Research Paper

Non-targeted Spatially Offset Raman Spectroscopy-based vanguard analytical method to authenticate spirits: White Tequilas as a case study

Christian Hazael PÉREZ-BELTRÁN✉ [a], Guadalupe PÉREZ–CABALLERO [b], José M. ANDRADE [c], Luis CUADROS-RODRÍGUEZ [a], Ana M. JIMÉNEZ-CARVELO✉ [a]

[a] Department of Analytical Chemistry, Faculty of Sciences, University of Granada, C/ Fuentenueva, s/n, E-18071 Granada (Spain).

[b] Laboratorio de Fisicoquímica Analítica y Especiación Química, Unidad de Investigación Multidisciplinaria. Facultad de Estudios Superiores Cuautitlán, Campo 4. Universidad Nacional Autónoma de México (México).

[c] Group of Applied Analytical Chemistry, University of A Coruña, Campus da Zapateira s/n, E-15071, A Coruña (Spain).

✉ Corresponding author (E-mail: christianpb@correo.ugr.es; Phone: +34 958 24 07 97)

✉ Corresponding author (E-mail: amariajc@ugr.es; Phone: +34 958 24 07 97)

Abstract

Adulteration and counterfeiting are ongoing problems for alcoholic drinks, including beers, wines, and spirits. To fight against them, official analytical methods need to be complemented with faster, trustworthy, non-invasive and *in-situ* ones, which have been named as vanguard methods, to increase the efficiency in the detection probability of truly adulterated alcoholic drinks. The analytical methodology proposed here synergistically combines a novel measurement analytical technique (spatially offset Raman spectroscopy, SORS) with chemometrics methods, i.e., principal component analysis (PCA), soft independent modeling of class analogies (SIMCA), partial least squares regression-discriminant analysis (PLS-DA), support vectors machine, (SVM) and quantitative partial least squares regression (PLSR). The applicability of the proposal is tested with Tequila to (i) differentiate among *100% agave* and *mixed* white packaged Tequilas, and (ii) to predict the alcoholic content. SORS spectra of 51 samples were obtained in the 300-2000 cm^{-1} range, from which classification and regression models were developed. The best classification performances were obtained with PLS-DA and SVM with 100% sensitivity, specificity and overall classification rate. PLSR exposed a better trend of the samples than PCA in the exploratory analysis; and yielded predictive models capable of foreseeing alcoholic contents with average errors lower than 4%. These results demonstrate the potential of this fast, *in-situ* analytical approach to be used as a vanguard analytical method to screen adulterated or counterfeited Tequilas and to assess the conformity of the alcoholic stated in the label.

Keywords

Chemometrics; Spirits fraud; Spatially offset Raman spectroscopy; Tequila authentication.

1. Introduction

Criminal activity against consumers continues non-stop, in fact European Union Intellectual Property Office (EUIPO) and European Union Agency for Law Enforcement Cooperation (EUROPOL) have indicated in a last report published in March 2022 that *the production of illicit food products, especially drinks, is increasingly professional and sophisticated* [**1**]. However, in terms of health and food safety, the weightiness of food and drink fraud will depend on the type of fraud. In some cases, the consequences are limited to consumer deception, since offenders pass off lower value products as higher value foods or drinks for illicit financial profit. Specifically in drinks the most frequent fraud is that committed in alcoholic beverages, so-called spirits. In fact, in the last two years, adulteration of this type of product has been detected, such as the case of the Whiskey fraud in Spain in 2020 [**2**] or the adulteration of alcoholic beverages in Santo Domingo in April 2022, which resulted in the death of several people [**3**].

There is a battery of recognized and well-described analytical methods for detecting different types of adulteration for each particular alcoholic beverage, most of them based on the identification and quantification of specific chemical markers. Despite traditional analytical methods proved to be reliable, accurate and are suitable tools for production control, they often do not comply with the principles of green chemistry, since they involve the use of environmentally unfriendly reagents, are time-consuming and frequently expensive, considering them as rearguard methods [**4**]. This gives opportunity for the development and application of alternative analytical methods, which are characterized by being miniaturized, transportable, simple, rapid, low-cost and capable of providing overall analytical information that is reliable and representative. The application of these type of alternative analytical methods, which have been named as vanguard methods, increase the efficiency of control laboratories since they make possible the analysis of only suspicious samples by rearguard methods [**4**]. The term vanguard method does not refer to the fact that the methodology presented in this study is highly recent and innovative, as might be inferred at first. It suggests that such a methodology could be applied as a first analytical approach to quickly process laboratory samples. In this sense, a vanguard method is often a forward screening method that allows the selection of suspect samples that will subsequently be subjected to a full backward analytical method, *i.e.*, a reaguard method.

In this sense, the use of non-targeted spectroscopic analytical techniques, such as conventional Raman or medium and near infrared spectroscopies, constitute established methodologies that fit most requirements to get vanguard analytical methods as they require minimum or null sample preparation. Despite of providing unspecific signals (spectroscopic instrumental fingerprints), they became popular to determine the composition/adulteration of food and beverages to ensure the authenticity and traceability [**5**]. One essential and inherent subsequent step after the application of spectroscopic techniques is the use of multivariate chemical data analysis or chemometrics, which together have created a synergistic and powerful analytical methodology that is regularly applied in the food industry to extract important and non-evident (or hidden) information from the raw spectra by developing mathematical models [**6**,**7**,**8**].

Quite recently, a new and more advanced Raman spectroscopy modality, termed spatially offset Raman spectroscopy (SORS), appeared and it shows highly promising capabilities for spirit quality and authenticity control. The fundamentals of SORS are like the conventional Raman spectroscopy, although in SORS the Raman signal is obtained at certain millimeters off the laser spot, making it possible to collect photons emitted from samples contained within opaque packaging materials [**9**]. This means that it is possible to carry out the analysis directly on the product within the container, without the need to alter the original package/sample, making SORS one of the few truly non-invasive analytical techniques. Even though this novel approach was first developed for the pharmaceutical industry, it expanded rapidly to the food industry to analyze packaged beverages in a fast and non-destructive manner [**9**]; for instance: Vodka, Gin and Whisky through their containers [**10**]. However, no applications have been found to authenticate Tequilas.

Tequila is a representative spirit from México that holds an Official Designation of Origin (DOT - *from the Spanish term 'Denominación de Origen Tequila'*), which is regulated by the Mexican Government and the Regulatory Council of Tequila (CRT) through the official Mexican standard NOM-006-SCFI-2012 [**11**]. Tequila can be classified in five classes according to their aging process in oak or holm oak containers: '*Silver or White'*, *'Aged'*, *'Extra-aged'* and *'Ultra-aged'* according to whether maturation lasts for <2 months, ≥2 months, ≥2 years or ≥3 years, respectively. *'Gold Tequila'* corresponds to commercial mixtures of White Tequila with Aged, Extra-aged or Ultra-aged Tequilas [**11**]. Additionally, two categories of Tequila can be distinguished: (i) 100% agave Tequila if only sugars from the juice of the *Agave Tequilana Weber blue variety* are used for the fermentation process, and (ii) 'mixed Tequilas' if any combination with other sources of reducing sugars (never more than 49%) are added to the process. The commonest commercial product is white Tequila, so, this paper focused on it.

Currently, many adulteration and counterfeiting cases are still reported, not only at Mexico but in other countries. The main adulteration practice is to substitute ethanol with methanol or, less frequently, with propanol, ethylene glycol, aldehydes and others [**12**]. In 2021, a production of 527 million of liters of Tequila was reported by the CRT whose quality and authenticity were evaluated using representative samples extracted from the distilleries and analyzed independently at the CRT. All the aforementioned classes of Tequila are inspected by the CRT using standardized analytical techniques, such as liquid and gas chromatography or atomic absorption spectroscopy, to adhere to current official analytical methods. Several quality parameters are determined, e.g., furfural, esters, aldehydes, methanol, higher alcohols, reducing and total sugars. An exemplary routine verification is whether the alcoholic content, using a digital densimeter method at 20ºC, which is established in the Mexican standard NMX-V-013-NORMEX-2019 [**13**], is between 35 and 55% (v/v).

