TrAC Trends in Analytical Chemistry. 2017; 94: 117-129

Authentication of tequilas using pattern recognition and supervised classification

G. Pérez-Caballero[a], J.M. Andrade[b], P. Olmos[a], Y. Molina[a], I. Jiménez[a], J.J. Durán[a], C. Fernandez-Lozano[c], F. Miguel-Cruz[d]

[a] *Unidad de Investigación Multidisciplinaria (UIM), Facultad de Estudios Superiores Cuautitlán, Universidad Nacional Autónoma de México, Km 2.5 Carretera Cuautitlán–Teoloyucan, San Sebastián Xhala, Cuautitlán Izcalli, Estado de México, CP 54714, Mexico*
[b] *QANAP Research Group, Analytical Chemistry, University of A Coruña, Campus da Zapateira s/n, E-15071, A Coruña, Spain*
[c] *Information and Communication Technologies Department, Faculty of Computer Science, University of A Coruña, A Coruña, 15071, Spain*
[d] *Consejo Regulador Del Tequila, Av. Patria No. 723, Jardines de Guadalupe, 45030 Zapopan, Jalisco, Mexico*

Abstract
Sales of reputed, Mexican tequila grown substantially in last years and, therefore, counterfeiting is increasing steadily. Hence, methodologies intended to characterize and authenticate commercial beverages are a real need. They require a combination of analytical characterization and chemometric tools. This work reports concisely on the former and focus on the chemometric tools employed so far in connection with them. Further, a practical case study presents the classification capabilities of nine supervised classification methods to differentiate white, rested, aged and extra-aged tequilas. The largest set of certified tequilas employed so far was considered. In general, non linear methods performed best than linear ones (accuracy higher than 94% in both training and validation). The case study demonstrates that it is possible to develop fast, cheap, easy to implement and reliable analytical methodologies to authenticate and classify samples of tequilas.

Keywords
Tequila; Supervised classification; Authentication; Dimensionality reduction; Machine learning

1. Introduction

Tequila is a Mexican alcoholic spirit worldwide renowned whose popularity rocketed in recent years both nationally and, mainly, internationally. Indeed tequila has become the fourth most consumed spirit in the world after whiskey, vodka and rum. To control its production and overall quality, the Mexican government set rules which, roughly speaking, were deployed by means of the so-called *Protected Designation of Origin* [1] which is recognized in major markets, like USA (North America Free Trade Agreement) and European Union [2], [3]. Tequila is elaborated from agave *Tequilana Weber*, blue variety, which is cultivated within specific, protected regions of Mexico that constitute the geographical Denomination of Origin Tequila – DOT, located in the States of Jalisco, Nayarit, Tamaulipas, Michoacán and Guanajuato.

The *Tequila Regulatory Council* (in brief, CRT, *Consejo Regulador del Tequila*) coordinates several organizations involved in the production and marketing of tequila. It is the only worldwide organization accredited to certify compliance with the Mexican standard for tequila (NOM-006-SCFI-2012) [1]. It also takes care of the overall quality of tequila through research and specialized studies. It is worth noting that the different types of tequila are not based on chemical or compositional characteristics, but on the verification/certification of the type of agave and resting time in oak casks, which are controlled by authorized inspectors.

The strong demand for tequila caused that for several years its production stepped from a traditional handcraft scale to an industrial one, which gave rise to some commercial problems. The high demand opened opportunities to adulteration practices as there may be economical benefits associated to counterfeit tequila. This urged authorities and interested parties to implement monitoring and authentication studies. This task is a real analytical challenge due to the sophistication of some adulterations and the decrease in the quality of many authentic tequilas after the introduction of low-cost and low-quality components. Sometimes, the problem is improper labeling (for example, indicating that a tequila is Aged, rather than Rested). Administrative actions were enforced, like the mandatory use of an official seal from the CRT stamped onto the bottle, but this is not enough to discourage some counterfeiters.

In Mexico alone, the tequila industry generates more than 70,000 jobs. In 2015, approximately 229 ML (millions of liters) were produced from which about 183 ML were exported to over 120 countries worldwide [4]. Adulteration and/or counterfeiting damage significantly the economy of companies that produce good quality tequila, affect their credibility and, worst, the satisfaction of the customer and –in extreme cases- threaten consumer's health.

Two major categories of tequila can be recognized: ‘100% agave’ and ‘mixed’. The former is obtained collecting the pinecones or stems (‘piñas’) of agave *Tequilana Weber* blue (7–9 years old), removing the leaves, cutting and cooking the stems, milling and pressing them to get the juice and submitting it to alcoholic fermentation with special yeasts, yielding 100% agave tequila. The second category contains sugars which do not proceed from agave, added before fermentation, although in a proportion not greater than 49%. This results in the so-called ‘mixed tequila’ or, simply, ‘tequila’, as it is sold worldwide (in general, ‘Tequila 100% agave’ is not easily available outside Mexico). To sum up, for a spirit be considered ‘Tequila 100% agave’ it must proceed only from blue agave and be bottled by authorized producers in the production plants located within the DOT geographical protected area.

Different types of tequila can be recognized according to their wooden aging time currently in French oak or white oak barrels (*Quercus ilex* and *Quercus alba* are recommended). ’*Silver*’ or ‘*white*’ tequila (‘Tequila Blanco’) is transparent, without aging. ‘*Rested*’ tequila (‘Tequila Reposado’) was aged between 2 months (at least) and 12 months. ‘*Aged*’ tequila (‘Tequila Añejo’) corresponds to tequila rested in wooden barrels for a minimum of 12 months (up to 24 months). Premium ‘*Extra-aged*’ tequila (‘Tequila Extra-añejo’) appears when aging lasts between 18 months and 4 years. Rested, Aged and Extra-aged tequilas exhibit a variety of colors, from amber

to toasted-coffee. Finally, '*Young*' or '*gold*' tequila ('Tequila Joven') results from blends of white tequila with other aged tequilas, it has a gold color and it has usually been added authorized substances to smooth its flavor and taste. Without doubt, the price of each type increases sharply with aging. For instance, silver tequila costs around 13 euros/bottle whereas aged tequila (100% agave) can be anything from 145 euros/bottle (premium brands).

In the following paragraphs a resume of the literature directly related to the characterization and comparison of different types of tequila is presented. Please, note that papers dealing only with technological, biotechnological or engineering aspects of tequila production are out of the scope of this work and they were discarded. The bibliographical source to perform the search was the ISI Web of Science (accessed in November 2016). The key word used to search throughout the titles was 'tequila'. A first refining criterion was to select papers only in the science technology, food science technology and chemistry fields. This yielded 108 references that were reviewed further to evaluate which contained relevant applications of analytical chemistry methodologies. This led to a final selection of 41 papers since 1996. Despite this number is not too high, there has been a continuous interest for studying tequila in last years, likely because of a general interest of citizens by the quality of the foods we consume. This fact, along with rising counterfeiting activities worldwide, suggests that the number of papers will continue to increase.

Fig. 1a shows the continuous research this issue underwent and how the number of related works has increased steadily. Fig. 1b indicates that the number of cites has also increased steadily; the 41 works selected for this study received 565 citations in 375 papers. A very first seminal paper from Benn and Peppard has been referenced 91 times since 1996.

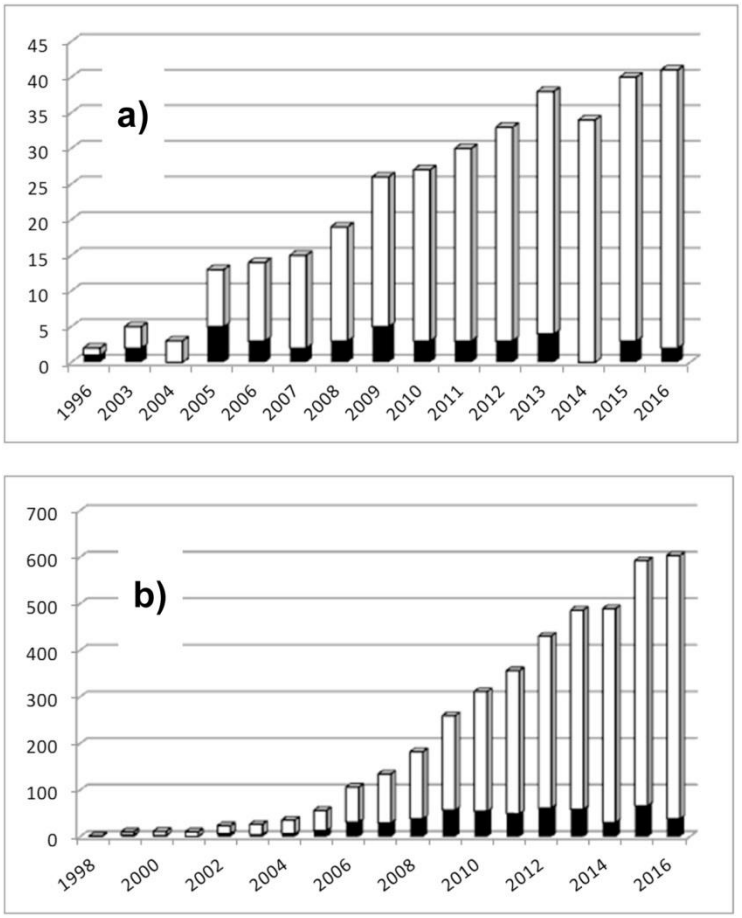

Fig. 1. Bibliographical search with 'tequila' as key word in the title (see text for more details on the search). Number of papers published per year –black- and accumulated figures –white- (a); number of citations per year –black- and accumulated values –white- (b).

Finally, it is worth noting that to the best of the authors' knowledge, this is the first review on the analytical characterization of tequila and its authentication. The review is organized according to the two major stages involved in an analytical strategy to address authentication: the analytical characterization itself and the chemometric tools used to treat the data. The second issue constitutes the main focus of this work.

1.1. Analytical characterization

With regards to the chemical characterization, tequila is a rich, complex mixture of compounds. Many of them are volatile (higher alcohols, aldehydes, fatty acids, esters, sulfur compounds, etc.), including small phenolic compounds as studied by liquid chromatography (LC) with ion-trap mass spectrometry (MS) [5]. They are responsible for its aroma and flavor and, so, they were studied by gas chromatography (GC) with flame ionization detection (FID) and MS detection [6], [7], [8], [9], or sulfur chemiluminiscence [6] detection (incidentally, this was the 1996 seminal paper on tequila authentication referred to above).

