, L., & Buydens, L.M.C. (2015). Chemometrics and qualitative analysis have a vibrant relationship. *Trends in Analytical Chemistry*, *69*, 34–51. https://doi.org/10.1016/j.trac.2015.02.015

Szymanska, E. (2018). Modern data science for analytical chemical data. *Analytica Chimica Acta*, *1028*, 1-10. https://doi.org/10.1016/j.aca.2018.05.038

Teye, E., Huang, X., Han, F., & Botchway, F. (2014). Discrimination of cocoa beans according to geographical origin by electronic tongue and multivariate algorithms. *Food Analytical Methods, 7*, 360-365. https://doi.org/10.1007/s12161-013-9634-4

Teye, E.,& Huang, X. (2015). Novel prediction of total fat content in cocoa beans by FT-NIR spectroscopy based on effective spectral selection multivariate regression. *Food Analytical Methods, 8*, 945-953. https://doi.org/10.1007/s12161-014-9933-4

Van der Veer, G., Van Ruth, S.M., & Akkermans, W. (2011). Guidelines for validation of chemometric models for food authentication. *RIKLT Report 2011.22*.

Vigneau, E., Coureoux, P., Symoneaux, R., Guérin, L., & Villière, A. (2018). Random forest: a machine learning methodology to highlight the volatile organic compounds involved in olfactory perception. *Food Quality and Preference*, *68*, 135–145. https://doi.org/10.1016/j.foodqual.2018.02.008

Vitale, R., Marini, F.,& Ruckebush, C. (2018). SIMCA modelling for overlapping classes: fixed or optimised decision threshold. *Analytica Chemistry,90*, 10738-10747. https://doi.org/10.1021/acs.analchem.8b01270

Wang, X., Huang, J., Fan, W., & Lu, H. (2015). Identification of green tea varieties and fast quantification of total polyphenols by near-infrared spectroscopy and ultraviolet-visible spectroscopy with chemometric algorithms. *Analytical Methods, 7,* 787-792. https://doi.org/10.1039/C4AY02106A

Wei, Z., Wang, J., & Wang, Y. (2010). Classification of monofloral honeys from different floral origins and geographical origins based on rheometer. *Journal of Food Engineering*, *96*, 469-479. https://doi.org/10.1016/j.jfoodeng.2009.08.028

Weng, S., Wang, F., Dong, R., Qiu, M., Shao, J., Huang, L., & Zhang, D. (2018). Fast and quantitative analysis of ediphenphos residue in rice using surface-enhanced raman spectroscopy. *Journal of Food Science*, *83*,1179-1185. https://doi.org/10.1111/1750-3841

Westad, F., & Marini, F. (2015). Validation of chemometric models – A tutorial. *Analytica Chimica Acta, 893*, 14–24. https://doi.org/10.1016/j.aca.2015.06.056

Witten, I.H., & Frank, E. (2005). *Data mining: practical machine learning tools and techniques*. (2nd ed.). San Francisco: Morgan Kaufmann Publishers.

Wood, J.E., Allaway, D., Boult, E.,& Scott, I.M. (2010). Operationally realistic validation for prediction of cocoa sensory qualities by high-throughput mass spectrometry. *Analytical Chemistry, 82*, 6048-6055. https://doi.org/10.1021/ac1006393

Wu, X., Kumar, V., Quinlan, J.R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A., Liu, B., Yu, P.S., Zhou, Z-H., Steinbach, M., Hand, D.J. & Steinberg. D. (2008) Top 10 algorithms in data mining. *Knowledge Information Systems, 14*, 1–37. https://doi.org/10.1007/s10115-007-0114-2

Xu, L., Ye, Z.H., Cui, H.F., Yu, X.P., Cai, C.B., Yang, H.W. (2012). Calibrating the shelf-life of chinese flavored dry tofu by FTIR spectroscopy and chemometrics: effects of data preprocessing and nonlinear transformation on multivariate calibration accuracy. *Food Analytical Methods, 5,* 1328-1334. https://doi.org/10.1007/s12161-012-9376-8

Xu, Y., Zomer, S., & Brereton, R. (2006). Support vector machine: a recent method for classification in chemometrics. *Critical Reviews in Analytical Chemistry*, *36*,177-188. https://doi.org/10.1080/10408340600969486

Yang, P.; Hwa-Yang, Y.; Zhou, B.B. & Zomaya, A.Y. (2010) A review of ensemble methods in bioinformatics. *Current Bioinformatics*, 5, 296-308. https://doi.org/ 10.2174/157489310794072508

Yu, J., Zhan, J., & Huang, W. (2017). Identification of wine according to grape variety using near-infrared spectroscopy based on radial basis function neural networks and least-squares support vector machines. *Food Analytical Methods, 10*, 3306-3311. https://doi.org/10.1007/s12161-017-0887-1

Yu, P., Low, M.Y., & Zhou, W. (2018). Design of experiments and regression modelling in food flavour and sensory analysis: a review. *Trends in Food Science & Technology*, *71*, 202–215. https://doi.org/10.1016/j.tifs.2017.11.013

Zhang, L., Li, P., Sun, X., Huang, J.H., Wang, X., Xu, B., Wang, X., & Ma, F., Zhang, Q., & Ding X. (2014).Classification and adulteration detection of vegetable oils based on fatty acid profiles. *Journal of Agricultural and Food Chemistry*, *62*, 8745-8751. https://doi.org/10.1021/jf501097c

Zheng, L., Watson, D.G., Johnston, B.F., Clark, R.L., Edrada-Ebel, R.L., & ELseheri, W. (2009). A chemometric study of chromatograms of tea extracts by correlation optimization warping in conjunction with PCA, support vector machines and random forest data modeling. *Analytica Chimica Acta, 642*,257-265. https://doi.org/10.1016/j.aca.2008