The studies found in the literature concerning the assessment of tequila authenticity are focused on (i) some chemical markers, (ii) a specific spectral region of interest (ROI), or (iii) Red, Green and Blue (RGB) color coordinates obtained after the Tequila analysis by chromatographic and spectroscopic analytical techniques [**14**,**15**,**16**,**17**,**18**,**19**]. For example, Contreras et al. [**20**] applied UV-Vis spectroscopy to identify adulterated and fake Tequilas (between white and rested tequila) or Perez-Beltran et al. [**21**] employed FTIR and data fusion approach for distinguishing between pure and mixed White Tequilas. However, surprisingly no studies have been found where the full RAMAN spectrum is used as an unspecific instrumental fingerprint but characteristic of each tequila together with chemometric tools for tequila authentication.

In this regard, the innovation of this work lies in developing a fast and non-invasive vanguard analytical method for the *in-situ* screening quality control of spirits using SORS. Its applicability is demonstrated to ensure Tequila from Mexico in the following terms: (i) discriminate White Tequilas (100% agave vs mixed), and to (ii) predict and verify the alcoholic content. For this, SORS spectra were used together chemometric tools to develop suitable classification and quantitation multivariate analytical methods. Classification methods were validated in terms of sensitivity, specificity, precision, negative prediction value, among other 21 classification performance metrics and estimated following the study published by Cuadros-Rodríguez et al. (2016) [**22**]. In addition, the quantitative method for determining the alcohol content was validated according to the ASTM E2617 standard [**23**].

2. Materials and methods

2.1. Tequila samples

A total of 51 White Tequila samples were provided by the CRT in México, and analyzed in Spain, as described in the 'spatially offset Raman spectroscopy (SORS) measurements' section. Thirty White Tequilas belonged to the 100% agave White Tequila category (TB - from the Spanish term 'Tequila Blanco') and twenty-one to the mixed White Tequilas (TBM - from the Spanish term 'Tequila Blanco Mixto'). The alcoholic content of all these samples was determined by the CRT using a digital densimeter at 20ºC [**13**].

2.2. Spatially offset Raman spectroscopy (SORS) measurements

Vaya Raman SORS equipment (Agilent Technologies, Santa Clara, CA, USA) was used. The excitation radiation was 830 nm with a maximum power laser of 450 mW, obtaining Raman spectra in the low frequencies range, from 350 to 2000 cm^{-1}, with 12 to 20 cm^{-1} spectroscopic resolution. The SORS measurements of the 51 white Tequila samples were performed directly through amber vials lasting 30 s, approximately.

2.3. Similarity analyses

In order to make sure that this methodology can be transferable to any other situation, similarity analyses were performed. SORS measurements were directly performed on four original bottled Tequilas marketed in Spain (2 mixed White Tequilas, 1 mixed Rested Tequila and 1 mixed Tamarind flavored White Tequila). Afterwards, 2 mL of each of them were transferred to amber glass vials, similar to those used to transport the Mexican Tequila samples, and measured. Once both spectra for each sample were acquired, the similarity among them was assessed calculating the corresponding nearness similarity index [**24**], which is based on the proximity of two vectors in space and is calculated from the standardized Euclidean distance, as depicted in Eq. (1).

$$\mathrm{NEAR}(\mathrm{X_{SORS}}, \mathrm{X_{CRS}}) = 1 - \left[\sqrt{\frac{(\mathrm{X_{SORS}} - \mathrm{X_{CRS}})\ \mathrm{x}\ (\mathrm{X_{SORS}} - \mathrm{X_{CRS}})^{\mathrm{T}}}{(\mathrm{X_{SORS}} + \mathrm{X_{CRS}})\ \mathrm{x}\ (\mathrm{X_{SORS}} + \mathrm{X_{CRS}})^{\mathrm{T}}}}\right] \qquad (1)$$

where $\mathrm{X_{SORS}}$ and $\mathrm{X_{CRS}}$ symbolized both SORS and conventional Raman spectra, respectively, and the superscript T denotes the transposed matrix [**25**].

2.4. Multivariate data analyses

SORS raw data were exported from CSV format (comma-separated values) to MATLAB environment (Mathworks, Massachusetts, USA, v. R2013b). The exported spectra contained 1651 variables, each. The training set was constituted by 41 samples (24 of TB type and 17 of TBM type) whilst the external validation set contained 10 different samples (6 TB and 4 TBM). Splitting was performed applying the Kennard-Stone selection method (so-called CADEX algorithm), which was deployed on the TB and TBM classes independently in order to select the samples of the validation set.

The multivariate data analyses were carried out using the PLS_Toolbox software (v. 8.6.1, 2019, Eigenvector Research In., Manson, WA, USA). The applied chemometric tools were principal component analysis (PCA) and partial least squares regression (PLSR) for exploratory analysis, soft independent modeling of class analogy (SIMCA), partial least squares-discriminant analysis (PLS-DA) and support vector machines (SVM) for classification, and PLSR was also used to quantify the alcoholic content of the samples. Mean centering and smoothing were used as pre-processing techniques depending on the multivariate method, as described in 'exploratory analyses' and 'classification analyses'. The proper number selection of the PCs and LVs of the models was based on the study of their root mean square error for calibration (RMSEC), or for prediction (RMSEP) and for cross-validation (RMSECV) plots, and the total explained variance, avoiding overfitting in each case.

3. Results and discussion

3.1. SORS analyses and characterization

When SORS analyses are performed, two measurements are acquired: one at zero offset and another one with a laterally spatial offset of 0.7 mm from the point of incidence of the laser to the collection point [**9**]. This separation favors the photons from the lower layers to be radiated from a spot laterally shifted from the incidence zone while the photons on the upper package are radiated from the same incidence zone [**26**]. Afterwards, internal pre-processing and normalization are performed by the equipment and a final Raman spectrum is obtained with no contribution of the container. The Raman spectra of the two categories of white Tequilas can be observed in Fig. 1.

Fig. 1

The intense peak located at 882 cm^{-1} and the peak at 1053 cm^{-1} are attributed to the stretching and deformation modes of the skeletal C-C-O moieties, whilst the peak at 1090 cm^{-1} is associated to the stretching mode of the C-O bond. The peaks at 1279 cm^{-1} and 1455 cm^{-1} are assigned to the deformation wagging mode and to the wagging mode of CH_2, respectively [**15**,**27**]. Additionally, the two small peaks around 1610 cm^{-1} and 1728 cm^{-1} are associated to the cyclic ketone structure, which is the basis of furanic compounds in Tequila. Noteworthy, those peaks are more intense for the TB category than for the TBM one, as TB proceeds only from fermentable sugars of the Agave Tequilana Weber blue variety (through the Maillard reaction [**28**] when cooked). On the contrary, TBM might or might not present these spectral Raman peaks because this category of Tequilas can be produced from mixtures of fermentable sugars, so that the production of furanic compounds might not occur [**29**].

These acquired signals (Raman spectra), which are here used to evaluate the authenticity and quality of White Tequilas, are non-specific instrumental fingerprints and make it necessary the application of multivariate data analyses, as described in the following subsections.

3.2. SORS and conventional Raman spectra similarity analyses

A point-by-point comparison, using the nearness similarity index (NEAR), among the four pairs of spectra (data vectors) corresponding to the Tequila samples marketed in Spain was performed to assess their similarity when the spectroscopic measurements are performed through the original Tequila glass bottle or through amber glass vials (used as reference). The expected NEAR results of the standardized Euclidean distance range from 0 to 1, being 1 the maximum similarity among the spectra. Fig. 2 displays the spectra of the four analyzed samples within their original glass bottles and the spectra of the samples transferred to the vial.

Fig. 2

As it can be observed in Fig. 2, each pair of overlapping spectra are similar at first glance and this fact is further confirmed when the Nearness similarity index is calculated, obtaining NEAR values >0.92, which indicates that both spectra are largely similar with almost null influence of the original glass bottles over the measurements (the remaining ca. 0.08% can be considered as random noise). According to these results, it is evident that the methodology presented here has potential application to the *in-situ* quality control and authentication analysis of Tequila.