As for any alcoholic beverage, the fermentation process is really critical to get olfactive and flavor sensations, however the process itself is very complex. Of course, control of sugar(s) fermentation is a key issue. For instance, fermentation can convert agave fructans into sugars (using commercial inulinases (inuline is a rough denomination for the agave sugars) and thermal acid hydrolysis), as it was studied for blue agave employing high performance liquid chromatography (HPLC) [10] and GC–MS, after a previous Maillard reaction [11]. However, tequila also needs a distillation process, which contributes to the final aromas; thus analytical control of this stage is important and it was accomplished using HPLC, an interesting result being that despite a short heat treatment is essential for the development of the organoleptic characteristics of tequila, the hydrolysis can be carried out just with enzymes (avoiding major changes in its taste or aroma) [12]. HPLC was also proposed in combination with evaporative light scattering detection to determine carbohydrates in tequila [13].

A comparative study of volatile compounds and sensory profiles by HPLC and GC–FID (as well as classical methods to measure reducing sugars and ethanol) showed that they were the result of fermentation of blue agave juice by different yeasts [14], [15], [16]. Maturation has also an obvious impact on the odor active compounds which determine the final aroma as it was studied recently by GC–MS [17]. Several markers to evaluate aging in oak barrels during maturation were proposed [18]. Different solvents and extraction methods in combination with GC–FID and GC–MS were also studied to determine volatiles [19].

Besides ethanol, the major components of the volatile fraction were reported to be 1-propanol, ethyl acetate, 2-methyl-1-propanol, 3-methyl-1-butanol and 2-methyl-1-butanol and the aldehydes 5-(hydroxymethyl)-2-furaldehyde and 2-furaldehyde [20]. Their detailed determination is carried out often by Gas Chromatography (GC), with several detectors although mass spectrometry (MS) outstands [6], [8], [9], [21], [22]. More specifically, it was found that methanol, 2-/3-methyl-1-butanol (typically metabolized from amino acids by yeasts during alcoholic fermentation), and 2-phenylethanol concentrations were lower in mixed than in 100% agave tequilas [10], [20]. A combination of GC–MS–MS was required to determine a carcinogenic substance in tequilas (ethyl carbamate) [23].

Sometimes dedicated sample treatments need to be developed before separation and quantitation (using, e.g., GC–MS): solid-phase microextraction (SPME) [7], [24], [25], sometimes with synthetic fibers [20], [26], headspace preconcentration [8] or a previous separation in a capillary column [27]. In relation to sample preparation to determine the volatile composition of tequila, three methods based on liquid-liquid batch and continuous extraction, as well as simultaneous distillation–extraction were evaluated [28]. Characterization studies have also been done using HPLC [5], [18], [22], [29], [30].

Some other complex analytical methods were employed, as: (i) surface plasmon resonance (SPR) (to differentiate between white, aged, and extra-aged tequilas) [31], (ii) stable $^{18}O/^{16}O$ and $^{13}C/^{14}C$ isotope ratios of ethanol (to distinguish alcohols derived from blue agave and from sugar cane or corn) using head space SPME–HRGC–IRMS (solid-phase microextraction–high resolution gas chromatography–isotope ratio mass spectrometry) [26], and (iii) analyzing D/H ratios of ethanol by SNIF-NMR [32] (site-specific natural isotopic fractionation by nuclear magnetic resonance). However, these techniques are available only in a limited number of laboratories, and their implementation is complex and costly.

Tequilas were also characterized by metals, using inductively coupled plasma (ICP), either with MS [29] or emission (AES) [33], [34] detection. Also, major ions (likely from the dilution water employed outside Mexico to fix the final alcoholic percentage) were determined by ion chromatography [35]. Also related with metals, the removal of Cu(II) from tequila was assessed with a suite of techniques, namely, infrared spectroscopy (FTIR), X-ray photoelectron spectroscopy (XPS) and GC–FID [36]. Anodic stripping voltammetry and atomic absorption spectrometry were used as well [37].

Recently a simple approach was presented to characterize tequilas employing their physicochemical properties (conductivity, density, pH, sound velocity and refractive index) and this was enough to detect products that do not met the basic quality requirements [38].

Molecular spectrometric analytical techniques outstand in terms of simplicity, economy and extraction of relevant information from the samples. Thus, UV–Vis [39], [40], [41], FTIR (medium region) [35], [42], and Raman [43] spectrometry, as well as chemiluminiscence [44] were applied to fight against tequila counterfeiting. Less frequent techniques are fluorescence spectroscopy (for which a portable UV fluorescence device was tested recently as a proof-of-concept to discriminate fake tequila from genuine ones) [45], [46] and photoacoustics [47].

It is worth noting that some excellent and relevant studies referred to above contain a relatively low number of samples and, hence, there is not a sound background to select a unique approach to authenticate tequilas (if, finally, that is possible). In fact, it was stated that a suite of analytical techniques (UV–Vis, GC–MS, etc) would be required to satisfactorily and fully characterize tequilas [39]. Confirmatory studies would be desirable to avoid possible laboratory biases in method development and also to analyse a comprehensive set of samples.

1.2. Chemometric tools

A general, common characteristic of the analytical techniques above is that they generate huge amounts of data that can nowadays be studied by chemometric tools to get multivariate models. This may constitute an important, fundamental strategy for improved quality control, authentication studies and classification and determination of the origin of tequilas, as it was reported for wine (see, e.g. Ref. [48] for a classical introductory review).

The chemometric algorithms applied to differentiate between types of tequila, original and fake tequilas, or to discriminate tequila from other products were diverse. They ranged in complexity from simple Student's *t*-tests to evaluate differences on: i) total phenolics, syringic, vanillic and protocatechuic acids between white and aged tequilas [5]; ii) biomass, reducing sugar and ethanol [16]; and iii) contents of ethyl carbamate [23]; correlation analysis to identify groups of samples from the same commercial brand [47], and ANOVA (analysis of variance) to determine differences between a set of production characteristics [16], [20], [27], [49] and ethyl carbamate in several beverages [23], to multivariate analyses. The latter include several approaches; namely:

i) Principal components analysis (PCA) to differentiate amongst types of tequilas [20], [29], [33], [35], [39], [41], [43], to find groups of tequilas and mezcal (a beverage obtained from another type of Agave) [40], to evaluate the maturation of tequila in barrels [27] (they differentiated white, aged and ripened transition tequilas during maturation), to differentiate brands of rested tequilas [50] or to evaluate ethanol in tequilas [43]. It was also employed to differentiate volatile components of mixed and 100% agave tequilas [19].
ii) Cluster analysis, curiously, was not broadly reported despite its simplicity. The most relevant example dealt with differentiating tequila brands not identified by the Regulatory Council [38].

iii) Linear discriminant analysis (LDA) was employed to differentiate five types of products (mezcal, silver, gold, aged and extra-aged tequilas) using 12 metals, being the most successful differentiation that discriminating mezcal from tequila [33]. Best results were obtained when discriminating silver tequilas from three production areas, using the same metals [34]. LDA was applied also to discriminate four brands of white tequila [41], and to evaluate aging markers derived from oak barrels in 15 samples of white, rested and aged tequilas [18].
iv) Partial least squares (PLS) was employed to develop multivariate regression models to predict the contents of three sugars in tequila and mezcal samples [40]. Its discriminant counterpart (PLS-DA) was applied to differentiate between 100% agave and mixed tequilas using UV–Vis spectra [39].
v) Finally, other complex tools, as error back-propagation artificial neural networks (ANN) [20], [33] and Support Vectors Machines (SVM) [34], [41] were used as well.

The present work is aimed to contribute to the development of reliable, fast and cheap analytical methods to characterize tequila and ascertain their quality. The review above resumed the state-of-the-art and it revealed that there are two major research ways: First, there is an obvious tendency to use HPLC and/or GC to characterize specific compounds or families of compounds that contribute to the organoleptic quality of tequila. Second, 'general purpose' molecular spectrometric techniques (mostly UV–Vis and IR) are employed to characterize tequilas and obtain like 'molecular fingerprints' that are used to authenticate them. It is clear that the latter option does not pretend to characterize tequilas in detail, but to get their overall chemical profiles (which are caused by all the constituents) and whose relevant information has to be mined either with unsupervised (exploratory) or supervised (classificatory) methods. However consensus has not been achieved on which chemometric tool(s) might be more useful to authenticate tequilas and, to the best of our knowledge, there is a lack of published papers comparing them when applied to differentiate tequilas. Taking into account the steady pace of adulteration of tequila worldwide, it would be desirable to streamline a simple (although reliable) methodology to screen its authenticity, applicable not only in common laboratories but –eventually- portable and even dedicated devices. This does not preclude the need for cutting-edge methods to go deeper when characterizing and monitoring the quality (composition) of tequilas. They would be really needed to confirm or elucidate complex situations [39].

In the next sections of this work a case study is presented to exemplify the 'hybridization' of a simple, cheap and fast analytical method (available in any laboratory) to a suite of chemometric methods. The goal here is to compare the behavior of various typical supervised classification models when the four types of the tequila (white, rested, aged and extra-aged) have to be differentiated. The chemometric tools will be introduced briefly (their complete details are out of the scope of this report), along with some of their advantages and disadvantages.

2. Experimental part (case study)

2.1. Samples

In total, 170 samples of tequila (65 white, 53 rested, 39 aged and 13 extra-aged) were included in this work; out of them, 18 were purchased in specialized Mexican liquor stores, from well-known tequila producers whose labels displayed the corresponding CRT quality seals (these were only aged and extra-aged tequilas) while the others were provided by CRT. It is worth mentioning that, to the best of our knowledge, this is the highest number of certified tequila samples considered in a unique report so far.

2.2. Instrumentation

UV–Vis spectra were measured in the 190–700 nm wavelength range, with a 1 nm nominal resolution and a 2-mm-thick quartz cell in a Lambda 35, Perkin-Elmer© double beam spectrophotometer. A synthetic blank (ethanol:water, 40:60 v/v) was subtracted automatically. No further spectral processing was done. The spectrophotometer was routinely checked for wavelength accuracy, stray light and intensity (absorbance).