Zheng, W., Fu, X., & Ying, Y. (2014). Spectroscopy-based food classification with extreme learning machine. *Chemometrics and* Intelligent *Laboratory Systems*,*139*, 42-47. https://doi.org/10.1016/j.chemolab.2014.09.015

Zielinski, A.A.F., Haminiuk, C.W.I., Nunes, C.A., Schnitzler, E., van Ruth, S.M., Granato, D. (2014). Chemical composition, sensory properties, provenance, and bioactivity of fruit juices as assessed by chemometrics: a critical review and guideline. *Comprehensive Reviews in Food Science and Food* Safety, *13*, 300-316. https://doi.org/10.1111/1541-4337.12060

**Table 1.** Advantages and disadvantages of some data mining/chemometric methods

| Method | Advantages | Disadvantages |
|---|---|---|
| PCA | This displays quickly and easily similarities and differences between the samples and the variables relationships. | It does not allow classifying and assigning a class to each sample. |
| kNN | User-friendly method of applying. | If there are more samples of one class than other class (skewed distribution of classes), it could cause a wrong classification of the samples since the dominant class controls the classification. |
| SIMCA | It is able to develop a binary classification model training only with the target-class since it defines an acceptance region that contains all the objects/samples of the target class. | In models trained with two classes or more might give rise to overlapping between acceptance regions which contain the samples of the different classes. Thus, some samples might be classified in one or more classes. |
| PLS-DA | The classification model is built quickly and easily, and the results use to be very successful. | If the separation between the regions of the different classes are not sufficiently evident could give rise to classification errors. |
| SVM | This can circumvent the technical difficulty when the separation between the regions of the different classes of the samples are not sufficiently evident. | For non-lineal SVM models, alternative kernel functions must be used. Thus, the development of the model is difficult, and a lot of Informatics resources are necessary. |
| DT / CART | This is not affected by outliers or non-linear relationships. The models are presented in a simplified and are easy-to-interpret way. | It performs poorly when training set is small in comparison with the number of classes, especially for continuous data. |
| Boosted DT | This achieves more accurate classification by decreasing bias. | The model could be overfitted, therefore might fail to fit new samples and predict them incorrectly. |
| Bagged DT / RF | This is very suitable for unstable models or for class imbalance problems since the variance is decreased. The | Understanding results is complex since the classification is not displayed as a graphical tree. |

| | | |
|---|---|---|
| | risk of overfitting is minimised. | |

**Acronyms:** PCA (principal component analysis), SIMCA (soft independent modelling by class analogy), kNN (k-nearest neighbours), PLS-DA (partial least squares – discriminant analysis), SVM (support vector machine), DT (decision trees), CART (classification and regression tree), RF (random forest).

**Table 2.** Pattern recognition methods implemented in some of the more used software in multivariate data analysis for food quality and authenticity data

| Software | Exploratory analysis | | Class-modelling | Discriminant analysis | | | Decision rules | | | Variable reduction | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | HCA | PCA | SIMCA | kNN | SVM | PLS-DA | ANN | CART | RF | MCR-ALS | PARAFAC |
| PLS_Toolbox | • | • | • | • | • | • | • | | | • | • |
| SOLO | • | • | • | • | • | • | • | | | • | • |
| SIMCA | • | • | • | | | • | | | | | |
| Unscrambler | • | • | | | • | • | | | | | |
| Pirouette | • | • | • | • | | | | | | | |
| PerClass | • | • | • | • | • | • | • | • | • | | |

**Acronyms:** PCA (principal component analysis), HCA (hierarchy cluster analysis), SIMCA (soft independent modelling by class analogy), kNN (k-nearest neighbours), PLS-DA (partial least squares – discriminant analysis), SVM (support vector machine), ANN (artificial neural network), CART (classification and regression tree), RF (random forest), PARAFAC (parallel factor analysis), MCR-ALS (multivariate curve resolution – alternating least squares).

**Table3.** Compilation of papers on food chemistry in which SVM, DT, CART, and RF have been used

| **Food** | **Purpose** | **Analytical technique** | **Tool** | **Ref.** |
|---|---|---|---|---|
| Apple | Determination of pesticides | Raman spectrometry | PLS, **SVM** | [Fan, Lai, Rasco, & Huang, 2015] |
| Beef | Evaluation of sensory quality | Electronic nose | PCA, DFA, **SVM** | [Papadopolou, Panagou, Mohareb, & Nychas, 2013] |
| | Evaluation of sensory quality | GC-MS | LDA, SIMCA, PLS-DA, **SVM** | [Arredondo et al., 2014] |
| | Evaluation of sensory quality | Electronic nose | **SVM** | [Mohareb, Papadopolou, Panagou, & Nychas, 2016] |
| Beer | Selection of the optimal number of parameters describing beer qualities | *Several chemical parameters* | **DT** | [Debska & Guzowska-Swider, 2011] |
| Cheese | PDO authenticity | SPME-MS | PCA, LDA, SIMCA, **SVM** | [Majcher, Kaczmarek, Pawlik, Pikul, & Jelén, 2015] |
| | Quality control | PTR-TOF-MS | **RF**, **SVM**, PDA, DPLS | [Fabris, et al., 2010] |
| Cocoa | Evaluation of sensory quality | Sensory tasting | PLS, **SVM**, MLR | [Wood, Allaway, Boult, & Scott, 2010] |
| | Quantification of the total fat content | FT-NIR spectrometry | PLS, **SVM** | [Teye & Huang, 2015] |
| | Classification according to their geographical origin | Electronic tongue | FDA, PCA, kNN, **SVM** | [Teye, Huang, Han, & Botchway, 2014] |
| Coffee | Evaluation of authenticity according to | ICP-MS | MLP, **SVM**, NB | [Barbosa et al., 2014] |