3.3. Exploratory analyses

An exploratory analysis was performed to screen the natural grouping of the 51 Tequilas. For this study, the spectral data was previously mean centered. First, a PCA was built considering 5 principal components (PCs) and explaining 75.9% of the cumulative variance, whose main scores plot, is displayed in Fig. 3. Nonetheless, it can be observed that the samples do not follow any specific trend among categories.

Fig. 3

Furthermore, PLSR was used to explore these samples. The model was built with 5 latent variables (LV) explaining 71.1% of the cumulative variance in the X block and 85.8% in the Y block. Fig. 4 shows the LV2 vs LV3 scores plot, where the TB category concentrates (although not unequivocally) in the upper-right region of the plot and the TBM category to the left. The different results among PCA and PLSR lies basically in the very nature of the PLS latent variables that capture both variance and correlation [**30**], yielding best results when PLSR is applied, as it was also found when looking for groups among FTIR fused data of 100% agave and mixed White Tequilas [**21**]. Additionally, there are some samples placed out of the 95% confidence limit that might be considered as outliers (see Figures 3 and 4), however, it was noticed through the normalized (or reduced) Hotelling T^2-leverages *vs.* Q residuals plot that those samples had a normal behavior, discarding the existence of outliers. Thus, all samples were included in the following data analyses.

Fig. 4

3.4. Classification analyses

The next step after the exploratory analysis was the development of non-targeted multivariate analytical methods to discriminate among TB and TBM. For all classification models, mean centering and smoothing (Savitski-Golay, 15 points for filter width and 1st order polynomial) were used as preprocessing techniques. Smoothing is a low-pass filter that removes high-frequency noise [**30**]. The target class is TB as it is the category with more probability to be adulterated due to its economic profit. The results of the final classification models are discussed next.

- One Class-SIMCA

The developed SIMCA models were generated using two strategies: (i) two input-class classification (2iC-SIMCA) models, in which the model is trained using two classes (TB and TBM), and (ii) one input-class classification (1iC-SIMCA) model, in which the model is trained only with the 'target class' (TB). Within the 1iC-SIMCA strategy, two options were evaluated: (a) using the aforementioned calibration and validation data sets and (b) augmenting the validation set using all the 21 TBM and the previous 6 TB samples. It was found that the 1iC-SIMCA approach presented the best results using 5 PCs.

The 1iC-SIMCA classification plot (Fig. 5a) depicts the normalized (or reduced) Hotteling's T^2 and Q statistics of the target class, at a 95% confidence level. Samples from the validation set with normalized T^2 and Q values < 1 (left-bottom quadrant) are those considered as the target class (TB), whereas samples with T^2 and Q values > 1 (right-bottom quadrant) are considered as non-TB (or TBM). In this sense, samples TBM13 and TBM102 are misclassified as TB and sample TB70 as TBM, indicating that further confirmatory analyses should be performed. These results are used to create the corresponding validation contingencies of the classification model, as shown in Fig. 5b.

Fig. 5

- PLS-DA

The PLS-DA model was built using 4 latent variables, which explained 78.3% and 44.1% of the cumulative variance of both X- and Y-variable blocks, respectively. A threshold value of 0.5 was established as a decision criterion for the classification of the samples; scores (weights) >0.5 correspond to TB and <0.5 to TBM, as can be observed in the classification plot represented by Fig. 6a. The validation contingencies of the PLS-DA classification model are shown in Fig. 6b. Note that all validation samples were correctly classified, even though some samples from the training set were misclassified. This demonstrates the powerful generalization capabilities of the PLS-DA model.

Fig. 6

- SVM

Support vectors machine (SVM) was performed using the radial basis function (RBF) kernel algorithm with the gamma and cost values studied in the 10^{-6}-10 and 10^{-3}-10^{2}

ranges, respectively, and PLS compression with 4 LVs. The classification results for both the training and validation samples are displayed in Fig. 7a. The results are almost the same as the PLS-DA ones, suggesting that sample TB70 should undergo further confirmatory analyses, since it is very close to the threshold value. The SVM validation contingencies are displayed in Fig. 7b.

Fig. 7

As a matter of comparison, the classification performance metrics for the classification models were calculated from the results of the validation contingencies (see Table 1) [**22**], considering TB as the target class. The most popular metrics are discussed here; however, the detailed explanation of each of them is out of the scope of this work and interested readers are kindly forwarded to ref [**22**] for specific details on this topic.

Table 1

In principle, satisfactory classifications lead to classification performance metrics close to 1 and bad models to 0. For instance, Table 1 shows that PLS-DA and SVM models have a sensitivity (SENS) = 1, whilst 1iC-SIMCA a and b yields SENS = 0.83, which indicates that PLS-DA and SVM models classify better the TB samples than 1iC-SIMCA. Specificity (SPEC) indicates that the TBM samples are correctly classified, being better for PLS-DA and SVM models with a value = 1 than for 1iC-SIMCA a and b with SPEC = 0.50 and 0.33, respectively. In fact, the 1iC-SIMCA b model, validated with all the TBM samples, provided worse classification results than 1iC-SIMCA a, validated with fewer TBM samples.

Additionally, the positive predictive value (PPV) (so-called precision) informs on the proportion of agreements in relation to all assigned values of TB class whilst the negative prediction value (NPV) takes into account the ratio between agreements and the total number of TBM samples. For PLS-DA and SVM those metrics were = 1, whereas for the 1iC-SIMCA a and b models PPV were = 0.71 and 0.26, and NPV = 0.67 and 0.88, respectively. The overall classification rate (OCR) was 100%, 100% and 83% for PLS-DA, SVM and 1iC-SIMCA, respectively, and the Matthews correlation coefficient (MCC) –which might be considered a compendium of the overall classification ability of the models– was 1.0, 1.0 and 0.36 for the same classification models.

When the validation set 'a' is applied on the 1iC-SIMCA model, the validation results are relatively good; however, the results are fictitious as this set does not represent the reality of the sample population. The good results are due to the fact that in the validation set 'a' only 4 TBM samples (non-target class) are considered, but when the number of TBM samples is increased (validation set 'b'), the model does not classify well. That is, the model classifies almost all TBM samples as belonging to the TB class, which is related to the results shown in the exploratory analysis and the no clustering tendency of the classes, so it is not possible to establish regions for each of them. Therefore, the SIMCA class modelling method is not suitable for the purpose of this study.

The classification ability of the models obtained in this study (PLS-DA and SVM models) are better than others previously reported for different purposes (despite a direct, straightforward comparison is not possible) applying PCA-linear discriminant analysis (LDA), with an overall classification rate (OCR) of 90.02%, SENS = 0.90 and SPEC = 0.96 [**17**]. Furthermore, in a previous study [**18**] in which nine models were built using mean-centered UV-Vis spectroscopic data to differentiate various classes of Tequila, it was found that nonlinear models behaved better than linear ones (EFFIC > 0.94).

In this context, it is worth noting that class modeling methods, such as 1iC-SIMCA, are particularly suitable for real-world authentication problems where the target class is always defined from the authentic or genuine product and is modeled with a large number of samples, since it is less common to find adulterated samples. This approach has a great potential when the ideal scenario with sufficient number of authentic samples (target class) are available, being capable to properly identify new samples obtained from non-authentic products and differentiate them from those specimens of genuine ones. However, for this particular study, the available samples to build a more reliable 1iC-SIMCA model were limited, since Tequila Blanco 100% agave is only produced in certain regions of México and the accessibility of a variety of samples is rather narrow. A good alternative to address this situation is the use of discriminant methods, such as PLS-DA and SVM, particularly in this study, because it aimed at classifying two mutually excluding classes ('100 agave' and 'mixto') of the same quality sort of tequila ('Tequila Blanco'). In fact, it was evidenced that the validation results of the 1iC-SIMCA model depend on the number and type of samples included in the test set, but PLS-DA and SVM models provided better ability to correctly classify samples from both classes. However, this discriminant strategy is not free from the drawback of misclassifying new samples coming from non-genuine products with some different composition from those already used in the training step, which is a risk that practitioners must evaluate and take into account when extending the application of the method.