Fig. 2a depicts the general appearance of the spectra. They were totally comparable to others from previous studies [40], [41]. Since white tequilas are clean and transparent, they only absorb in the UV region, with a clearly defined Gaussian-like band (centred around 280 nm whose right tail approaches zero around 325 nm); in addition, they present an increasing slope towards 200 nm. Rested, aged and extra-aged tequilas are more complex because the band ca. 280 nm become less resolved and gets partially overlapped with a highly-absorbing band centred around 205 nm, close to the operational limit of the instrument. Besides, they present an unresolved band in the 350–375 nm region whose right tail gets zero only after 400 nm. The complex spectra of rested, aged and extra-aged tequilas can be explained because of their maturation in wooden barrels for several months (as detailed in the introductory part). Hence, it is expected that many compounds migrate from the wood to tequila to yield complex flavors and tastes (some of these compounds evolve during maturation). Although three furanic compounds were demonstrate to influence greatly white tequilas (furfural, 2-acetylfurfural and 5-methylfurfural) [40], tequilas are pretty much complex and a recent study identified up to 327 compounds, which were monitored through the production process [9]. An increase on the concentration of higher alcohols, acids, esters, aldehydes, color and turbidity was reported with maturation [27].

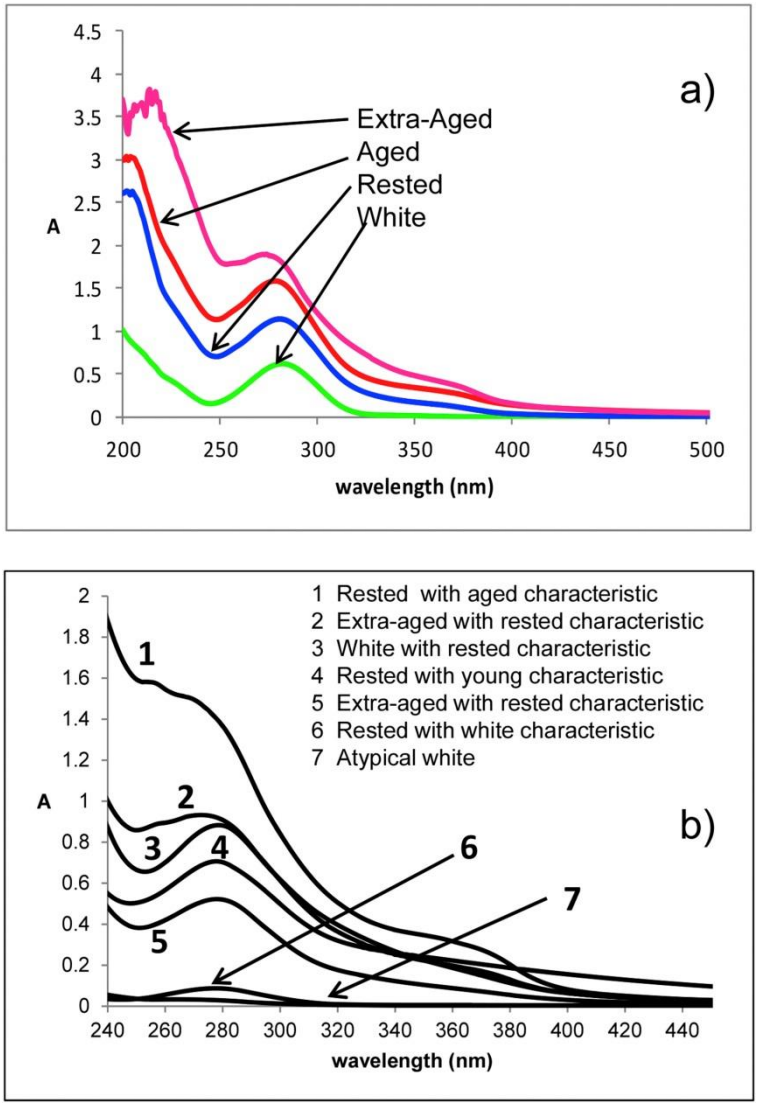

Fig. 2. General appearance of the UV–Vis spectra for each type of tequila (a) and some examples of tequilas labeled as from a class but whose spectra suggest pertainance to other class (b).

Note that the variability on the composition of tequilas (and, therefore, on their spectra) depend on many factors like the maturity of the agave, cultivation field, yeast strain used on the fermentation process, which modify the sugar composition, ethanol and volatile components of tequila [49], the cooking process, the severity of the two distillation processes, bottling (to avoid losing volatile flavors), etc. Finally, note that the use of the barrels can also have a relevant role as oak French barrels are expensive and they tend to be used as many times as possible, despite the release of the polyalcohols to the spirit is not constant throughout all their lifetime.

2.3. Chemometric methods

In the following only some schematic descriptive details are given on the chemometric methods considered in the work. Readers are kindly forwarded to the references in order to get more technical details.

The first step before developing any model should be to search for atypical (and/or outlying) samples. In authentication studies it is worth not just deleting samples with an outlying behavior but studying why they behaved anomalously (or differently). When large datasets are generated by the analytical techniques this cannot be done visually and, so, unsupervised pattern recognition methods offer their excellent ability to compress information and unravel trends. This is especially important when the 'nature' (class, quality level, etc.) of the samples is decided in a previous step –usually, out of the control of the researcher. It is worth noting that in many published works dealing with studies on tequila we felt that this step was not considered at all or not exploited fully and, many times, there was a lack of chemical interpretation of the results that were derived from chemometrics. Of course, this is not always possible because some supervised classification techniques do not offer ready-to-use information (like SVM), but when that is easily available it should be a must (typically, when exploring the data with PCA studies, hierarchical clustering and, even, LDA). That was one of the reasons why a part of the Results and Discussion section was devoted to this –in our opinion-, relevant issue.

Principal Components Analysis (PCA) and hierarchical Cluster Analysis (CA) are neither supervised nor classification algorithms but unsupervised tools to discover trends into the variables and the samples. They were employed here because they should be the first stage in every multivariate study. PCA is a variable-reduction technique intended to visualize major trends in the variables (mostly, their relationships, throughout a detailed study of the loadings) and in the samples (visualizing the scores). Further, PCA scores can be used as a starting point for many other techniques. Cluster analysis is a simple technique based on defining a similarity measure –distance- and a grouping criterion whose typical objective is to discover groups of samples, although groups of variables can be searched for, as well.

Potential functions were proposed by Forina et al. [51] as a probabilistic and distribution-free class modeling technique. The boundaries between the classes are defined assuming a multivariate normal distribution which, grossly speaking, is obtained by considering the sum of the density functions which surround each multivariate point (sample) belonging to that class in the training set. The calculations associated to this approach are not trivial (see Ref. [51] for more details) and the graphical output deserves irregular shapes. Simplified potential curves are an approach proposed to simplify the determinant method of potential functions in order to make its calculations simpler and more straightforward. The major difference between potential functions and simplified potential curves is that in the latter a two-dimensional Gaussian region is constructed for each class using two PC scores (usually, but not necessarily, the PC1 and PC2 scores), without considering a density function around each sample. The means and standard deviations required for the Gaussians are derived from the PC scores of each corresponding class samples. More details can be found elsewhere [52]. The use of PC-scores in the simplified potential curves approach precludes the use of noisy information present in the original data. From a pragmatic viewpoint, in our experience, the simplified approach works well whenever the classes do not overlap and occupy different spatial regions. An isoprobability function can be defined for each class and new (unknown) samples would be objectively classified according to the probability of belonging to each group. The isoprobability regions derived from potential curves are smooth, symmetrical (as they correspond to a typical two-dimensional Gaussian) and simple to interpret. Despite the differences in the two techniques, in essence, both look for a separated region in the experimental space that can be assigned only to a group of samples. Potential functions can be calculated with the '*Classification toolbox*' while simplified potential curves were implemented in '*GenEx*' (see section 2.4).

The K-Nearest Neighbor classification method (KNN) is a nonparametric pattern recognition method that classifies a sample in the category to which the majority of its '*K*' nearest neighbor samples belongs to. A definition for 'neighbour' is, therefore, required for which different options exist. The Euclidean distance was selected here after some preliminary assays. *K* uses to be an odd number to ensure that a majority vote is obtained locally. The determination of the optimal value for *K* is the most important part of this process [53]. We selected it after an optimization step where cross-validation procedures were used to look for the *K* value that minimized the classification error of both the training and validation sets.

SIMCA is a supervised classification method based on disjoint PCA models obtained for each class in the training set. Unknown samples are then compared to the class models and assigned to classes according to their analogy with the training samples. A new sample will be recognized as a member of a class if it is similar enough to the other members. Similarity is measured according to a statistical 'border' which is calculated according to the model residuals of the calibration (training) samples [54].

PCA–DA stands for a combination of PCA and linear discriminant analysis (DA). DA is a standard classification method based on determining multivariate linear discriminant functions, which maximize the ratio of between-class variances and minimize the ratio of within-class variances. The factors obtained in DA are termed discriminant functions or canonical variables. Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) are used depending on the linear or non-linear class separation and on the reliability of the class covariance matrices. DA requires the number of variables not to exceed the number of objects of each class. If this does not occur, as it is usually the case, an alternative consists of applying DA on the scores obtained from Principal Components Analysis (PCA), hence the PCA–DA acronym [54].

Partial Least Squares (PLS) is, likely, one of the most used multivariate regression algorithms in chemistry, and it was adapted for classification purposes. Main difference between regression and classification models is that in the latter the dependent variables (the 'Y-block') codify which objects belong to each class. When dealing with G classes, the class vector is of size G (with values of 0 – not of this class- and 1 – of this class-). The main advantage of PLS-DA is that the relevant sources of data variability are modeled by the so-called Latent Variables (LVs), which are linear combinations of the original variables and, consequently, it allows for graphical visualization and understanding of the different data patterns and relations by means of the latent variable scores and loadings. For the unknowns, PLS outputs will not have the form (0,0,…,1,…0) but real values in the range between 0 and 1. Hence, a classification rule must be applied; the object can be assigned to the class with the maximum value in the Y vector or, alternatively, a threshold between zero and one can be determined for each class [55].

CART (Classification and Regression Tree) classify samples after developing a tree-like structure, which partitions the data into mutually exclusive groups (nodes) each as pure or homogeneous as possible concerning a response variable [56], in the present work the code of the class to which the sample belongs to. The idea is to create a sequence of simple decisions (based on experimental values of a set of variables) by which the class of a sample can be predicted. The tree starts by a root node including all objects, which is divided in child nodes by binary splitting. Then, each child node is considered a root node and the process continues until a node contains either objects from a unique class or just one sample. Each split is based on a simple rule derived from an explanatory variable (typical rules are of the type: if variable 'x' < 2, then the sample(s) is (are) included in node 'y' – eventually a class). More technical details can be found elsewhere [56].

