Contents lists available at ScienceDirect

Journal of Hazardous Materials

journal homepage: www.elsevier.com/locate/jhazmat

Multivariate chemometric discrimination of cigarette tobacco blends based on the UV–Vis spectrum of their hydrophilic extracts

Dimosthenis L. Giokas*, Nicholaos C. Thanasoulias, Athanasios G. Vlessidis

Laboratory of Analytical Chemistry, Department of Chemistry, University of Ioannina, 45110, Ioannina, Greece

ARTICLE INFO

Article history:
Received 29 April 2010
Received in revised form 28 August 2010
Accepted 31 August 2010
Available online 15 September 2010

Keywords:
Chemometrics
Classification and regression trees
Discriminant analysis
Principal components analysis
Tobacco
UV–Vis spectrophotometry

ABSTRACT

The application of UV–Vis spectrophotometry as an alternative or complementary approach to the classification of tobacco products is presented in this work for the first time. Two hundred fifty samples from five different cigarette brands composed of single and mixed tobacco blends were examined for that purpose on the basis of the UV–Vis spectrum of their aqueous extracts. Data transformation based on the normalization of absorbance intensities as a function of sample weight was employed in order to account for differences in the relative intensities of each sample. Principal components analysis (PCA) was used to extract outlier cases and sample classification was then pursued with the aid of discriminant analysis (DA) suggesting that a reduced number of variables (thirteen out of seven hundred initially available) could provide perfect classification (100% correct assignations) of samples containing single tobacco species or different blends and a fair classification of samples with similar composition (80% correct assignations) yielding an overall 95.7% correct classification. To this pursue, classification and regression trees were found to afford perfect classification of all samples using only a few logic rules based on appropriate split conditions at the expense of inserting 15 variables in the model.

1. Introduction

Although tobacco is represented by 67 species belonging to the *Nicotiana* genus of the nightshade family (*Solanaceae*), it is mainly represented by five species (Virginia, Oriental, Burley, Maryland and Rustica) which predominate the world production due to their wide use in tobacco industry [1]. Each of these species has different chemical composition (i.e. oil, sugar, nicotine content, etc.) and they are used either separately or most often in various blends in commercial tobacco products.

The large revenue generated by the worldwide tobacco commerce has led to the emerge of counterfeit and shoddy products with potentially increased hazards due to the uncontrolled production process (e.g. unregulated pesticide use) or treatment (e.g. unverified use of additives). As a consequence, there is a continuous demand to arbitrate the authenticity and control the quality of tobacco delivered to the market. Furthermore, beyond quality concerns, characterization of tobacco products may be an important form of trace evidence in forensic science since it can provide corroborative evidence regarding the presence of a person at the place of an event. Therefore, the characterization of tobacco products is a multifarious issue that merit investigation.

Empirically, the quality or the type of tobacco can be monitored by human sensory responses on the basis of subjective criteria like aroma, flavour or color. However, these criteria cannot be used objectively to support the authenticity of a product or discriminate it from other related products. For this reason, several efforts have been devoted to the provision of scientifically sound criteria that can provide an objective discrimination and characterization of tobacco products. Instrumental techniques have inevitably been brought into this effort. Methods based on gas and liquid chromatography, usually combined with mass spectrometric (MS) detection for the identification of major organic constituents [2–4] and inductively coupled plasma spectrometry for the determination of a multitude of inorganic elements [5,6], have been successfully employed to obtain a more accurate assessment of tobacco samples. However, tobacco contains over 3000 organic compounds and a large number of trace elements which are practically impossible to determine [5–7]. Moreover, the mixture of different tobacco species, the variety of grades and the relative use of additives in commercial products render straightforward comparisons an intricate task [3,7]. Therefore, less rigorous and cumbersome procedures are required which are able to provide a rapid, yet accurate, screening of tobacco products within a reasonable analysis time and experimental effort.