| | | | | |
|---|---|---|---|---|
| | the trace element | | | |
| | Classification according to geographical origin | NIR and FTIR spectrometry | **SVM** | [Bona et al., 2017] |
| Fruit | Classification of persimmons according to geographical origin | FT-NIR spectrometry | HCA, PCA, **SVM** | [Khanmohammadi et al., 2014] |
| | Classification of jujube fruit according to geographical origin | NIR spectrometry | PCA, LDA, **SVM**, ANN | [Guo, Ni, & Kokot, 2016]] |
| Grape | Quality control | Imaging spectrometry | **SVM** | [Liu, & Whitty, 2015] |
| | Detection of adulteration | Dielectric sensor | PCA, HCA, LDA, **SVM** | [Naderi-Boldaji et al., 2018] |
| Ginseng | Classification according to geographical origin | FT-MIR, NIR | **RF** | [Li, Zhang, & Wang, 2018] |
| Honey | Classification according to floral and geographical origin | Rheometry | PCA, PLS, PCR, **SVM** | [Wei, Wang, & Wang, 2010] |
| | Classification according to geographical origin | GC x GC-TOF | SIMCA, DPLS, LDA, **SVM** | [Stanimora et al., 2010] |
| | Classification according to geographical origin | Electronic tongue | PCA, HCA, PLS, **SVM** | [El Alami et al., 2018] |
| | Classification according to phenolic composition | HPLC-UV | PLS-DA, **SVM** | [Kemal, de B Harrington, Sahin, Demir, & Gunes, 2017] |
| | Classification according to botanical origin | viscosimetry, UV-Vis spectrometry, HPLC-IR, HPLC-UV | **CART** | [Popek, Halagarda, & Jursa, 2017] |
| | Classification according to botanical and geographical origin | ICP-MS | LDA, **CART** | [Chudzinska & Baralkiewicz, 2011] |

| | | | | |
|---|---|---|---|---|
| Juices | Detection of adulteration of tomato juices | Electronic nose and tongue | CDA, **SVM**, PCR | [Hong & Wang, 2014] |
| | Discrimination of organic grape juice from conventional grape juice | ICP-MS | **SVM**, **CART,** MPL | [Maione et al., 2016a] |
| | Detection of additives | Electronic nose | PLS, **SVM**, **RF** | [Qiu & Wang, 2017] |
| Licors | Quality control | HS-SPME-MS | PLS, **SVM** | [Cheng, Fan & Yan, 2013] |
| Olive oils | Classification according to geographical origin | NIR and MIR spectrometry | **SVM** | [Devos, Downey, & Duponchel, 2014] |
| | Detection and quantification of adulteration of extra virgin olive oil with other edible vegetable oils | Raman spectrometry | PLS, **SVM** | [Dong, Zhang, Zhang, & Wang, 2012] |
| | Classification of oil blends according to the vegetable oil used for blending and prediction of the proportion of olive oil used in each blend | GC-MS | **CART**, **RF** | [Ruiz-Samblás, Cadenas, Pelta, & Cuadros-Rodríguez, 2014] |
| | Discrimination of olive oil from other edible vegetable oils | HPLC-CAD | PCA, kNN, PLS-DA, OCPLS, SIMCA, **SVM** | [Jiménez-Carvelo, Pérez-Castaño, González-Casado, & Cuadros-Rodríguez, 2017a] |
| | Discrimination of olive oil from other edible vegetable oils | HPLC-CAD | PCA, PLS-DA, OCPLS, kNN, SIMCA, **SVM** | [Jiménez-Carvelo, González-Casado, Pérez-Castaño, & Cuadros-Rodríguez, 2017b] |
| | Discrimination of olive oil from other edible vegetable oils and quantification of the proportion of olive oil in blends with other vegetable oils. | FTIR and Raman spectrometry | PCA, PLS-DA, OCPLS, kNN, SIMCA, **SVM** | [Jiménez-Carvelo, Osorio, Koidis, González-Casado, Cuadros-Rodríguez, 2017c] |

| | | | | |
|---|---|---|---|---|
| | Quantification of olive oils in blends with other vegetable oils | HPLC-CAD | PLS, **SVM** | [Jiménez-Carvelo, González-Casado, & Cuadros-Rodríguez, 2017d] |
| | Classification according to geographical origin | HPLC-IR, GC-FID | PCA, **DT** | [Nasibov, Kantarci, Vahaplar, & Kinay, 2016] |
| | Classification according to geographical origin | ICP-MS | PLS-DA, **SVM**, **RF** | [Sayago, González-Domínguez, Beltrán, & Fernández-Recamales, 2018] |
| | Identification and classification of Turkish olive oils according to geographical origin | Electronic Nose | NB, kNN, LDA, ANN, **SVM** | [Ordukaya & Karlik, 2017] |
| | Quality control | Synchronous spectrofluorimetry | LDA, QDA, RDA, kNN, **SVM, RF** | [Dankowska & Kowalewski, 2018] |
| | Classification of extra virgin olive oils according to their cultivars | HPLC-DAD | PCA, MCR, PLS-DA, NPLS-DA, **RF** | [Jiménez-Carvelo, Cruz, Olivieri, González-Casado, & Cuadros-Rodríguez, 2019] |
| Pepper | Determination of pesticides | Raman spectrometry | **SVM** | [Li, Sun, Pu & Jayas, 2017] |
| Rice | Discrimination of organic rice from conventional rice | ICP-MS | **SVM** | [Barbosa, et al., 2016] |
| | Classification according to geographical origin | ICP-MS | **SVM**, **RF**, ANN | [Maione, Lemos Batista, Campiglia, Barbosa, & Barbosa, 2016b] |
| | Detection of adulterations | MS | **SVM**, **RF**, kNN | [Kyu et al., 2017] |
| | Classification according to geographical origin | Raman spectrometry | kNN, SIMCA, PLS-DA, **SVM** | [Feng, Zhang, Cong, & Zhu, 2013] |