Strongly related to CART, Random Forest (RF) combines many decision trees to get a class prediction. It increases the accuracy and the interpretation of the results as small changes in the data do not usually modify the final interpretation, opposite to CART. Since many trees are considered, the final output is the mode of the predictions for all individual trees. Each tree is grown on an independent bootstrap sample (randomly selected, with replacement) from the training data (here, two-thirds of the tequilas), currently known as 'the bag'. The nodes are split by considering the best predictor variable (here, wavelength) among a subset of predictors randomly chosen at that node (hence, the term 'random'). The remaining data (ca. one-third of the tequilas) are said to be 'out of the bag' and serve as a test set for this particular tree and gives information about the estimated error rate and a variable importance rate. Thus, RF are ensemble methods whose final outcome is achieved by promediating the outputs of many trees. In principle, the more trees you use the better the results should be. However, the improvement on the predictions (classifications) of new samples decreases as the number of trees increases, so that at a certain point the gain in the prediction performance from a high number of trees will be lower that the computational cost (in time) to learn the new trees. Main advantages of RF are its robustness towards overfitting and its tendency to converge always when the number of trees is large [57]; RF can handle high-dimensional datasets, numerous missing values and unbalanced datasets [58].

Another advantage of RF is that a 'proximity matrix' can be calculated by counting the number of times that two samples are placed in the same terminal node of the same tree of RF, divided by the number of trees in the forest (to normalize the results). In this way, it is possible to visualize clusters of samples and explain them in accordance to the variables defining each one. Furthermore, a so-called 'Gini variable importance index' (in brief, 'Gini') [58] can be calculated to assess the (relative) importance of a variable on the final classification process. Roughly, it measures the total increase of impurity in a variable, when this variable has been selected for splitting. Every time a split of a node is made considering a variable (say, *m*) the Gini impurity criterion for the two child nodes is less than the parent node (because in each child node, the branches contain less diverse samples). Adding up the Gini decrement for each individual variable over all trees in the forest gives a variable importance index (which is not just a plain summatory, but a weighted one); technical details were given elsewhere [58].

Support Vectors Machines for classification (SVM-C or, simply, SVM) were developed to model two-class problems which are not linearly separable (e.g. with LDA) with the aim of classifying future unknowns. The key idea of SVM is to represent the original (usually inseparable) classes of samples in a higher dimensional space (the so-called 'feature space'). The term 'higher' refers to the original dimensionality of the problem, i.e. the variables that will be used to classify the samples. They can be either the analytical measured variables or a simplified view of them, like the principal components scores [59]. The key idea is to extend the space where the samples are represented by one or several extra dimensions so that, hopefully, the extended space will be useful to separate the classes (this is termed non-linear mapping). This can be done mathematically by combining the experimental variables according to a set of mathematical functions, called kernel functions, among which the most common ones are the linear and the radial basis (RBF) functions [60], [61], [62] although, sometimes, polynomial kernels are also used.

An appealing characteristic of SVM is that the a priori complex step of non-linear mapping of the variables to a feature space can be calculated in the original space by the kernel functions after some key parameters are optimized, remarkably a penalty parameter, which requires not only to develop a model but also to validate it with a set of samples not used at all to get the model. Overfitting must be avoided as it was demonstrated that SVM are prone to easily overfit the data [60], [61], [62]. Nowadays, efforts are being made to apply SVM to complex problems with more than two classes (for which they were originally developed) although the two classical approaches 'one-vs-all' and 'one-vs-one' are commonly applied [63]. The first was employed here, and it consists on the development of as many two-class models as different classes there are. In each

model a class is opposed to all the other ones. Finally, the pertainance of unknowns to the class(es) is defined most commonly by votation.

Due to the high dimensionality of many datasets nowadays PCA is a usual first step before SVM (which is applied to the first two scores components). However, feature selection arises as a powerful option to select a reduced subset of experimental features which lead to best results (without altering the original meaning of the variables employed to deploy the SVM). There are three major approaches for feature selection in Machine Learning [64]: ***i) filters***: they use a statistical measure to assign a scoring to each variable; then they are ranked and either selected to be kept or removed from the dataset. The methods are often univariate and consider the feature independently, or with regard to a dependent variable, this is why they tend to select redundant variables; ***ii) wrapper***: they evaluate subsets of variables and consider the selection of a set of features as a search problem, where different combinations of variables are prepared, evaluated and their performance compared. A predictive or classification model is used to evaluate each combination of variables and some kind of 'score' is given based on model accuracy. The search process may be methodical, stochastic, or it may use heuristics, like forward and backward iterations to add or remove features; ***iii) embedded***: they are inherent to the development of the model by learning which features contribute most to the accuracy of the model while the model is being created. Usually, they introduce additional constraints into the optimization of a predictive algorithm (such as a regression algorithm) that bias the model toward lower complexity (fewer coefficients).

The latter option constitutes a well-known, successful approach because those methods perform feature selection as part of the model construction process. A good example is the so-called SVM-RFE algorithm. This was developed originally for ranking genes in a cancer classification problems according to the hyperplane decision value of a SVM [65]. In brief, in SVM-RFE the analytical variables are ranked (scored) according to a coefficient, following an iterative process called Recursive Feature Elimination, RFE, which removes variables (one or more at each iteration) according to their lowest rank (or 'score' – do not confound it with the principal components scores of the samples) until the highest performance is achieved. The ranking criterion (score) is derived just when the SVM hyperplane is computed. At that moment, a weight vector is calculated to define the hyperplane and the magnitude of each element of the vector denotes the importance of the corresponding variable in the process (if the value of the coefficient for a particular variable is zero, the variable is not important for the classification problem). Then, the variables are sorted according with their power (ability) to discriminate among the classes [64].

2.4. Software

The 'Classification Toolbox for Matlab' (v.3.1) [55] was used for potential functions, PCA–DA, KNN, SIMCA and CART. GenEx© (MultiD Analysis AB, Gothenburg, Sweden) was used for PCA, Hierarchical clustering, potential curves and SVM. Random Forest and SVM-RFE were made with the Random Forest [66], Kernlab [67] and Caret [68] R packages [69].

The performance of the models was evaluated using traditional, well-known statistics [55]:

i) Precision of a class represents the capability of a classification model to not include samples of other classes in the considered class (it is the ratio between the samples of the g-th class correctly classified and the total number of samples assigned to that class).
ii) Sensitivity of a class is the ability of the model to correctly recognize samples belonging to that class.
iii) Specificity of a class describes the ability of the model to reject samples of all other classes in the considered class.

iv) Accuracy is the ratio of correctly assigned samples (it can be defined for a class alone or for the overall model considering the correctly assigned samples in all classes).
v) Class non-error rate (NER) is the average of the specificity and sensitivity of the class, whereas the model NER is the average of all class non-error rate.
vi) Class error rate (ER) or, just class error, is the complement of the NER (ER = 1−NER).

3. Results and discussion

3.1. Unsupervised pattern recognition

PCA (mean centred data, 250–450 nm) reveals that 3 major groups appear rather differentiated in the PC1–PC2 scores subspace (Fig. 3, 99.6% explained variance). They correspond roughly to the three classes of tequila defined by CRT; however, some samples became located in wrong groups.

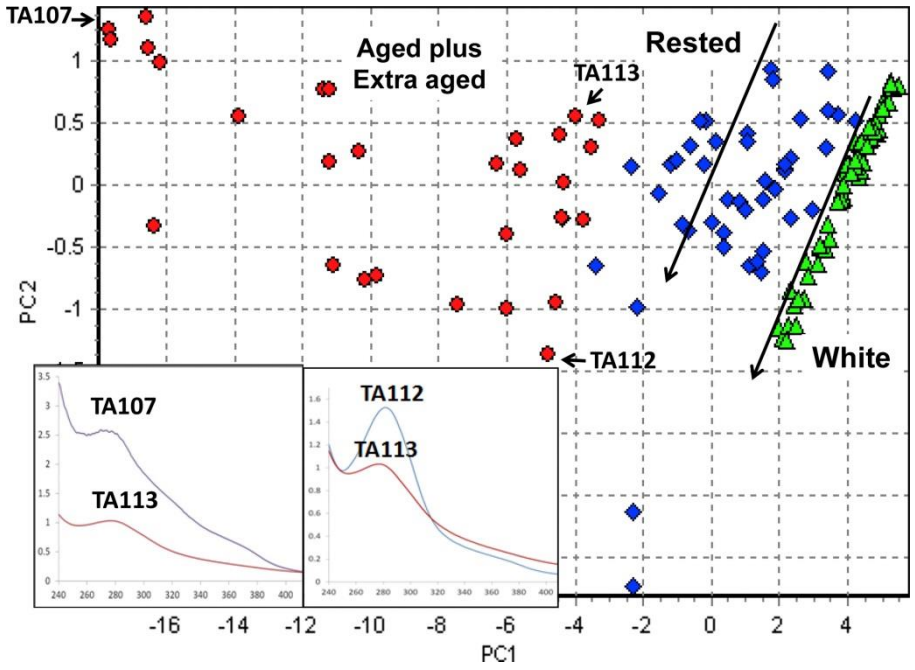

Fig. 3. PCA scores plot depicting the three classes of tequilas. The insets reveal the major differences on the spectra along PC1 and PC2 (signaled samples).

As mentioned in section 2.3, one of the very first steps when searching for classification models is to look for outliers and/or atypical samples. Here, it was found that the spectra of some tequilas labeled as white were indeed much more typical of rested tequilas, or with a totally different spectral profile than the other white tequilas in the main band around 280 nm (see Fig. 2b for two typical examples). Similarly, some rested tequilas had spectral characteristics more typical of white or, even, aged ones (Fig. 2b); and also a few aged and extra-aged tequilas showed misleading characteristics (or labeling). These situations are quite normal in food authentication studies because of the natural heterogeneity of the samples, differences in processing, raw materials, time into the barrels, etc. In some cases they may correspond to mislabeled samples or,

even, typing errors. In order to develop class models they should be avoided so that the typical performance of the classes may be considered. In case false negatives appear in future (e.g. a tequila declared as aged that is classified as white), they would be subjected to more detailed studies to confirm its legal adequacy to the standard; likely, using any measurement state-of-the-art approach reviewed in the first section.

The studies and results presented hereinafter refer to the data set without the outlying samples. Besides, aged and extra-aged tequilas will be considered in a unique group because of the low number of extra-aged tequilas (only 13) and their similar behavior (preliminary tests could not differentiate them), as it happened in other studies [40].