3.5. Alcoholic content quantitation

A PLSR-based quantitation analytical method was calibrated to predict the alcoholic content of the Tequila samples. As detailed above, the reference values were obtained by the CRT following the official method. The PLSR model was built using mean centering to preprocess the spectra and including 5 LVs in the model which explained 73.6 and 97.1% of the cumulative variance for the X- and Y-variable blocks, respectively. Fig. 8 compares the PLSR predicted alcoholic contents against the total alcoholic content reported by the CRT. The evaluation of this model was performed with the quantitation performance metrics, as observed in Table 2.

Fig. 8

Table 2

The first quantitation performance metric is the coefficient of determination (R^2) with a value = 0.971, evidencing a good fitting. The following four metrics are related to different sorts of errors the model might present (root mean square error, mean absolute error, median absolute error and standard error of validation), all of them with values less than 4%; the sixth metric is the standard deviation of validation residuals (SDV = 2.7%), indicating that the agreement of the predictions of the empirical model with the reference values is high, which results in a quite good predictive ability.

Note that PLSR has been previously applied to predict the alcoholic content of different Tequilas using FTIR, obtaining very good results [**19**]. Moreover, a vector network analyzer with an open-ended coaxial probe kit was used for the same purpose [**31**].

PLSR has also been applied to quantitate the furfural, 2-acetylfuran and 5-methylfurfural content in White Tequilas and Mezcals samples with acceptable results [**29**]. It would have been interesting to compare the results obtained here with those of another report in which SORS was applied to study the adulteration of Vodka, Gin and Whisky with methanol, but prediction of the alcoholic content was not considered [**10**].

4. Conclusions

Economic losses for the industry of alcoholic beverages and societal health problems are two relevant consequences of the adulteration and counterfeiting of commercialized spirits, which have not ceased over the years. To streamline the authentication surveillance of these products, current official rearguard methods need to be complemented with vanguard, faster and reliable *in-situ* screening analytical methods. In this regard, the present study reports for the first time the combination of the SORS analytical technique and chemometrics to discriminate between 100% agave and mixed White Tequilas and to predict their alcoholic content. It should be noted that the potential of the *in-situ* non-invasive SORS measurement implemented here has been verified by means of a similarity analysis. This demonstrated that the spectra obtained after analyzing Tequilas through the original bottle and through amber vials are almost the same, obtaining nearness indexes close to 1. Afterwards, models were developed and assessed with several classification performance metrics, which indicated that satisfactory classifications and predictions were achieved. PLS-DA and SVM presented the best OCR = 100%, evidencing that the combination of SORS and some chemometric methods is able to discern among 100% agave and mixed White Tequilas. Finally, a PLSR quantitation model demonstrated an excellent ability to predict the alcoholic content of the samples.

The approach presented here offers an alternative analytical method for routine authentication tasks undergone by official regulatory bodies. It is reliable and fast for *in-situ* screening purposes and, can complement and accelerate the quality control and authentication processes of commercial spirits, such as Tequila.

Conflicts of interest

The authors declare that they have no conflict of interest.

Acknowledgements

The authors are deeply grateful to the Mexican 'Consejo Regulador del Tequila' (CRT) for providing the samples for this study, as well as to the 'Sánchez Baldiviezo' Association for facilitating the samples for the similarity study. One author, C.H. Pérez-Beltrán acknowledges also the scholarships from the Autonomous University of Sinaloa (México) and from the 'Asociación Universitaria Iberoamericana de Posgrado' (AUIP) and the 'Consejería de Transformación Económica, Industria, Conocimiento y Universidades' of the Regional Government of Andalusia (Spain) for a research stay.

References *(warning: section change – don't remove this line)*

[1] C. de Bolle, C. Archambeau. Intellectual property crime. Threat assessment 2022. EUIPO. (2022). https://doi.org/10.2814/830719.

[2] Fourteen arrested and 300,000 bottles of counterfeit whisky seized. (2020). https://www.elconfidencial.com/espana/2020-12-10/catorce-detenidos-intervenidas-300-000-botellas-whisky-falso-guardia-civil_2866480/ (*Spanish version*). Accessed 27 July 2022.

[3] Six people have died as a result of adulterated alcohol. (2022). https://www.elcaribe.com.do/destacado/seis-personas-han-fallecido-a-causa-del-alcohol-adulterado-en-2022/ (*Spanish version*). Accessed 27 July 2022.

[4] M. Valcárcel, S. Cárdenas, Vanguard-rearguard analytical strategies, Trends. Anal. Chem. 24 (2005) 67-74. https://doi.org/10.1016/j.trac.2004.07.016.

[5] A.M. Jiménez Carvelo, S. Martin Torres, L. Cuadros Rodríguez, A. González Casado, Food Authentication and Traceability, in: C.M. Galanakis (Ed.), Nontargeted fingerprinting approaches, Academic Press, 2021, pp. 163-194. https://doi.org/10.1016/B978-0-12-821104-5.00010-6.

[6] P. Oliveri, C. Malegori, E. Mustorgi, M. Casale, Comprehensive Chemometrics – Chemical and Biochemical Data Analysis, in: S. Brown, R. Tauler, B. Walczak (Eds.), Application of chemometrics in the food sciences, Elsevier, 2020, pp. 99-111. https://doi.org/10.1016/B978-0-12-409547-2.14748-1.

[7] A.M. Jiménez Carvelo, L. Cuadros Rodríguez, Data mining/machine learning methods in foodomics, Curr. Opin. Food Sci. 37 (2021) 76-82. https://doi.org/10.1016/j.cofs.2020.09.008.

[8] A.M. Jiménez Carvelo, A. González Casado, M.A. Bagur González, L. Cuadros Rodríguez, Alternative data mining/machine learning methods for the analytical evaluation of food quality and authenticity – A review, Food Res. Int. 122 (2019) 25-39. https://doi.org/10.1016/j.foodres.2019.03.063.

[9] A. Arroyo Cerezo, A.M. Jiménez Carvelo, A. González Casado, A. Koidis, L. Cuadros Rodríguez, Deep (offset) non-invasive Raman spectroscopy for the evaluation of food and beverages – A review, LWT-Food Sci. Technol. 149 (2021) 111822. https://doi.org/10.1016/j.lwt.2021.111822.

[10] D.I. Ellis, R. Ecccles, Y. Xu, J. Griffen, H. Muhamadali, P. Matousek, I. Goodall, R. Goodacre, Through-container, extremely low concentration detection of multiple chemical markers of counterfeit alcohol using a handheld SORS device, Sci. Rep. 7 (2017) 12082. https://doi.org/10.1038/s41598-017-12263-0.

[11] Mexican Official Standard NOM-006-SCFI-2012, Alcoholic Beverages -Tequila- Specifications, National Advisory Committee on Standardization, User Safety, Commercial Information and Trade Practices (CCNNSUICPC), Mexican Government. https://www.crt.org.mx/images/documentos/Normas/NOM_006_SCFI_2012_Ingles.pdf (accessed 13 June 2022).

[12] D.G. Barceloux, R. Bond, E.P. Krenzelok, H. Cooper, J.A. Vale, American academy of clinical toxicology practice guidelines on the treatment of methanol poising, J. Toxicol. Clin. Toxicol. 40 (2002) 415-446. https://doi.org/10.1081/CLT-120006745.

[13] Mexican Standard NMX-V-013-NORMEX-2019, Bebidas alcohólicas-determinación del contenido alcohólico (por ciento de alcohol en volumen a 20ºC) (% Alc. Vol.) - Métodos de ensayo (prueba) (*in Spanish*). National Advisory Committee on Standardization of the Economy Secretariat (CCONNSE), Mexican Government.

[14] L.I. Espinosa Vega, A. Belio Manzano, C.A. Mercado Ornelas, I.E. Cortes Mestizo, V.H. Méndez García. Aging spectral markers of tequila observed by Raman spectroscopy, Eur, Food Res. Technol. 245 (2019) 1031-1036. https://doi.org/10.1007/s00217-018-3203-4.