Up to date, many studies have demonstrated the usefulness of NIR spectroscopy to carry out a qualitative and quantitative analysis of tobacco samples in combination with various chemometric techniques that aid to relate sample properties to the observed spectra [8–10]. Depending on the available samples and the intended application a satisfactory discrimination of various tobacco samples has

* Corresponding author. Tel.: +30 26510 08400; fax: +30 2651008781.
E-mail address: dgiokas@cc.uoi.gr (D.L. Giokas).

0304-3894/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.jhazmat.2010.08.126

been reported [8–12], especially with regards to the identification of cultivation areas [8,10], tobacco varieties [13] or different commercial tobacco brands [9,14,15].

However, despite the progress made in instrumentation, the purchase of NIR spectrometer remains expensive, especially when compared to other spectroscopic detectors. Furthermore, a major disadvantage of NIR spectrometry is the dependence on time consuming and laborious calibration procedures and the complexity in the choice of data treatment [16,17]. Although this problem could be amortized by transferring the calibrations from the master instrument to several slaves, no specific methodology has yet gained widespread acceptance due to optical differences between the instruments [17]. As a result different calibration procedures are required depending on the available training data set and the intended application [8–15]. Although these procedures are efficient, they significantly increase the computational effort and in several occasions require too sophisticated statistical techniques which are not always available in commercial statistical software. Furthermore, NIR spectroscopy has generally weak sensitivity to minor constituents [16], as compared to other spectrometric detectors, therefore minor changes in sample composition can hardly be detected, which limits its applicability in exclusionary hierarchy of sample authenticity and especially in forensic investigations were sample availability may be limited.

Beyond NIR, the exploitation of other techniques like UV–Vis spectrophotometry in tobacco classification could offer an efficient and cost-effective alternative. Nevertheless, its analytical application towards this direction has been overlooked despite the fact that information retrieved from UV–Vis spectra have been successfully applied in many occasions. Using classification techniques which are readily accessible in most statistical packages and minimum sample pre-treatment, UV/Vis spectra interpretation has been successfully employed in a large array of discrimination assays e.g. for the discrimination of soils [18], blue ball-point pen inks [19], wines [20], pharmaceutical active compounds [21], etc. Surpisingly, its use in tobacco classification and authenticity has not received any attention despite the fact that color is an important property of almost all tobacco products.

With the above in mind, the aim of this study is two-fold. The first is to examine the analytical utility of UV–Vis spectrometry, in combination with appropriate chemometric techniques, as an alternative and cost-effective tool for the classification and discrimination of different tobacco samples. The second is to evaluate the possibility of discriminating among samples with similar composition within the same experimental and computational approach. For this reason, five commercially available cigarette brands containing various tobacco mixtures were selected and the UV–Vis absorbance spectra of their aqueous extracts were recorded. These spectra were then used to construct a model, based on principal component analysis (PCA) and linear discriminant analysis (LDA) that could characterize and classify each brand. The results suggested that LDA of the UV–Vis spectra could be a valuable tool in tobacco discrimination affording a very good classification of products of a single tobacco type and a fair approximation of products with similar blends. In the latter case, classification trees offer an efficient tool for the discrimination of samples with similar composition using two more variables than DA.

2. Materials and methods

2.1. Sample preparation and measurements

Five commercially available cigarette brands belonging to the same commercial category (formerly known as "lights") were used in the study. To avoid reference to the cigarette brand based on the initial letter they were randomly coded as: B, C, G, J and W. Brands B and J were selected as representatives of products using a single tobacco species while brands C, W and G as representatives of products composed of tobacco blends of similar composition (from a qualitative perspective since no detailed data were available). Brand J uses a tobacco species which is also used in C, W and G (at unknown proportion) while brand G uses a similar blend as C and W, the latter two being imported from the same manufacturer. For each brand, 5 packets were purchased and 10 cigarettes from the same batch were sampled from each packet. Samples were homogenized and approx. 0.1 g of tobacco was extracted with 500 mL distilled water in triplicate. The extract was centrifuged at 45,000 rpm for 15 min. The resulting supernatant liquid was carefully sampled with a Pasteur pipette, diluted 3 times, and its absorbance was measured on a JENWAY 6405 UV–Vis spectrophotometer in 1.00 cm quartz cell against distilled water as the blank. The scanning range was 200–900 nm at 1 nm intervals resulting in 700 experimental variables (absorbance values) for each sample. This procedure was run in triplicate for each tobacco sample and the results were averaged.