Fig. 3 reveals that not only the three classes of tequilas can be differentiated grossly but that an inner pattern exists within each group. White tequilas became ordered in a very distinct narrow, elongated pattern. Tequilas located at the upper extreme show the lowest absorbances whereas those in the lowest extreme correspond to highest absorbances. The same behavior appears for rested tequilas. The group formed by aged plus extra-aged tequilas is a bit more complex, however, they show an interesting distribution: considering samples with approximately the same PC1 scores (e.g. TA113 and TA112), those with highest PC2 scores exhibit lower peak intensities around 280 nm (see insets in Fig. 3). On the contrary, samples with almost equal PC2 scores reveal hugely different baseline slopes (TA113 and TA107); the more displaced the samples are to the left side, the more intense their absorbances are (see insets in Fig. 3). This, of course, is a direct consequence of the loadings profiles (Fig. 4). In general, the profile of the PC1-loadings takes account of the change in the slope of the 'baseline' underlying all spectra, this feature is responsible for 97.8% of the information. The loadings for PC2 are clearly associated to the spectral band around 280 nm (1.8% of the information). Less relevant mathematically, although interesting from a chemical viewpoint, is that PC3 (0.3% of the information) and PC4 (0.1% of the information) seemed associated mostly to the slope before 240 nm, and to the unresolved band in the 290–340 nm region, respectively (see Fig. 4).

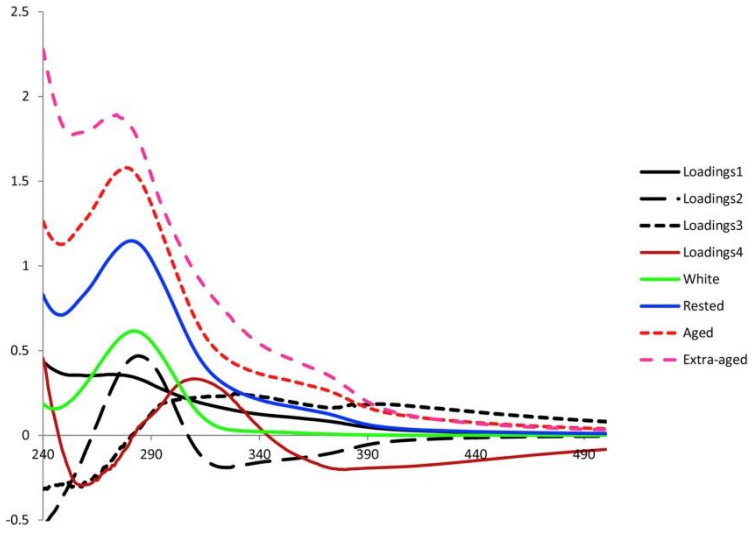

Fig. 4. Comparison of original tequila spectra (one of each class: white, rested, aged, extra-aged) and the first four factor loadings.

Hierarchical cluster analysis rendered very satisfactory results, not only using the first 6 component scores as independent variables to avoid masking the underlying patterns due to the high correlation between the spectral variables [56] but considering the spectra themselves. The groups were slightly more appealing using the spectra than the scores, likely because some minor details got hidden when the last PCs were deleted. A possible justification for the dendrograms being almost the same regardless of using either the spectral data or the scores might be the spectral differences between the three main classes of tequilas (white, rested and aged plus extra-aged) in some spectral regions. The dendrogram depicted in Fig. 5 was obtained using the Ward's method and the Manhattan distance (mean centred data, variables in the 240–450 nm range).

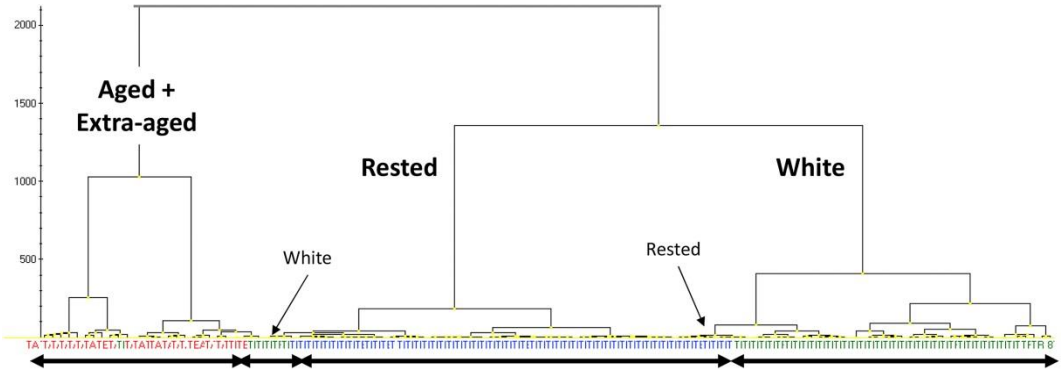

Fig. 5. Dendrogram obtained for the collection of tequilas (see text for more details).

The dendrogram shows three major groups of samples. One devoted exclusively to the aged and extra-aged tequilas (but for a rested tequila, T70 which is included here), a second cluster with rested tequilas (although it includes a subcluster with white tequilas) and a third major group is associated to white tequilas (although with a subgroup of rested ones).

3.2. Supervised classification

3.2.1. K-nearest neighbors

KNN yielded very good results, with satisfactory performance characteristics for calibration and validation (Table 1). The model was developed considering mean centred UV–Vis spectra, K = 3 neighbors (as decided by venetian-blinds cross validation, 10 cancellation groups, on the calibration set), and the Euclidean distance as a measure of similarity. The non-error-rate was 0.98 (or 98%), whereas the error-rate in cross-validation was 0.02; due to three misclassified rested tequilas; namely, T56 and T63 (considered as white) and T70 (considered as aged/extra-aged). In validation, two rested samples (T130 and T135) were classified as aged (because their spectra looked like aged samples, indeed).

Table 1. Resume of the performance parameters for the different models (mean centred spectra, venetian blinds cross-validation, 10 segments).

Method		Calibration (cross validation)				Validation			
		Accuracy	Precision	Sensitivity	Specificity	Accuracy	Precision	Sensitivity	Specificity
KNN	C1	0.98	0.96	1.00	0.97	0.89	1.00	1.00	1.00
	C2		1.00	0.93	1.00		1.00	0.71	1.00
	C3		0.96	1.00	0.99		0.71	1.00	0.85
Potential functions	C1	0.96	0.98	1.00	0.98	0.88	1.00	1.00	1.00
	C2		1.00	0.85	1.00		1.00	0.67	1.00
	C3		0.88	1.00	0.96		0.71	1.00	0.82
PCA–DA	C1	0.97	1.00	0.98	1.00	0.89	1.00	1.00	1.00
	C2		0.98	0.93	0.99		1.00	0.71	1.00
	C3		0.90	1.00	0.97		0.71	1.00	0.85
SIMCA	C1	0.84	1.00	0.74	1.0	0.83	1.00	0.83	1.00
	C2		0.74	0.86	0.84		0.83	0.71	0.91
	C3		0.79	1.00	0.93		0.71	1.00	0.85
PLS-DA	C1	0.81	0.81	0.81	0.89	1.00[a]	1.00	1.00	1.00
	C2		0.81	0.71	0.88		1.00	1.00	1.00
	C3		0.80	1.00	0.93		1.00	1.00	1.00
CART	C1	0.98	0.98	0.98	0.96	0.94	1.00	1.00	0.83
	C2		0.98	0.95	1.0		1.00	0.86	1.00
	C3		0.99	0.99	0.99		0.83	1.00	0.92
Random Forest	C1	0.98	1.00	1.00	1.00	0.94	1.00	1.00	1.00
	C2		0.98	0.98	0.99		0.86	1.00	0.92
	C3		0.96	0.96	0.99		1.00	0.83	1.00
SVM[b] white vs all	C1	0.99	0.98	1.00	0.99	1.00	1.00	1.00	1.00
	C2		1.00	0.99	1.00		1.00	1.00	1.00
SVM Rested vs all	C1	0.99	1.00	0.98	1.00	0.89	1.00	0.71	1.00
	C2		0.99	1.00	0.98		0.85	1.00	0.71
SVM Aged vs all	C1	0.98	0.96	0.96	0.99	0.94	0.83	1.00	0.92
	C2		0.99	0.99	0.96		1.00	0.92	1.00
SVM-RFE	C1	0.99	1.00	1.00	1.00	0.94	1.00	1.00	1.00
	C2		0.98	1.00	0.99		0.86	1.00	0.92
	C3		1.00	0.96	1.00		1.00	0.83	1.00

C1 = white, C2 = rested and C3 = aged + extra aged tequilas (but for SVM).
[a] 56% of unassigned samples.
[b] SVM without cross-validation; C1 represents the class under study (see text for details).

3.2.2. Potential functions

Results for Potential Functions and Simplified Potential Curves were unsatisfactory due to a large overlap between the isoprobability regions of the classes (Fig. 6). Potential Curves were more affected than Potential Functions because the former consider the PC1–PC2 scores of each group of tequila to deploy a classical bivariate Gaussian. The Gaussians for rested and aged plus extra-aged tequilas overlapped seriously and, also the rested Gaussian expanded to the region of white tequilas. Therefore, calibration and validation were quite unsuccessful. Potential Functions performed a bit better thanks to its capability of yielding irregular shapes (Fig. 6) although disappointingly many samples (24%!) became unclassified in calibration. Worst figures were obtained always for the rested samples (C2 in Table 1); e.g. the ability of the model to correctly recognize rested tequilas (sensitivity) was only 67%. Accuracy for the overall model was only 88%.

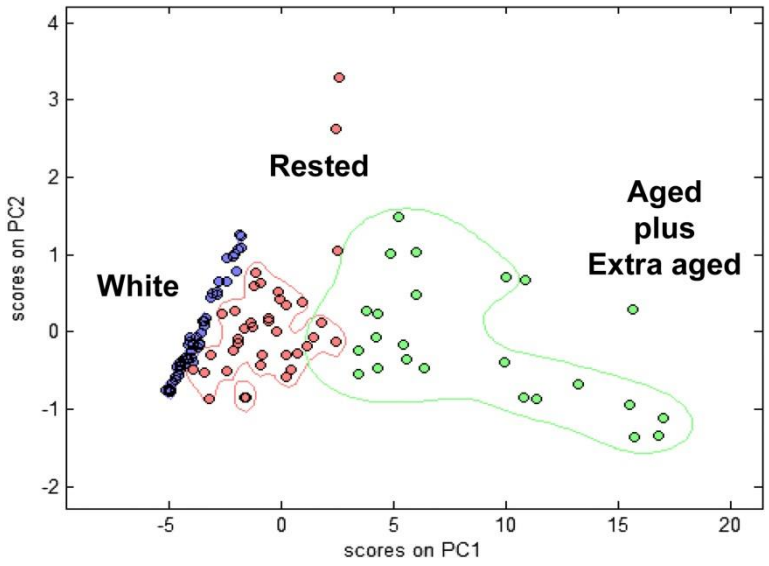

Fig. 6. PC1–PC2 scores scatterplot and general shape of the isoprobability regions obtained for the Potential Functions.