[15] C. 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 Mol. Biomol. Spectrosc. 61 (2005) 2657-2662. https://doi.org/10.1016/j.saa.2004.10.008.

[16] C. Fernández Lozano, M. Gestal Pose, G. Pérez Caballero, A.L. Revilla Vázquez, J.M. Andrade Garda, Quality Control in the Beverage Industry, in: A. Grumezescu, A.M. Holban (Eds.), Multivariate classification techniques to authenticate Mexican commercial spirits, Academic Press, 2019, pp. 259-287. https://doi.org/10.1016/B978-0-12-816681-9.00008-4.

[17] A. Gómez, D. Bueno, J.M. Gutiérrez, Electronic eye based on RGB analysis for the identification of tequilas, Biosensors 11 (2021) 68-83. https://doi.org/10.3390/bios11030068.

[18] G. Pérez-Caballero, J.M. Andrade, P. Olmos, Y. Molina, I. Jiménez, J.J. Durán, C. Fernández-Lozano, F. Miguel-Cruz, Authentication of tequilas using pattern recognition and supervised classification, Trends Anal. Chem. 94 (2017) 117-129. https://doi.org/10.1016/j.trac.2017.07.008.

[19] 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. 53 (2005) 2151-2157. https://doi.org/10.1021/jf048637f.

[20] U. Contreras, O. Barbosa García, J.L. Pichardo Molina, G. Ramos Ortíz, 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) 2356-2362. https://doi.org/10.1016/j.foodres.2010.09.001.

[21] C.H. Pérez Beltrán, V.M. Zuñiga Arroyo, J.M. Andrade, L. Cuadros Rodríguez, G. Pérez Caballero, A.M. Jiménez Carvelo, A sensor-based methodology to differentiate pure and mixed white tequilas based on fused infrared spectra and multivariate data treatment, Chemosensors 9 (2021) 47-59. https://doi.org/10.3390/chemosensors9030047.

[22] L. Cuadros Rodríguez, E. Pérez Castaño, C. Ruiz Samblás, Quality performance metrics in multivariate classification methods for qualitative analysis, Trends Anal. Chem. 80 (2016) 612-624. https://doi.org/10.1016/j.trac.2016.04.021.

[23] ASTM E2617-17. Standard practice for validation of empirically derived multivariate calibrations, ASTM International, 2017.

[24] R. Pérez Robles, N. Navas, S. Medina Rodríguez, L. Cuadros Rodríguez, Method for the comparison of complex matrix assisted laser desorption ionization-time of flight mass spectra. Stability of therapeutical monoclonal antibodies, Chemometr. Intell. Lab. Syst. 170 (2017) 58-67. https://doi.org/10.1016/j.chemolab.2017.09.008.

[25] F. Stilo, A.M. Jiménez Carvelo, E. Liberto, C. Bicchi, S.E. Reichenbach, L. Cuadros Rodríguez, C. Cordero, Chromatographic fingerprinting enables effective discrimination and identitation of high-quality Italian Extra-virgin olive oils, J. Agric. Food Chem. 69 (2021) 8874-8889. https://doi.org/10.1021/acs.jafc.1c02981.

[26] P. Matousek, I.P. Clark, E.R.C. Draper, M.D. Morris, A.E. Goodship, N. Everall, M. Towrie, W.F. Finney, A.W. Parker. Subsurface probing in diffusely scattering media using spatially offset Raman spectroscopy, Appl. Spectrosc. 59 (2005) 393-400. https://doi.org/10.1366/0003702053641450.

[27] F. Li, Z. Men, S. Li, S. Wang, Z. Li, C. Sun, Study of hydrogen bonding in ethanol-water binary solutions by Raman spectroscopy, Spectrochim. Acta A Mol. Biomol. Spectrosc. 189 (2018) 621-624. https://doi.org/10.1016/j.saa.2017.08.077.

[28] 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 (2002) 806-812. https://doi.org/10.1021/jf0110295

[29] 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 (2010) 1077-1087. https://doi.org/10.1590/S0103-50532010000600018.

[30] B.M. Wise, N.B. Gallagher, R. Bro, J.M. Shaver, W. Winding, R.S. Koch, Chemometrics Tutorial for PLS_Toolbox and Solo, Eigenvector Research, Inc. Wenatchee, WA, USA, 2006.

[31] T.K. Kataria, M.E. Sosa Morales, J.L. Olvera Cervantes, A. Corona Chavez, Dielectric properties of tequila in the microwave frequency range (0.5-20 GHz) using coaxial probe, Int. J. Food Prop. 20 (2017) S377-S384. https://doi.org/10.1080/10942912.2017.1297949.

Table 1. Summary of classification performance metrics for 1iC-SIMCA, PLS-DA and SVM models.

	1iC-SIMCA		**PLS-DA**	**SVM**
	a	**b**		
Metrics	*Target class (100% agave White Tequila, TB)*			
Sensitivity (SENS)	0.83	0.83	1.00	1.00
Specificity (SPEC)	0.50	0.33	1.00	1.00
False positive rate (FPR)	0.50	0.67	0.00	0.00
False negative rate (FNR)	0.17	0.17	0.00	0.00
Positive predictive value (PPV) (precision)	0.71	0.26	1.00	1.00
Negative predictive value (NPV)	0.67	0.88	1.00	1.00
Youden index (YOUD)	0.33	0.17	1.00	1.00
Positive likelihood rate (LR(+))	1.67	1.25	–	–
Negative likelihood rate (LR(-))	0.33	0.50	0.00	0.00
Classification odds ratio (COR)	5.00	2.50	–	–
F-measure (F)	0.77	0.40	1.00	1.00
Discriminant power (DP)	0.39	0.22	–	–
Efficiency (or accuracy) (EFFIC)	0.70	0.44	1.00	1.00
Misclassification rate (MR)	0.30	0.56	0.00	0.00
AUC (correctly classified rate) (CCR)	0.67	0.58	1.00	1.00
Gini coefficient (Gini)	0.33	0.17	1.00	1.00
G-mean (GM)	0.65	0.53	1.00	1.00
Matthews' correlation coefficient (MCC)	0.36	0.15	1.00	1.00
Chance agreement rate (CAR)	0.54	0.39	0.52	0.52
Chance error rate (CER)	0.48	0.35	0.48	0.48
Kappa coefficient (KAPPA)	0.35	0.09	1.00	1.00
PROB (TB/TB)	0.71	0.26	1.00	1.00
PROB (nTB/nTB)	0.67	0.88	1.00	1.00
PROB (TB/nTB)	0.33	0.13	0.00	0.00
PROB (nTB/TB)	0.29	0.74	0.00	0.00

The hyphen "–" signifies that the performance feature cannot be determined since it involves a division between zero.

a and b: models validated using 10 (6 TB and 4 TBM) and 27 (6 TB and 21 TBM) samples as external validation sets, respectively.

Table 2. Performance metrics in the quantitation of the alcoholic content of the Tequila samples that constitute the validation set.

Metrics	**Value (%)**
Coefficient of determination (R^2)	0.971
Root mean square error (RMSE)	3.32
Mean absolute error (MAE)	1.82
Median absolute error (MdAE)	2.61
Standard error of validation (SEV)	3.14
Standard deviation of validation residuals (SDV)	2.65

Figure legends

Figure 1. Raman spectra of a '100% agave' White Tequila sample (TB) and a 'mixed' White Tequila (TBM) one.

Figure 2. Similarity plots of four sample pairs of White Tequila (S1-S4) measured through the original bottle (BS) and amber vial (VS), considered as the reference spectrum.

Figure 3. Exploratory PC1 vs PC2 scores plot from the 51 samples PCA model showing two different categories of White Tequilas. TB: 100% agave White Tequila (n=30) and TBM: mixed White Tequilas (n=21).

Figure 4. Exploratory LV2 vs LV3 scores plot from the 51 samples PLS model showing two different categories of White Tequilas. TB: 100% agave White Tequila (n=30) and TBM: mixed White Tequilas (n=21).