| | | | | |
|---|---|---|---|---|
| | Quality control | GC-MS | **RF** | [Mahdavi, Farimani, Fathi, & Chassempour, 2015] |
| | Quantification of Ediphenphos | Raman spectrometry | PCA, PLS, **RF** | [Weng et al., 2018] |
| Sugar | Quality control | NIR spectrometry | **SVM** | [Ramírez-Morales, Rivero, Fernández-Blanco, & Pazos, 2016] |
| | Authenticity evaluation | ICP-MS | NB, **RF** | [Barbosa et al., 2015] |
| Tea | Discrimination of green and black tea | Voltammetry | PCA, **SVM** | [Liu et al., 2014] |
| | Classification between different kinds of tea | HPLC-UV | PCA, **SVM**, **RF** | [Zheng et al., 2009] |
| | Discrimination of five varieties of green tea and quantification of polyphenolic compounds | NIR and UV-Vis spectrometry | PCA, PLS, **RF** | [Wang, Huang, Fan, & Lu, 2015] |
| | Classification according to geographical origin | ICP-AES | LDA, PLS-DA, **DT** | [Ni, et al., 2018] |
| | Classification according to botanical and geographical origin | UV-Vis spectrometry | kNN, SIMCA, PLS-DA, PCA-LDA, **CART** | [Gonçalvez et al., 2016] |
| Tequila | Discrimination of tequila from traditionally processed mescal | (DIESI)LTP-MS | PCA, **RF** | [Martínez-Jarquín, Moreno-Pedraza, Cázarez-García, & Winkler, 2017] |
| | Differentiation of different kinds of tequila | UV-Vis spectrometry | kNN, SIMCA, PCA, PLS, **CART**, **RF**, **SVM** | [Pérez-Caballero et al., 2017] |
| | Differentiation of different kinds of tequila | UV-Vis spectrometry | QDA, PLS-DA, PLS-KERNEL, **SVM,** CPANN | [Andrade, Ballabio, Gómez-Carracedo, & Pérez-Caballero,, 2017] |

| | | | | |
|---|---|---|---|---|
| | Detection of adulterations | UV-Vis spectrometry | PCA, LDA, **SVM** | [Contreras, et al., 2010] |
| | Classification according to the geographical origin | ICP-AES | PCA, LDA, **SVM** | [Ceballos-Magaña, *et al.*, 2012] |
| Tofu | Study of its shelf-life | FTIR spectrometry | PLS, **SVM** | [Xu et al., 2012] |
| Vegetable oils | Quantification of fatty acid compounds | Raman spectrometry | **SVM** | [Dong, Zhang, Zhang, & Wang, 2013] |
| | Detection of adulteration | GC-MS | PCA, **RF** | [Zhang et al., 2014] |
| | Discrimination of vegetable oils according to their quality | GC-MS | PCA, HCA, **RF** | [Ai et al., 2014] |
| | Detection of adulteration | GCxGC-TOF | PCA, HCA, **RF** | [Hu et al., 2014] |
| Vinegar | Authenticity evaluation | Electronic tongue | **SVM, RF,** BPANN | [Liu, Wang, Wang, & Li, 2013] |
| | PDO authenticity | Spectrofluorimetry | PARAFAC, PLS-DA, **SVM** | [Ríos-Reina et al., 2017] |
| | Quality control | NIR spectrometry | **LS-SVM,** BPANN, PLS | [Ji-yong et al., 2013] |
| | Quality control | Vis/NIR spectrometry | PLS, **LS-SVM** | [Bao et al., 2014] |
| Wine | Authentication based on the grape variety | GC-MS | **DAG tree**, OPLS-DA, SIMCA | [Springer, 2019] |
| | Authenticity evaluation | FTIR spectrometry | PCA, HCA, LDA, **CART** | [Loannou-Papayianni, Kokkinfta, & Theocharis, 2011] |
| | Assurance of the authenticity according to the grape variety and different family compounds | GC-FID; GC-FPD; GC-MS | **SVM**, **RF**, MLP, kNN, NB | [Gómez-Meire, Campos, Falqué, Díaz & Fdez-Riverola, 2014] |
| | Classification according to the geographical origin | HPLC-DAD | **SVM** | [da Costa, Castro, & Barbosa, 2016] |

| | | | |
|---|---|---|---|
| Classification according to the geographical origin | HPLC-DAD; HPLC-DAD-MS | **SVM** | [Costa, García Llobodanin, Alves Castro, & Barbosa, 2018] |
| Classification according to geographical origin | UV/Vis/NIR spectrometry | LDA, SIMCA, **SVM** | [Martelo-Vidal, Dominguez-Agis, & Vázquez, 2013] |
| Classification of white wine from different brands according to their elemental profile | ICP-MS | **SVM** | [Jurado, Alcázar, Palacios-Morillo, & de Pablos, 2012] |
| Classification according to their grape variety | NIR spectrometry | RBFNN, **SVM** | [Yu, Zhan, & Huang, 2017] |
| Evaluation of sensory quality | GC-MS | **RF** | [Vigneau, Coureoux, Symoneaux, Guérin, & Villière, 2018] |
| PDO authenticity | HPLC-DAD | LDA, SIMCA, **SVM** | [Martelo-Vidal & Vázquez, 2016] |
| Quality Control | UV-Vis spectrometry | PCA, **SVM** | [Liu, Pan & Zhang, 2018] |