2.2. Validation data

Two sets of data were used to validate the results. In the first, a few packages from the initial batch that was used for calibration were left intact and opened 3 months later, in order to minimize the effect of air, moisture and possibly temperature that could alter their composition during storage. The second dataset comprised of samples purchased 3 months later from local stores.

3. Results and discussion

The extraction of tobacco products can be made with the aid of various water-miscible or water-immiscible organic solvents as well as water [22]. The extraction medium and the procedure used to retrieve tobacco components strongly influence the quality of the measurements. In this work, the water soluble fraction of cigarette tobacco was assessed mainly due to the simplicity of the extraction but also due to the fact that several tobacco products are available in moisted form (smokeless tobacco), therefore their hydrophilic extracts better reflect the tobacco components inserted into the human organism via smokeless application.

3.1. Data treatment

The raw absorbance values for each sample (A) were divided by weight so that they represented absorbance per g of tobacco (A_g). The use of A_g values instead of raw absorbance was necessary in order to overcome the problem of differences in the weights of the samples used to prepare the solutions and normalize the spectra per unit mass so that they could become comparable.

As most of the examinations performed in this work were parametric statistical techniques, it was necessary for the data (absorbance per unit mass) to be normally distributed. For this reason the $\log_{10}(A_g)$ values were calculated to ensure normality of the data [18,19,23,24]. Therefore, from now on, the $\log_{10}(A_g)$ values will be referred to as the original variables.

3.2. Variable selection

The average spectrum for each cigarette brand (as A_g vs. λ) is depicted in Fig. 1. From a macroscopic point of view, comparison of the average normalized spectra showed that differences in the absorption profile of each sample were evident only in the range between 225 and 350 nm (Fig. 1 inset), while no visual differences could be inferred at longer wavelengths. In our effort to

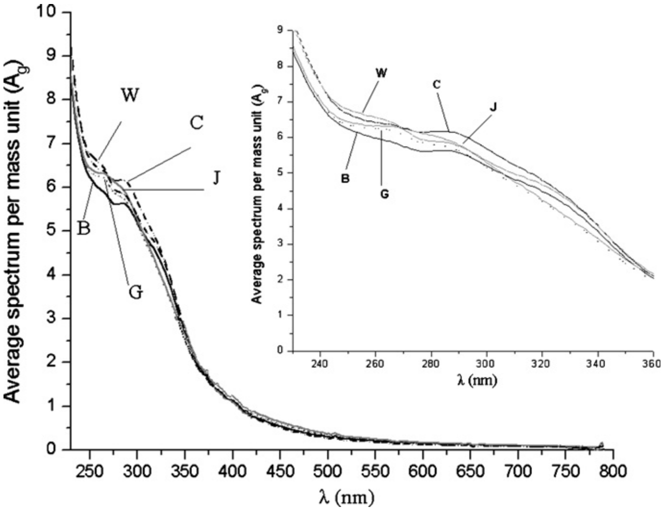

Fig. 1. Average absorbance spectra normalized per sample mass (A_g) vs. wavelength (λ) for all samples. Inset graph: Magnified view of A_g vs. λ in the UV region.