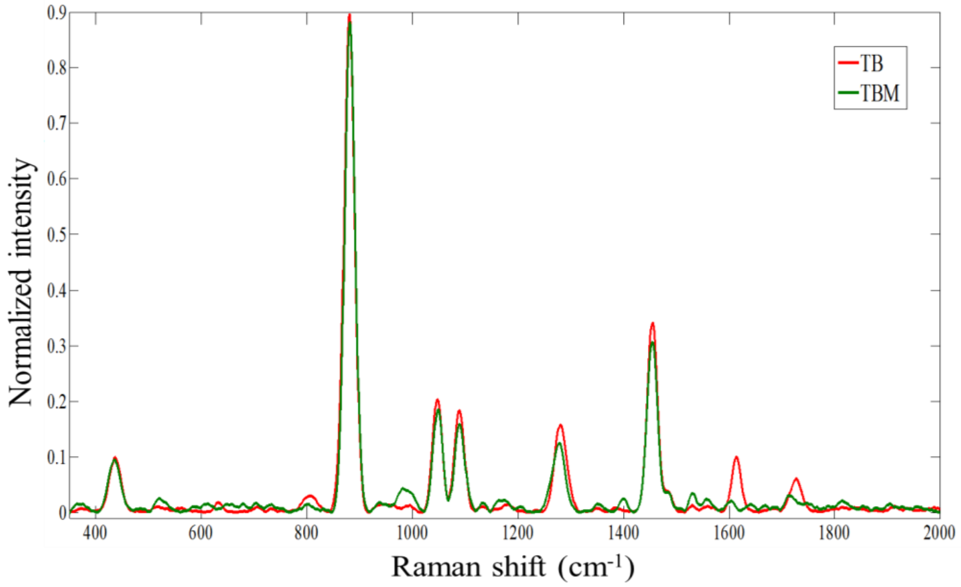
Figure 5. (a) Classification plot (a) and (b) validation contingencies for the one input-class SIMCA classification model. Class 1: target class (TB: '100% agave' Tequila); class 2: non-target class (TBM: 'mixed Tequila') (The magenta-marked samples in figure 5a are the misclassified samples).

Figure 6. (a) Classification plot and (b) validation contingencies for the PLS-DA classification model. Class 1: target class (TB: '100% agave' Tequila); class 2: non-target class (TBM: 'mixed Tequila'). (The dashed line in figure 6a indicates the 0.5 threshold level).

Figure 7. (a) Classification plot and (b) validation contingencies for the SVM classification model. Class 1: target class (TB: '100% agave' Tequila); class 2: non-target class (TBM: 'mixed Tequila'). (The dashed line in figure 7a marks the 0.5 threshold level).

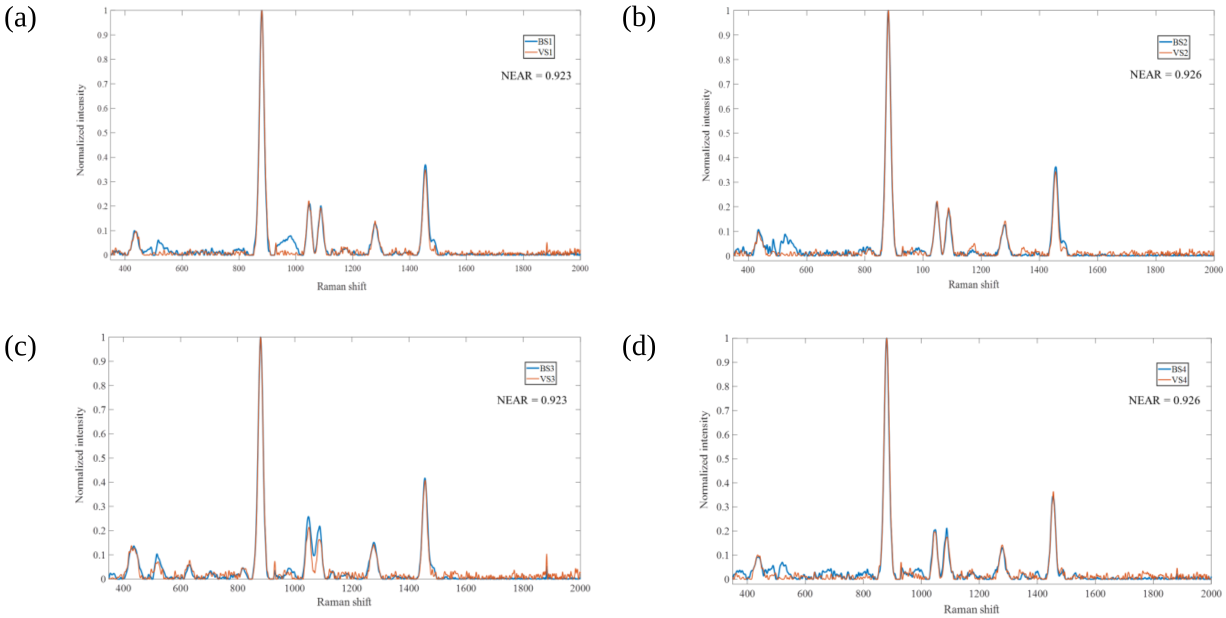
Figure 8. PLSR alcoholic predictions (% v/v) for White Tequila samples. (a) Calibration curve, and (b) alcoholic content plot of the validation set samples. The circles are colored according to the predicted alcoholic content from the vertical color scale. Each sample displays the predicted value against the real value of alcoholic content, which is underlined.

<Figure 1>

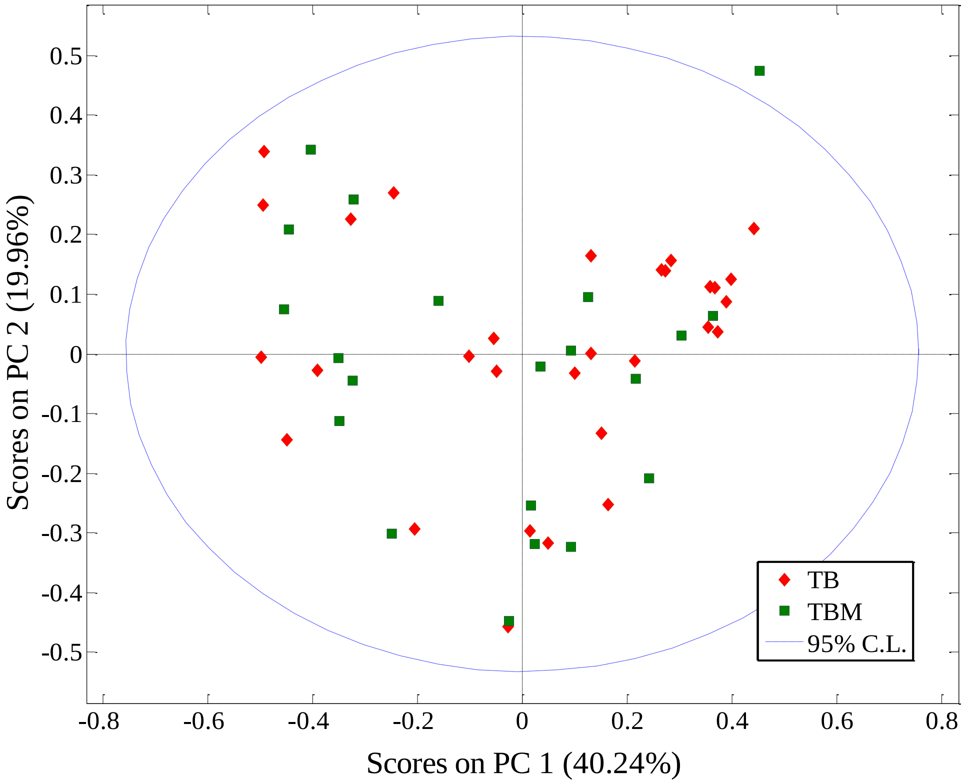


TB
TBM
Normalized intensity
Raman shift (cm-1)
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
400
600
800
1000
1200
1400
1600
1800
2000