3.2.3. SIMCA

Models were developed considering a PC for each type of tequila (mean centred data), after preliminary cross-validation studies. The amount of explained information was 99% (white tequilas), 87% (rested tequilas) and 96% (aged and extra-aged tequilas). In SIMCA, unknowns are assigned to a class whenever their distance is lower than a critical value calculated for that class. The performance parameters were not good (Table 1), with a low 84% overall accuracy, high error rate (13% for cross-validation training; 15% for validation with the external set) and 18 (c.a. 15%) not assigned samples in calibration. The other performance characteristics can be seen on Table 1.

Although SIMCA has traditionally been considered a *de facto* reference method results here were not good; in our opinion due to the quite different inner variance of the samples in each group. The problem of the bad performance of SIMCA under this circumstance had already been pointed out [70].

3.2.4. Discriminant analysis

PCA–DA was selected to handle more variables than samples for discriminant analysis (2 principal components were selected by cross-validation). In particular, Quadratic Discriminant analysis was applied because the variance was different between the classes.

The error rate was really good for both calibration (3% using cross-validation) and validation (1%). The performance parameters were very satisfactory for calibration, but not so good for external validation, only 89% accuracy. As in most trials in this work, the model performed worse for the rested class: 71% sensitivity (ability of the class model to recognize truly rested tequilas), with 4 rested tequilas being classified as aged (which made the precision of the aged plus extra-aged class to be poor, 71%), and 2 unclassified samples. One white tequila was considered as rested.

With regards to PLS-DA, three latent variables (LV) were considered using venetian blinds cross-validation (10 segments, mean centred data). The amount of explained information was about 100% for the X-block although only 25% for the Y-block. This might justify the bad results shown in Table 1 for both calibration and validation. Despite validation seems excellent, it is not indeed because the model let 56% of the validation samples unclassified. This also happened in calibration, for which 53.7% of the samples became unassigned. Other number of latent variables were tried although unsuccessfully. Therefore, common PLS-DA was discarded.

3.2.5. Classification trees and random forest

When CART was applied to the UV–Vis data (240–450 nm) only two variables were required to achieve a satisfactory classification. If the absorbance at 324 nm is lower than 0.0473, a child node contains only White tequilas. The other node contains rested and aged + extra-aged tequilas. They are differentiated by the absorbance at 240 nm; if it is greater than 1.057, aged + extra-aged tequilas are obtained (plus a misclassified rested sample); otherwise they are rested tequilas. The confusion matrices for calibration, leave-one-out cross validation (LOOCV) and validation (an independent set of samples) yielded excellent values for precision, sensitivity and accuracy (see Table 1).

It is worth noting that despite three samples were misclassified in calibration (coded as T24, T56 and T70) they had spectra which, indeed, corresponded to the CART assigned class. Hence, some kind of error in labeling or codification might have occurred with them (as mentioned above when discussing about PCA and some other samples, as T130 and T131).

RF was performed considering a standard value of 500 trees (as mentioned above, the number of trees is not critical, as it does not lead to overfitting), the best predictor variable to split the nodes was selected from a subset of *zz* variables per level (*zz* varied randomly from 2 to 17, the latter being the square root of the total number of spectral variables). Best results were achieved with 4 randomly selected variables and an ‘out-of-bag’ (validation) error of 1.63% (i.e., the average error for the samples left out of the training stages).

Fig. 7 depicts the PC1–PC2 subspace of a PCA made on the proximity matrix derived from the RF (calibration samples) to which the validation samples were projected. There, it can be seen that RF is really good in differentiating the three classes of tequilas; only two training samples (T70 and T120) and a validation sample (T130, as for the other methods) were misclassified. However, note that the two misclassified calibration samples are indeed further from the classes indicated into their labels. In an attempt to foresee the most important variables on the final classification process, a ‘Gini’ index plot was obtained (Fig. 8). There, it can be observed that the variables that contribute most to decrease the heterogeneity of the groups of tequilas are situated before the main spectral peaks, between 333 and 380 nm (Gini index > 0.98). They correspond essentially to the region where white tequilas reached a baseline, aged plus extra-aged tequilas show an unresolved spectral band (which gives rise to a clear shoulder, as mentioned in section 2.1) and rested tequilas have an intermediate behavior (Fig. 1).

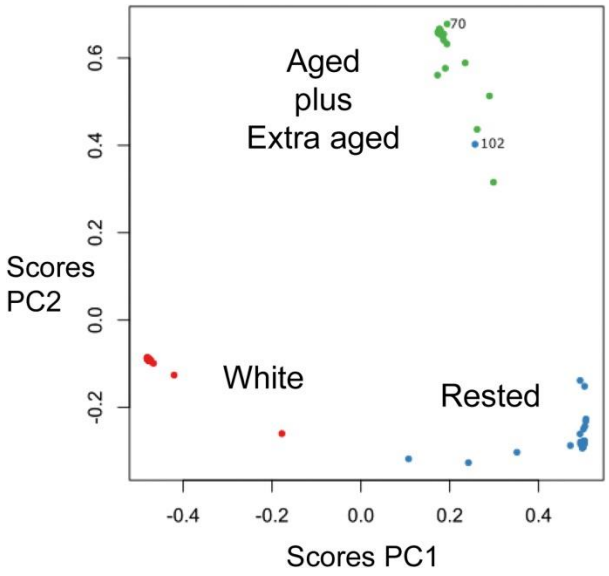

Fig. 7. PC1–PC2 scores scatterplot derived from the proximity matrix obtained by random forest. The two misclassified validation samples have been labeled.

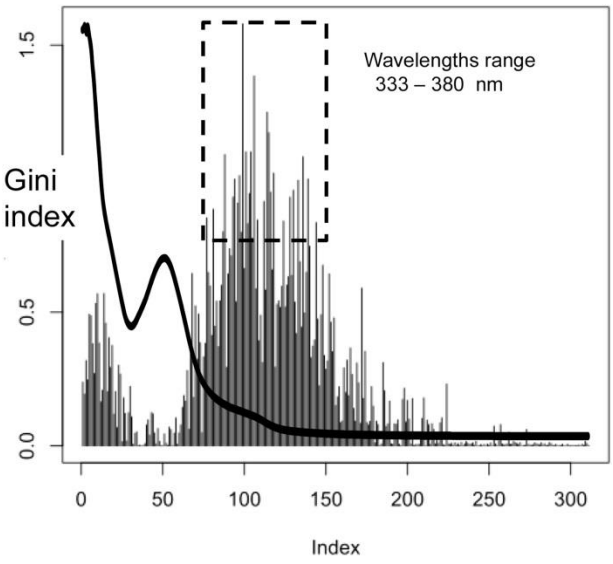

Fig. 8. Gini index plot indicating the most relevant variables associated to the RF model. The continuous line depicts a typical spectrum.

3.2.6. Support Vectors Machines

SVM was applied considering the 'one-vs-all' strategy (a class was confronted to all other samples). Thus, in Table 1 C1 represents the class containing all the samples we are interested at, whereas C2 contains all other samples. Two classes seemed separable by linear kernels (white and aged-plus-extra-aged tequilas), opposite to rested tequilas which relied in between the other two types and, therefore, a Gaussian SVM was required. The parameters to model the classes were: white tequilas: a linear kernel, with a penalty parameter C = 10 (1000 iterations); aged plus extra-aged tequilas: a linear kernel with C = 1000 (1000 iterations); rested tequilas: sigma = 2 (values from 1 to 4 were assayed), C = 100. In all cases the values assayed for the penalty parameter, C, were 0.1, 1, 10, 100 and 1000. Polynomial kernels were attempted (degrees 2, 3 and 4) although unsuccessfully. Fig. 9 depicts the SVM models selected above, with both the calibration and the validation samples for each class.

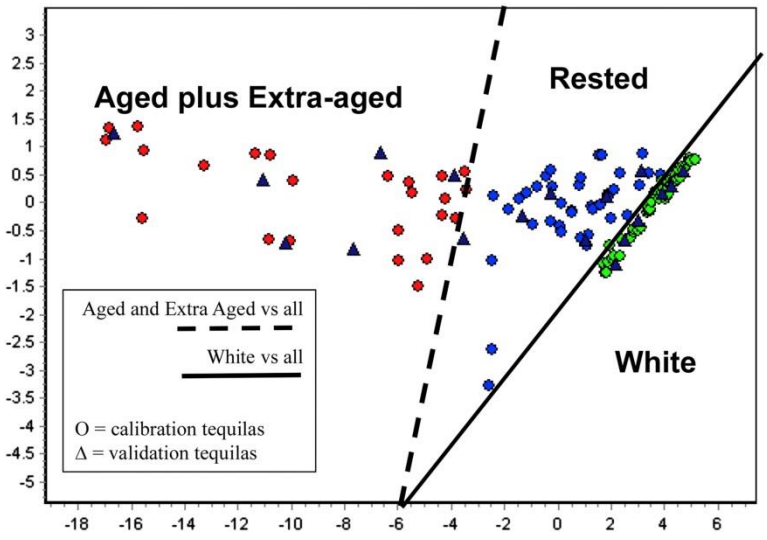

Fig. 9. Borders of the SVM linear models to differentiate white and aged tequilas.

Table 1 indicates the performance parameters, where from it can be deduced that white tequilas were modeled with no error neither in calibration nor in validation; aged plus extra-aged tequilas can be modeled fine (no error) although a misclassification was found for the validation samples (T130, which corresponded to the same sample discussed in the previous models). Finally, rested tequilas are modeled fine (only a sample became classified as white) and two validation samples (T130, as expected, and T135) were considered as aged. Hence, the performance parameters were good. In calibration (training) all figures were above 0.95, which denotes good ability to accept samples belonging to a class and, also, reject others from different classes. Accuracy and precision values were close to 1 (100% success). The validation performance for the group of rested tequilas worsened up to 71% sensitivity and 89% accuracy.

SVM-RFE was also applied considering the 'one-vs-all' strategy and a linear kernel. The penalty parameter was tested at 0.1, 0.01, 1, 5, 10, 50, 100 and 200. The size of the subset of features varied from 2^2 to 2^8 (the exponential ranged from 2 to 8). The best results (see Table 1) were obtained with C = 50 and 16 variables; 6 variables in the 498–503 nm region and 10 variables in the 540–550 nm range. Both subsets are at the extreme of the right tail of the spectra, with an unclear chemical interpretation although they are capable of unraveling differences between the classes (as mentioned in section 2.1 and the discussion on random forest). The

average values for white, rested and aged tequilas differ at that spectral region (e.g., −0.012 μA, 0.004 μA and 0.024 μA for white, rested and aged tequilas, respectively), which justifies its selection.