**Acronyms**: BPANN (back propagation artificial neural network), CAD (charged aerosol detector), CART (classification and regression tree), CDA (canonical discriminant analysis), CPANN (counter propagation artificial neural networks), DAD (diode array detector), DAG (directed acyclic graph), DIESI (direct-injection electrospray ionisation), DFA (discriminant function analysis), DPLS (discriminant partial least squares), DT (decision tree), FID (flame ionization detector), FDA (Fisher's discriminant analysis), FPD (photometric flame detection), FTIR-HATR (Fourier transform infrared spectroscopy - horizontal attenuated total reflectance), FT-NIR (Fourier transform-near infrared), GC (gas chromatography), HCA (hierarchical cluster analysis), HPLC (high performance liquid chromatography), HS (head space), ICP (inductively coupled plasma), IR (refractive Index), kNN (k-nearest neighbour), LDA (linear discriminant analysis), LTP ( low-temperature plasma), LS-SVM (least-squares support vector machine), MLP (multilayer perceptron), MS (mass spectrometry), NB (naïve Bayes), OCPLS (one class partial least squares), PARAFAC (parallel factor analysis), PCA (principal component analysis), PCR (principal component regression), PDA (penalised discriminant analysis), PLS (partial least squares), PLS-DA (partial least squares-discriminant analysis), PTR (proton-transfer-reaction), QDA (quadratic discriminant analysis), RDF (regularized discriminant analysis), RF (random forest), SIMCA (soft independent modelling by class analogy), SPME (solid-phase microextraction), SVM (support vector machine), TOF (time of flight), UV (ultra violet), Vis (visible).

**Figure caption**

**Figure 1**. General overview of conventional chemometric pattern recognition methods.

**Figure 2**. (a) Diagram showing the most–usual data mining chemometric methods used in food quality and authenticity; (b) Simple graphical description of how some of the most-usual data mining/chemometric methods carry out the classification.

**Figure 3**. Trends in publications in the area of food chemistry in which SVM, RF, DT, and CART have been used.

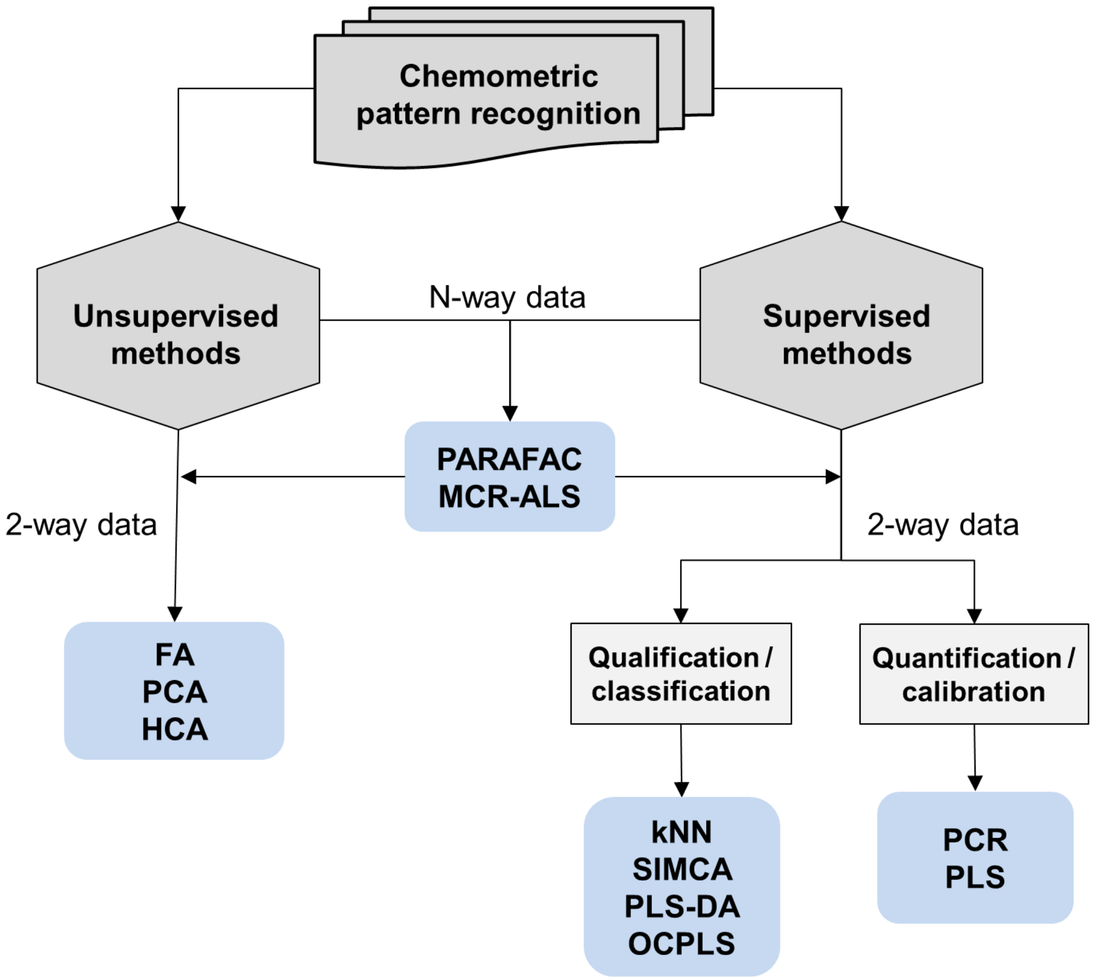
**Figure 4**. Use of the concepts of data mining and machine learning terms in the last 10 years.

**Figure 5**. Schematic illustration of the ensemble methods (D: data-set; sD: data-subset). The light cyan colour shows the decision taken by the classifier tree. In this example the target class is represented by a rectangle in red colour.

**Figure 6**. Graphical description of how RF performs the classification. A probability node (i.e., root node) is represented by a circle and shows the probability of certain results; a decision node is represented by a square and shows a decision to be made; finally, a terminal node shows the result of a decision path (see text for additional explanations on the operation of RF classification).