<Figure 2>

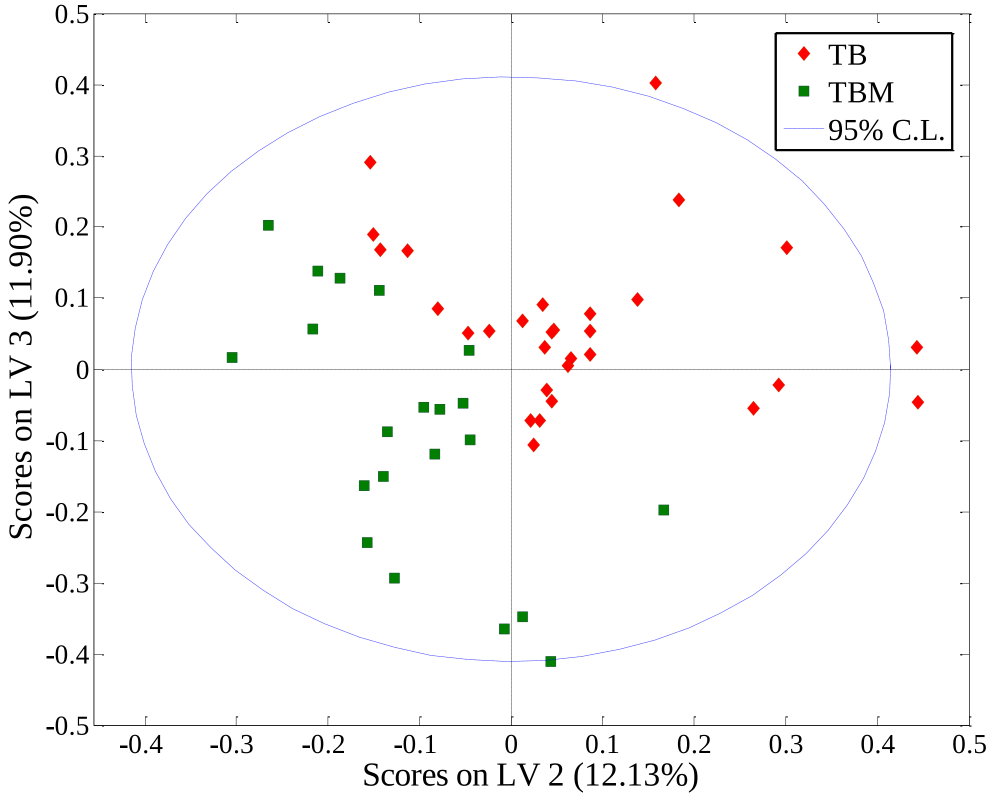

(a)
BS1
VS1
NEAR = 0.923
Normalized intensity
Raman shift
(b)
BS2
VS2
NEAR = 0.926
Normalized intensity
Raman shift
(c)
BS3
VS3
NEAR = 0.923
Normalized intensity
Raman shift
(d)
BS4
VS4
NEAR = 0.926
Normalized intensity
Raman shift

<Figure 3>



v2.4-clean

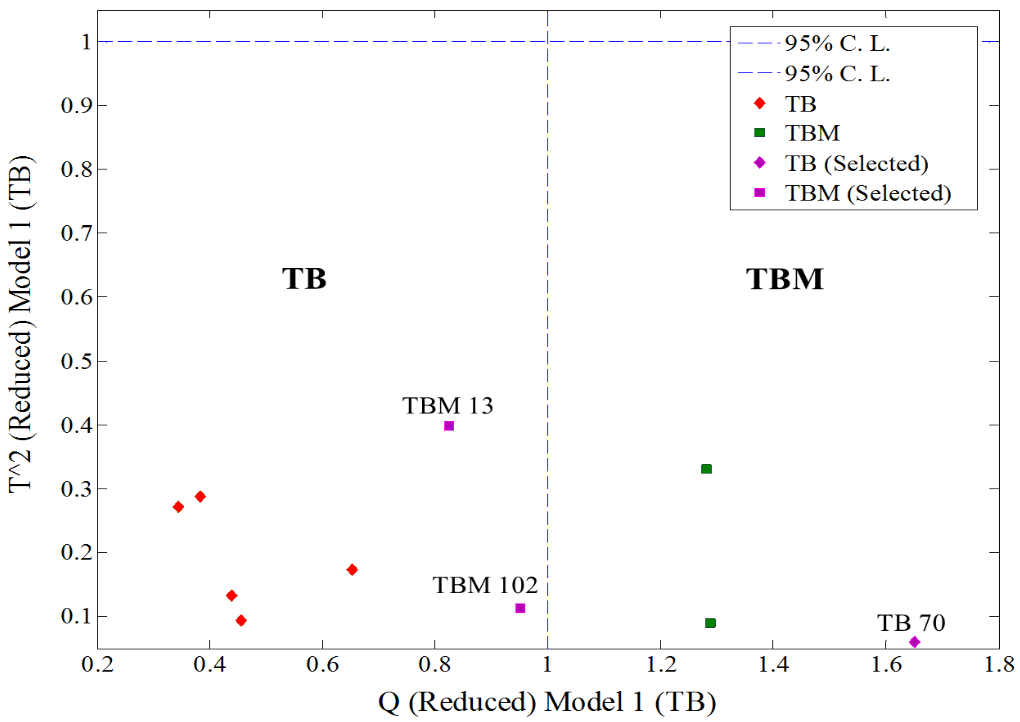
<Figure 4>



TB
TBM
95% C.L.
Scores on LV 3 (11.90%)
Scores on LV 2 (12.13%)
0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
-0.5
-0.4
-0.3
-0.2
-0.1
0
0.1
0.2
0.3
0.4
0.5

<Figure 5>

(a)



(b)

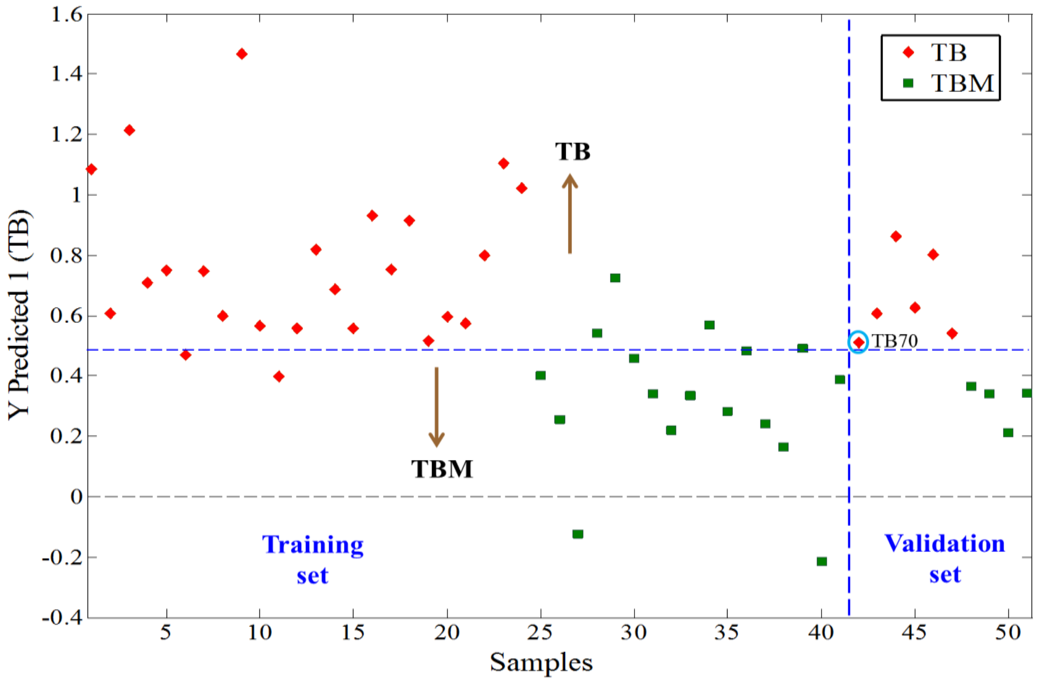
Assignation	Class 1	Class 2	
	6	**4**	**10**
Inconclusive (I)	0	0	**0**
Class 2 (TBM)	1 (8.33%)	2 (25%)	**3**
Class 1 (TB)	5 (41.67%)	2 (25%)	**7**