As for most other models, samples T130 (validation) and T70 (calibration) were misclassified (labeled as rested, classified as aged).

4. Conclusions

The first objective of this work was to review the analytical characterization efforts undergone so far to authenticate tequilas and differentiate among the three most important commercial types. It was found that chromatographic techniques outstand in terms of determining specific compounds and, likely, this will continue in a near future, probably in combination with dedicated sample preparation techniques (like solid-phase microextraction). However, these techniques are not too suited to be deployed in routine analysis with large collection of samples or field studies and, so, general-purpose measurements have been proposed to characterize the general chemical features of the samples, mostly using spectral measurements. In general, all analytical methods require chemometric tools to gather the most relevant information that allow for objective authentication and/or classification of a sample.

The case study presented here deals with the largest number of certified tequilas considered so far (to the best of our knowledge) and it shows that linear models are, in general, not the optimum choice to handle spectral data (in particular, UV–Vis) and, so, more advanced methods need to be used.

K-nearest neighbors, a quite simple classification method, yielded very good results, comparable to Quadratic Discriminant Analysis (with a previous PCA dimensionality reduction). Best results were obtained using Classification and Regression Trees (CART), Random Forest (RF) and Support Vectors Machines (SVM), with accuracies higher than 0.98 and 0.94 (calibration and validation, respectively). For the latter, two strategies were tried: *one-vs-all* approach (one model per class) and recursive feature elimination (a model for the three classes), although without relevant differences.

It is relevant to chemically interpret the models and although this is quite straightforward with linear techniques, this topic has been misconsidered frequently (even when PCA results were reported). More complex methods are not easy to interpret but, at least, they underline the importance of particular regions of the spectra to classify tequilas (e.g. the Gini index from RF or the set of variables selected by SVM-RFE). In our case study, the most relevant spectral region was that in the visible part, at the tails of the spectral bands (whose chemical assignations are very complex, although they seem related to the compounds extracted from the oak barrels employed for tequila maturation).

Another interesting issue is that the combination of spectroscopy and chemometrics allows for the identification of certified samples whose spectral characteristics and the subsequent classifications revealed an outlying behavior. In our view, they might have been labeled incorrectly, although more studies are needed here because the quality of some tequilas is certified according to some rules that determine how the product is handled, rather on chemical characteristics.

In our view, the results presented here agree with most previous studies and demonstrate that spectroscopy measurements (in particular, UV–Vis ones) combined with chemometric methods (in particular, non linear classification methods) constitute a reliable and practical tool to authenticate tequilas and to classify them according to their typical commercial types. This approach is fast,

simple to implement in industries and/or regulation organizations, it does not almost consume reagents (and, so, takes account of the green chemistry principles) and portable.

Acknowledgments

The Galician Government, "Xunta of Galicia", is acknowledged for its support to the QANAP group (Programa de Consolidación y Estructuración de Unidades de Investigación Competitiva, GRC2013-047).

G. P-C acknowledges a research grant from the UNAM to partially support this work.

C. F-L., acknowledges a Juan de la Cierva fellowship grant from the Spanish Ministry of Economy, Industry and Competitiveness (Ref. FJCI-2015-26071).

References

[1] NOM-006-SCFI-2012. Bebidas alcohólicas-Tequila-Especificaciones (p. 20). México, Diario Oficial de la Federación (2012)
[2] Federal Register, Title 27, 5.22(g) (1973)
[3] Off. J. Eur. Comm., L152 (1997), pp. 16-26
[4] Tequila Regulatory Council. General information available free at: https://www.crt.org.mx/ (last access 22.12.16).
[5] A. Alcazar-Magana, K. Wrobel, J.C. Torres-Elguera, A.R. Corrales-Escobosa, K. Wrobel, Determination of small phenolic compounds in tequila by liquid chromatography with ion trap mass spectrometry detection, Food Anal. Methods, 8 (2015), pp. 864-872
[6] S.M. Benn, T. Peppard, Characterization of tequila flavor by instrumental and sensory analysis, J. Agric. Food Chem., 44 (1996), pp. 557-566
[7] A. De León-Rodríguez, L. González-Hernández, A. De La Rosa, P. Escalante-Minakata, M.G. López, Characterization of volatile compounds of mezcal, an ethnic alcoholic beverage obtained from Agave salmiana, J. Agric. Food Chem., 54 (4) (2006), pp. 1337-1341
[8] De León-Rodríguez, P. Escalante-Minakata, M. Jiménez-García, G. Ordoñez-Acevedo, J.L. Flores Flores, A.P. Barba de la Rosa, Characterization of volatile compounds from ethnic agave alcoholic beverages by gas chromatography-mass spectrometry, Food Technol. Biotechnol., 46 (4) (2008), pp. 448-455
[9] N. Prado-Jaramillo, M. Estarrón-Espinosa, H. Escalona-Buendía, R. Cosío-Ramírez, S.T. Martín-del-Campo, Volatile compounds generation during different stages of the Tequila production process. A preliminary study, Food Sci. Technol., 61 (2015), pp. 471-483
[10] E. Waleckx, A. Gschaedler, B. Colonna-Ceccaldi, P. Monsan, Hydrolysis of fructans from Agave tequilana Weber var. azul during the cooking step in a traditional tequila elaboration process, Food Chem., 108 (1) (2008), pp. 40-48
[11] N.A. Mancilla-Margalli, M.G. López, Generation of Maillard compounds from inulin during the thermal processing of Agave tequilana Weber Var. Azul, J. Agric. Food Chem., 50 (4) (2002), pp. 806-812
[12] A. Avila-Fernández, X. Rendón-Poujol, C. Olvera, F. González, S. Capella, A. Peña-Alvarez, A. López-Munguía, Enzymatic hydrolysis of fructans in the tequila production process, J. Agric. Food Chem., 57 (12) (2009), pp. 5578-5585
[13] M. Jacyno, Analysis of complex carbohydrate profiles in tequila using evaporative light scattering detection, LC GC N. Am. February (2006), p. 41
[14] D.M. Díaz-Montaño, M.L. Délia, M. Estarrón-Espinosa, P. Strehaiano, Fermentative capability and aroma compound production by yeast strains isolated from Agave tequilana Weber juice, Enzyme Microb. Technol., 42 (7) (2008), pp. 608-616
[15] I.W. González-Robles, M. Estarrón-Espinosa, D.M. Díaz-Montaño, Fermentative capabilities and volatile compounds produced by Kloeckera/Hanseniaspora and Saccharomyces yeast strains in pure and mixed cultures during Agave tequilana juice fermentation, Antonie van Leeuwenhoek, 108 (3) (2015), pp. 525-536

[16] D.M. Díaz-Montaño, E. Favela-Torres, J. Córdova, Improvement of growth, fermentative efficiency and ethanol tolerance of Kloeckera africana during the fermentation of Agave tequilana juice by addition of yeast extract, J. Sci. Food Agric., 90 (2010), pp. 321-328
[17] I.W. Gonzalez-Robles, D.J. Cook, The impact of maturation on concentrations of key odour active compounds which determine the aroma of tequila, J. Inst. Brew., 122 (3) (2016), pp. 369-380
[18] A. Muñoz-Muñoz, C. Grenier, H. Gutiérrez-Pulido, J. Cervantes-Martínez, Development and validation of a high performance liquid chromatography-diode array detection method for the determination of aging markers in tequila, J. Chromatogr. A, 1213 (2008), pp. 218-223
[19] D.K. Aguilera-Rojo, S.T. Martin-Del-Campo, R. Cosio-Ramirez, H. Escalona-Buendia, M. Estarron-Espinosa, Identification of distinctive parameters between Tequila mixto and Tequila 100% agave by gas chromatography, J.L. LeQuere, P.X. Etievant (Eds.), Flavour Research at the Dawn of the Twenty-first Century (2003)
[20] S.G. Ceballos-Magaña, F. De Pablos, J.M. Jurado, M.J. Martín, A. Alcázar, R. Muñiz-Valencia, R. Gonzalo-Lumbreras, R. Izquierdo-Hornillos, Characterisation of tequila according to their major volatile composition using multilayer perceptron neural networks, Food Chem., 136 (2013), pp. 1309-1315
[21] D. Lachenmeier, E. Sohnius, R. Attig, M. López, Quantification of selected volatile constituents and anions in Mexican Agave spirits (Tequila, Mezcal, Sotol, Bacanora), J. Agric. Food Chem., 54 (11) (2006), pp. 3911-3915
[22] A. Peña-Alvarez, S. Capella, R. Juárez, C. Labastida, Characterization of three Agave species by gas chromatography and solid-phase microextraction-gas chromatography-mass spectrometry, J. Chromatogr. A, 1027 (1–2) (2004), pp. 131-136
[23] D.W. Lachenmeier, F. Kanteres, T. Kuballa, M.G. López, J. Rehm, Ethyl carbamate in alcoholic beverages from Mexico (Tequila, Mezcal, Bacanora, Sotol) and Guatemala (Cuxa): Market Survey and Risk assessment, Int. J. Environ. Res. Public Health, 6 (1) (2009), pp. 349-360
[24] A. Pena-Alvarez, S. Capella, R. Juarez, C. Labastida, Determination of terpenes in tequila by solid phase microextraction-gas chromatography-mass spectrometry, J. Chromatogr. A, 1134 (1–2) (2006), pp. 291-297
[25] B. Vallejo-Córdoba, A. González-Córdova, M.C. Estrada-Montoya, Tequila volatile characterization and ethyl ester determination by solid-phase microextraction gas chromatography/mass spectrometry analysis, J. Agric. Food Chem., 52 (18) (2004), pp. 5567-5571
[26] B. Aguilar-Cisneros, M. López, E. Richling, F. Heckel, P. Schreier, Tequila authenticity assessment by headspace SPME-HRGC-IRMS analysis of $^{13}C/^{12}C$ and $^{18}O/^{16}O$ ratios of ethanol, J. Agric. Food Chem., 50 (26) (2002), pp. 7520-7523
[27] J.E. López-Ramírez, S.T. Martín-del-Campo, H. Escalona-Buendía, J.A. García-Fajardo, M. Estarrón-Espinosa, Physicochemical quality of tequila during barrel maturation. A preliminary study, CyTA-J. Food, 11 (3) (2013), pp. 223-233
[28] S.T. Martin-del-Campo, H.E. Gomez-Hernandez, H. Gutierrez, H. Escalona, M. Estarron, R. Cosio-Ramirez, Volatile composition of tequila. Evaluation of three extraction methods, CYTA-J. Food, 9 (2) (2011), pp. 152-159
[29] Rodríguez Flores, J. Alberto Landero Figueroa, K. Wrobel, K. Wrobel, ICP-MS multi-element profiles and HPLC determination of furanic compounds in commercial tequila, Eur. Food Res. Technol., 228 (2009), pp. 951-958
[30] R. Muñoz, K. Wrobel, Determination of aldehydes in tequila by high performance liquid chromatography with 2,4-dinitrophenylhydrazine derivatization, Eur. Food Res. Technol., 221 (6) (2005), pp. 798-802
[31] G. Martínez, D. Luna, D. Monzón, R. Valdivia, Optical method to differentiate tequilas based on angular modulation surface plasmon resonance, Opt. Lasers Eng., 49 (2011), pp. 675-679
[32] Bauer-Christoph, N. Christoph, B. Aguilar-Cisneros, M. López, E. Richling, A. Rossmann, P. Schreier, Authentication of tequila by gas chromatography and stable isotope ratio analyses, Eur. Food Res. Technol., 217 (2003), pp. 438-443
[33] S.G. Ceballos-Magaña, J.M. Jurado, Quantification of twelve metals in tequila and mezcal spirits as authenticity parameters, J. Agric. Food Chem., 57 (2009), pp. 1372-1376
[34] S.G. Ceballos-Magaña, J.M. Jurado, R. Muñiz-Valencia, A. Alcázar, F. De Pablos, M.J. Martín, Geographical authentication of tequila according to its mineral content by means of support vector machines, Food Anal. Methods, 5 (2012), pp. 260-265
[35] D.W. Lachenmeier, E. Richling, M.G. López, W. Frank, P. Schreier, Multivariate analysis of FTIR and ion chromatographic data for the quality control of tequila, J. Agric. Food Chem., 59 (2005), pp. 2151-2157