681 **<Figure 1>**

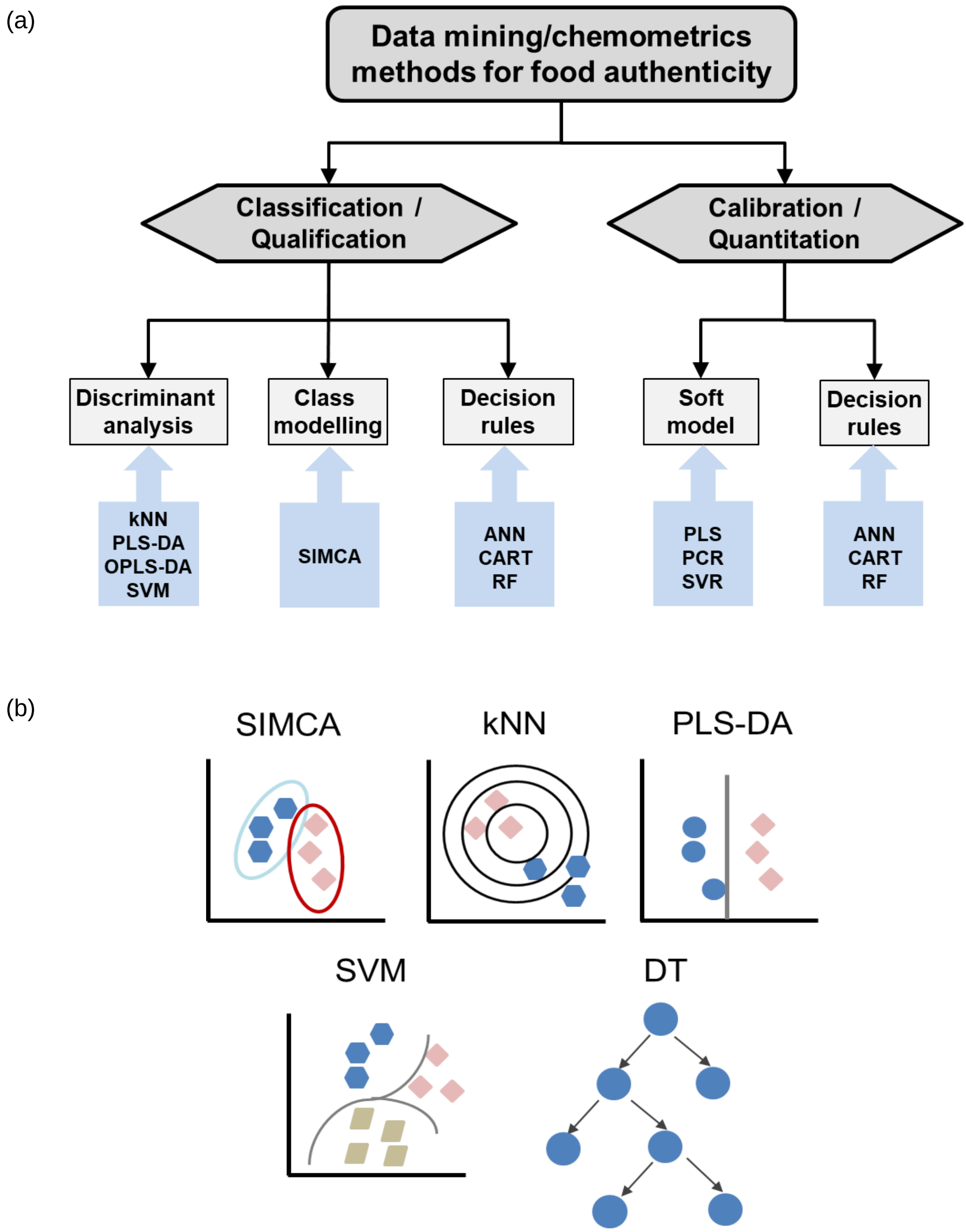
682



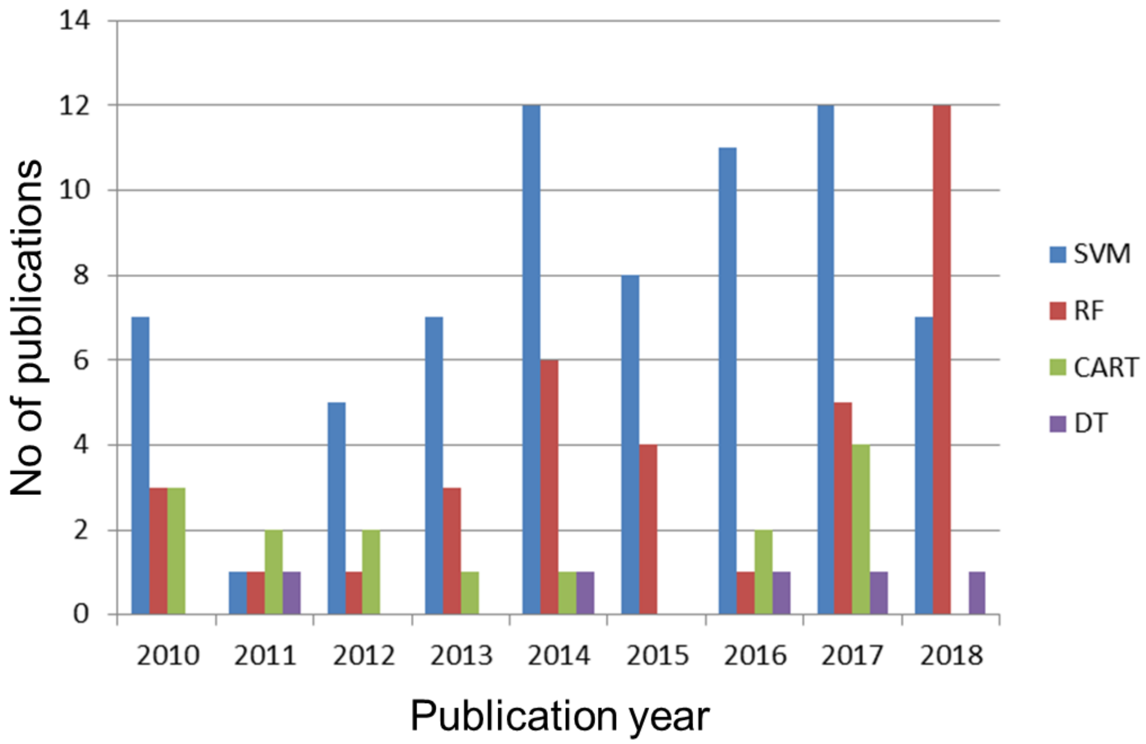

683

684

**<Figure 2>**

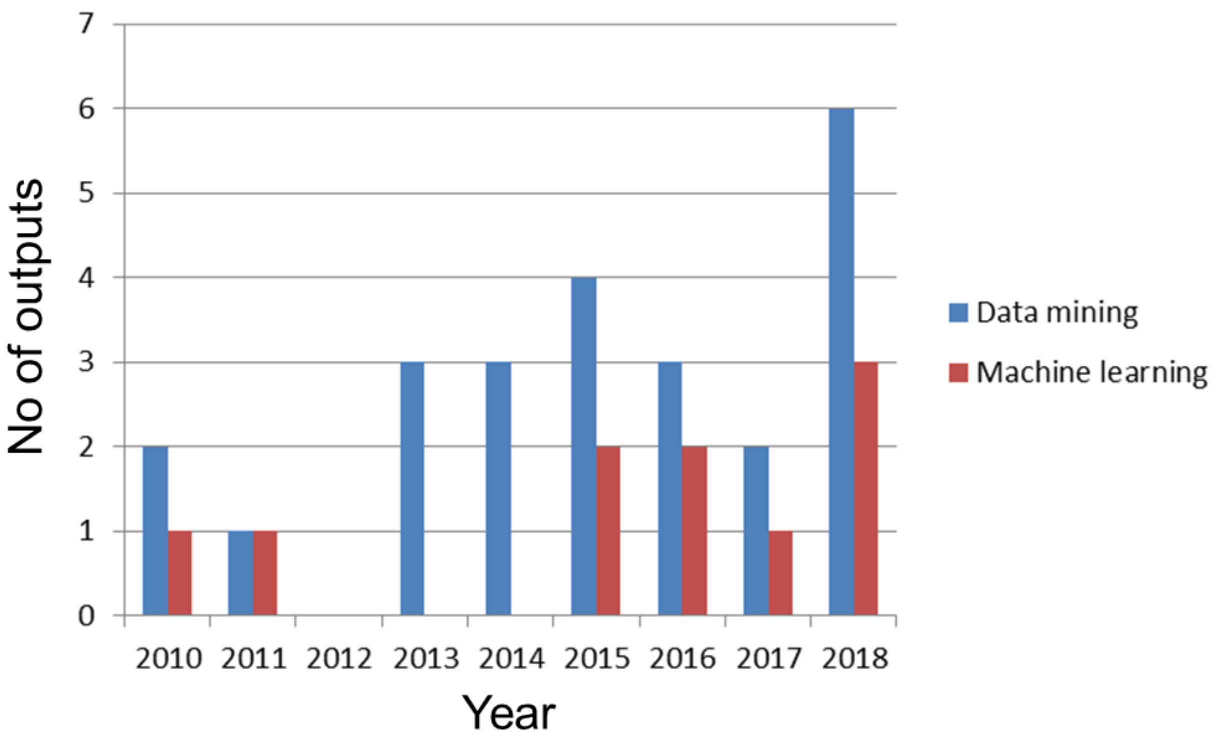

(a)
Data mining/chemometrics methods for food authenticity
Classification / Qualification
Calibration / Quantitation
Discriminant analysis
Class modelling
Decision rules
Soft model
Decision rules
kNN
PLS-DA
OPLS-DA
SVM
SIMCA
ANN
CART
RF
PLS
PCR
SVR
ANN
CART
RF
(b)
SIMCA
kNN
PLS-DA
SVM
DT

**<Figure 3>**


No of publications
14
12
10
8
6
4
2
0
2010
2011
2012
2013
2014
2015
2016
2017
2018
Publication year
SVM
RF
CART
DT