Actual

v2.4-clean

<Figure 6>

(a)



(b)

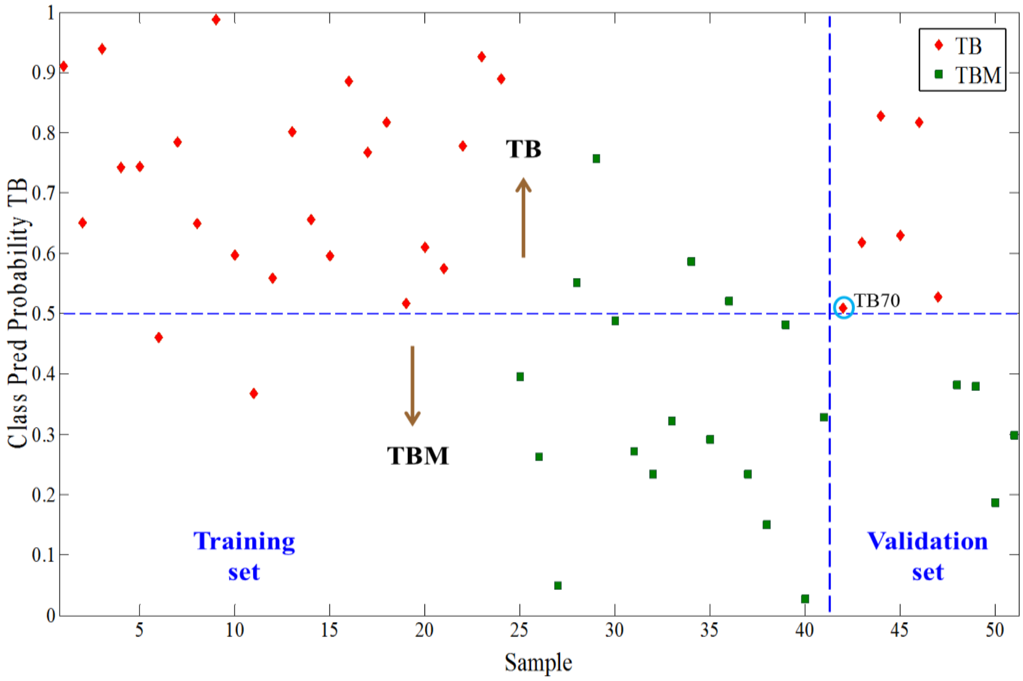
Assignation	Class 1	Class 2	
	6	**4**	**10**
Inconclusive (I)	0	0	**0**
Class 2 (TBM)	0	4 (50%)	**4**
Class 1 (TB)	6 (50%)	0	**6**

Actual

v2.4-clean

<Figure 7>

(a)



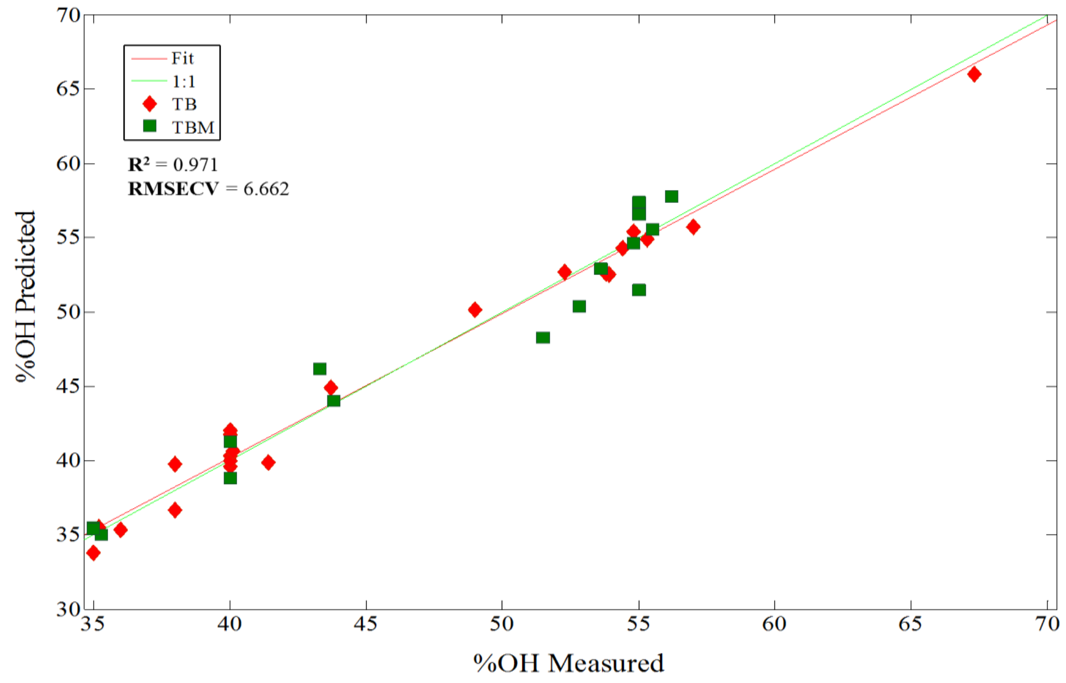
(b)

Assignation	6	4	10
Inconclusive (I)	0	0	0
Class 2 (TBM)	0	4 (50%)	4
Class 1 (TB)	6 (50%)	0	6
Actual	Class 1	Class 2	

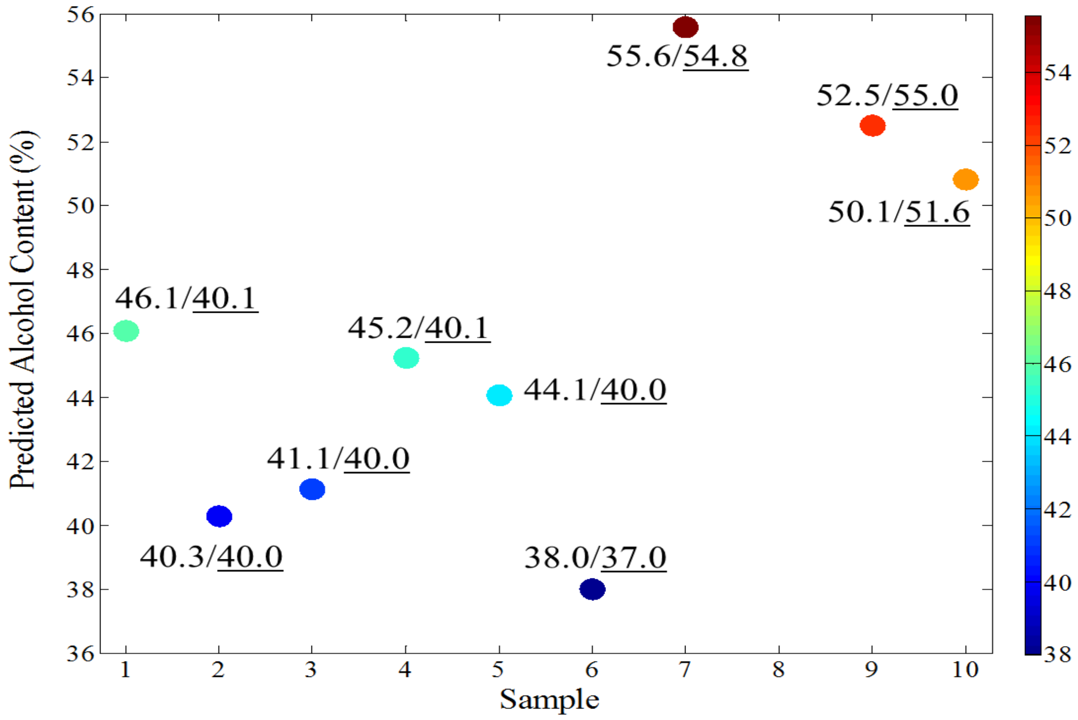
v2.4-clean

<Figure 8>

(a)



(b)



v2.4-clean