[36] A. Carreon-Alvarez, A. Herrera-Gonzalez, N. Casillas, R. Prado-Ramirez, M. Estarron-Espinosa, V. Soto, W. De la Cruz, M. Barcena-Soto, S. Gomez-Salazar, Cu (II) removal from tequila using an ion-exchange resin, Food Chem., 127 (2011), pp. 1503-1509
[37] Carreon-Alvarez, N. Casillas, J.G. Ibanez, F. Hernandez, R. Prado-Ramirez, M. Barcena-Soto, S. Gomez-Salazar, Determination of Cu in tequila by anodic stripping voltammetry, Anal. Lett., 41 (3) (2008), pp. 469-477
[38] Carreon-Alvarez, A. Suarez-Gomez, F. Zurita, S. Gomez-Salazar, J.F.A. Soltero, M. Barcena-Soto, N. Casillas, P. Gutierrez, E.D. Moreno-Medrano, Assessment of physicochemical properties of tequila brands: authentication and quality, J. Chem. (2016), 10.1155/2016/6254942 (Open access paper)
[39] O. Barbosa-García, G. Ramos-Ortiz, J.L. Maldonado, J.L. Pichardo-Molina, M.A. Meneses-Nava, J.E. Landgrave, J. Cervantes-Martínez, UV-Vis absorption spectroscopy and multivariate analysis as a method to discriminate tequila, J. Spectrochim. Acta A, 66 (2007), pp. 129-134
[40] A.C. Muñoz-Muñoz, J.L. Pichardo-Molina, G. Ramos-Ortíz, O. Barbosa-García, J.L. Maldonado, M.A. Meneses-Nava, N.E. Ornelas-Soto, A. Escobedo, P.L. López-de-Alba, Identification and quantification of furanic compounds in tequila and mezcal using spectroscopy and chemometric methods, J. Braz. Chem. Soc., 21 (6) (2010), pp. 1077-1087
[41] U. Contreras, O. Barbosa-García, J.L. Pichardo-Molina, G. Ramos-Ortiz, J.L. Maldonado, M.A. Meneses-Nava, N.E. Ornelas-Soto, P.L. López-de-Alba, Screening method for identification of adulterate and fake tequilas by using UV-Vis spectroscopy and chemometrics, Food Res. Int., 43 (2010), pp. 2356-2362
[42] U. Arzberger, D.W. Lachenmeier, Fourier Transform Infrared Spectroscopy with multivariate analysis as a novel method for characterizing alcoholic strength, density, and total dry extract in spirits and liqueurs, Food Anal. Methods, 1 (2008), pp. 18-22
[43] Frausto-Reyes, C. Medina-Gutiérrez, R. Sato-Berrú, L.R. Sahagún, Qualitative study of ethanol content in tequilas by Raman spectroscopy and principal component analysis, Spectrochim. Acta A, 61 (2005), pp. 2657-2662
[44] M.J. Navas, A.M. Jiménez, Chemiluminescent methods in alcoholic beverage analysis, J. Agric. Food Chem., 47 (1) (1999), pp. 183-189
[45] N. Leesakul, S. Pongampai, P. Kanatharana, P. Sudkeaw, Y. Tantirungrotechai, C. Buranachai, A new screening method for flunitrazepam in vodka and tequila by fluorescence spectroscopy, Luminiscence, 28 (1) (2012), pp. 76-83
[46] J.M. de la Rosa Vázquez, D.A. Fabila-Bustos, L.F.J. Quintanar-Hernández, A. Valor, S. Stolik, Detection of counterfeit tequila by fluorescence spectroscopy, J. Spectrosc. (2015), 10.1155/2015/403160 (Open access journal)
[47] A. Ruiz-Pérez, J.I. Pérez-Castañenda, R. Castañeda-Guzmán, S.J. Pérez-Ruiz, Determination of tequila quality by photoacoustic analysis, Int. J. Thermophys., 34 (2013), pp. 1695-1708
[48] Arvanitoyannis, M. Katsota, E. Psarra, E.H. Soufleros, S. Kallithraka, Application of quality control methods for assessing wine authenticity: use of multivariate analysis (chemometrics), Trends Food Sci. Technol., 10 (1999), pp. 321-336
[49] L. Pinal, E. Cornejo, M. Arellano, E. Herrera, L. Núñez, J. Arrizon, A. Gschaedler, Effect of Agave tequilana age, cultivation field location and yeast strain on tequila fermentation process, J. Ind. Microbiol. Biotechnol., 36 (2009), pp. 655-661
[50] U. Contreras-Loera, O. Barbosa-García, G. Ramos-Ortíz, J.L. Pichardo-Molina, M.A. Meneses-Nava, J.L. Maldonado, Identificación y discriminación de Tequilas reposados in situ para la protección de marca, Nova Sci., 1 (2) (2009), pp. 22-32
[51] M. Forina, C. Armanino, R. Leardi, G. Drava, A class-modelling technique based on potential functions, J. Chemom., 5 (1991), pp. 435-453
[52] A. Carlosena, J.M. Andrade, X. Tomás, E. Fernández, D. Prada, Classification of edible vegetables affected by different traffic intensities using potential curves, Talanta, 48 (1999), pp. 795-802
[53] M. Otto, Chemometrics, Wiley-VCH, Weinheim, Germany (2007)
[54] D.L. Massart, B.G.M. Vandeginste, L.M.C. Buydens, S. De Jong, P.J. Lewi, J. Smeyers-Verbeke, Handbook of Chemometrics and Qualimetrics, Part A, Elsevier, Amsterdam (1997)
[55] Ballabio, V. ConsonniClassification tools in chemistry. Part 1: linear models. PLS-DA, Anal. Methods, 5 (2013), pp. 3790-3798
[56] F. Qestier, R. Put, D. Coomans, B. Walczak, Y. Vander Heyden, Chemom. Intell. Lab. Syst., 76 (2005), pp. 45-54
[57] Y. Liu, S. Tang, C. Fernandez-Lozano, C.R. Munteanu, A. Pazos, Y. Yi-zun, T. Zhiliang, H. González-Díaz, Experimental study and Random Forest prediction model of microbiome cell surface hydrophobicity, Expert Syst. Appl., 72 (2017), pp. 306-316 (Open access journal) http://dx.doi.org/10.1016/j.eswa.2016.10.058)

[58] L. Breiman, Random forests, Mach. Learn., 45 (1) (2001), pp. 5-32
[59] J.M. Andrade-Garda, M. Gestal-Pose, F.A. Cedrón-Santaeufemia, J. Dorado-de-la-Calle, M.P. Gómez-Carracedo, Multivariate regression using artificial neural networks and support vector machines, J.M. Andrade-Garda (Ed.), Basic Chemometric Techniques in Atomic Spectroscopy (second ed.), RSC, London (2013)
[60] R.G. Brereton, G.R. Lloyd, Support vector machines for classification and regression, Analyst, 135 (2010), pp. 230-267
[61] H. Li, Y. Liang, Q. Xu, Support vector machines and its applications in chemistry, Chemom. Intell. Lab. Syst., 95 (2010), pp. 188-198
[62] J. Luts, F. Ojeda, R. van-de-Plas, B. de Moor, S. van-Huffel, J.A.K. Suykens, A tutorial on support vector machine-based methods for classification problems in chemometrics, Anal. Chim. Acta, 665 (2010), pp. 129-145
[63] M.P. Gómez-Carracedo, R. Fernández-Varela, D. Ballabio, J.M. Andrade, Screening oil spills by mid-IR spectroscopy and supervised pattern recognition techniques, Chemom. Intell. Lab. Syst., 114 (2012), pp. 132-142
[64] Y. Saeys, I. Inza, P. Larrañaga, A review of feature selection techniques in bioinformatics, Bioinformatics, 23 (19) (2007), pp. 2507-2517
[65] Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines, Mach. Learn., 46 (2002), pp. 389-422
[66] A. Liaw, M. Wiener, Classification and regression by randomForest, R. News, 2 (3) (2002), pp. 18-22
[67] A. Karatzoglou, A. Smola, K. Hornik, A. Zeileis, Kernlab – an S4 package for kernel methods in R, J. Stat. Softw., 11 (9) (2004), pp. 1-20
[68] M. Kuhn (contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang and Can Candan). Caret: Classification and Regression Training. R Package Version 6.0-68. https://CRAN.R-project.org/package=caret, 2016.
[69] R Core Team. R: a Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. Available at: https://www.R-project.org/ (last access 22.12.16).
[70] B. Mertens, M. Thompson, T. Fearn, Principal component outlier detection and SIMCA: a synthesis, Analyst, 119 (1994), pp. 2777-2784