693 **<Figure 4>**

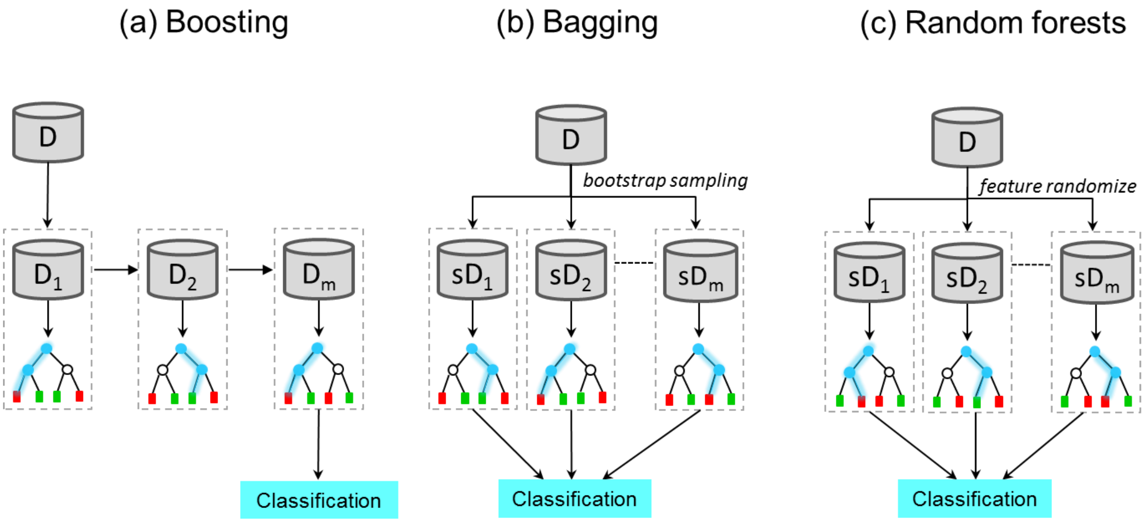
694



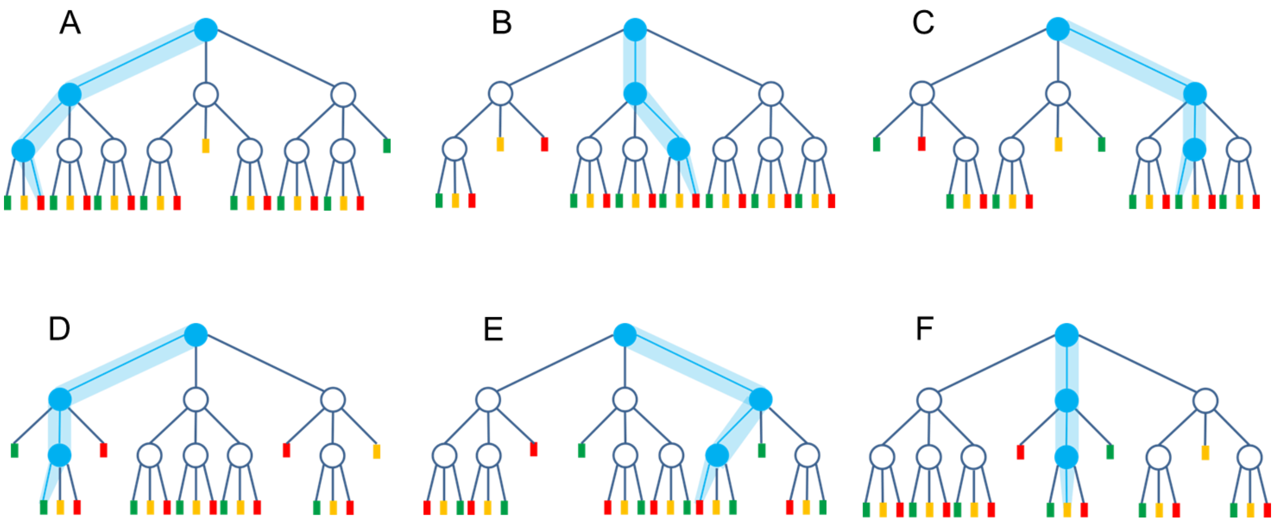

695

696

**<Figure 5>**


(a) Boosting
(b) Bagging
(c) Random forests
D
D$_1$
D$_2$
D$_m$
Classification
D
bootstrap sampling
sD$_1$
sD$_2$
sD$_m$
Classification
D
feature randomize
sD$_1$
sD$_2$
sD$_m$
Classification

**<Figure 6>**



704

**Ana M. JIMÉNEZ-CARVELO** is graduate chemist in 2013 from the University of Granada (Spain). She obtained her Master's Degree in Advances in Food Quality and Technology, one year later. Ph.D. in Analytical Chemistry in 2018 entitled 'Analytical study of the transesterified fraction of the olive oil. Application in problems of olive oil authentication '. During this period, she got two research stays at Institute for Global Food Security, Queen's University (Northern Ireland, UK) with the aim of improving her knowledge in new analytical techniques, and in the Department of Analytical Chemistry in University of Rosario (Argentina) with the goal of acquire experience in the use of advanced chemometric methods. Her research interest includes liquid chromatography (HPLC-CAD, HPLC-DAD and UHPLC-(Orbitrap)MS), spectroscopic techniques (NIR, ATR-FTIR, Raman) and chemometrics (classification and quantification multivariate techniques applied on first and second order data).

**Antonio GONZÁLEZ-CASADO**. Tenured Professor at the Department of Analytical Chemistry (University of Granada, Spain), expert in the field of Chemical Metrology and Qualimetrics (CMQ). He teaches analytical chemistry in Undergraduate Chemistry and Food Technology and Master's Degrees in Chemistry. His most significant R&D area of interest included the development of quality assurance protocol (calibration, validation, uncertainty estimation, etc.) on analytical process. His research fields also include the production of certified reference materials of olive oils for quality control. His working is currently focused on the analytical control for food quality and authenticity, particularly on vegetable (olive) oils using chemometrics tools (multivariate data analysis, MDA) from unspecific chromatographic data ("fingerprinting").

**Mª Gracia BAGUR-GONZÁLEZ**. Tenured Associate Professor at the Department of Analytical Chemistry (University of Granada, Spain), expert in the field of Chemical Metrology and Qualimetrics (CMQ). She teaches analytical chemistry in Undergraduate Chemistry and Environmental Sciences and Master's Degrees in Chemistry. Also coordinates the academic activity of the Master's Degree of Chemistry. Her most significant R&D areas of interest include the use of chemometrics tools (multivariate data analysis, MDA) for: (i) the evaluation of the environmental impact of abandoned metallic mining areas; and (ii) the analytical control aimed at food quality and authenticity, particularly on vegetable (olive) oils and fat spreads from unspecific chromatographic data ("fingerprinting").

**Luis CUADROS-RODRÍGUEZ**. Full Professor at the Department of Analytical Chemistry (University of Granada, Spain), expert in the field of Chemical Metrology and Qualimetrics (CMQ). He teaches analytical chemistry in Undergraduate and Master's Degrees in Chemistry and Chemical Engineering. His most significant R&D area of interest included the development of quality assurance protocol (calibration, validation, uncertainty estimation, etc.) on analytical process. He has also developed the use of multivariate process optimization by applying statistically designed experiments on analytical methods. His working is currently focused on the analytical control for food quality and authenticity, particularly on vegetable (olive) oils using chemometrics tools (multivariate data analysis, MDA) from unspecific chromatographic data ("fingerprinting").

705