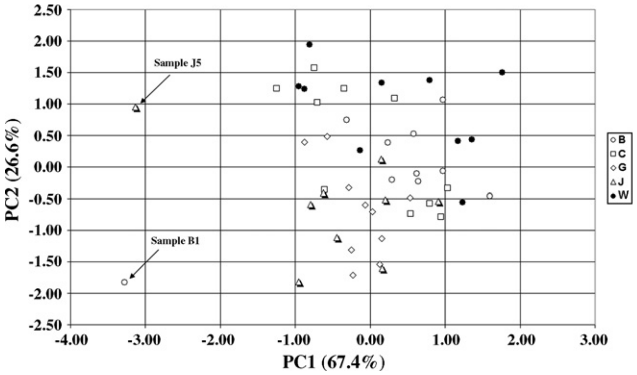

Fig. 2. Principle component graph for the detection of outlier cases (○) B, (□) C, (◇) G, (△) J, (●) W.

retrieve maximum information from the absorbance spectra, the 1st derivative of all samples was calculated but it did not reveal any pattern that could be employed to differentiate among the samples. Furthermore, since these samples were composed of a mixture of various blends, comparison of the average spectra could not be considered as a safe procedure for tobacco discrimination. That is because these spectra represent only the between sample variance with nothing been known about the within sample variance. Therefore, it was decided to select spectra from the entire UV–Vis region rather than isolating specific wavelength maxima that could import bias in the analysis.

Working with the entire dataset is impractical since the spectra were recorded in the range of 200–900 nm at 1 nm intervals meaning that 700 variables were recorded for each individual sample. To afford data reduction and decide which variables should be retained, *K*-means cluster analysis on the variables over the objects (tobacco samples) was performed. This feature reduction technique forms clusters with variables carrying similar information about the objects and the most representative variables can be chosen based on their proximity to the cluster centroids. However, as the number of selected clusters increases the variables become highly correlated which results in singular correlation matrices. In addition to this limitation, one has to use as few variables as possible when running classification tests with discriminant analysis (DA) in order to avoid capitalizing on change. A useful rule of thumb dictates that *m* variables should be used when *n* objects exists so that the criterion: $n/m > 3$, is satisfied. [18,25]. To compromise both limitations, 15 clusters were calculated in order to use the relevant variables with the closest proximity to group centroids, i.e. the absorbance of the solutions at 207, 220, 254, 321, 360, 398, 416, 441, 467, 496, 531, 566, 616, 691 and 776 nm.

3.3. Principal component analysis and data interpretation

Before proceeding with sample classification, it was necessary to remove outliers from the data. That is because they can negatively affect the results of DA through an overestimation of the within sample variance, which results in retaining the null hypothesis of no difference between group means. In multivariate systems, PCA can aid the observation of outliers by projecting the data in a two-dimensional plane after Varimax rotation of the first two extracted components [19]. The plane shown in Fig. 2 suggests that samples B1 and J5 exhibited noticeable deviation from their relative groups and should be considered as outliers.

After removal of the outliers from the dataset, PCA was applied again to investigate if there is any pattern in the data that can be used to extract information from the recorded spectra. The scree plot for the given dataset (graph not shown) showed that only the first two components complied both with the Kaiser criterion and satisfied the scree test, accounting for 92.3% of the total variance. Based on these data, the unrotated loadings of the experimental variables for the first two PC's were extracted. However, the unrotated factor loadings did not reveal any clear pattern among the variables. On the other hand, when these loadings were plotted again after Varimax rotation (Fig. 3) a good discrimination among the variables was obtained which can be assigned to the Vis region for the first principal component (PC1) and to the UV region for the second principle component (PC2). The Varimax rotated factor loadings and the coefficients of the extracted components gathered in Table 1 show that the contributions of the variables in the components are in agreement with the factor loadings. Interestingly, the contribution of the variables of the second principle component (UV spectra) are higher than those of the first principle component (Vis spectra) suggesting that the PC2 contributes more to the explanation of the observed variance. Furthermore, the $\log_{10}$ transformed normalized absorbance values at 398 and 416 nm exhibited factor loadings below 0.8, so they were not further considered since they do not seem to contribute significantly to the explanation of the observed variance. Finally the efficiency of PCA was re-assessed using the Barlett's sphericity test on the correlation matrix returning a $\chi^2 = 1888.8$ which is statistically significant at $p = 0.05$ and 120 degrees of freedom indicating that the variables are not orthogonal

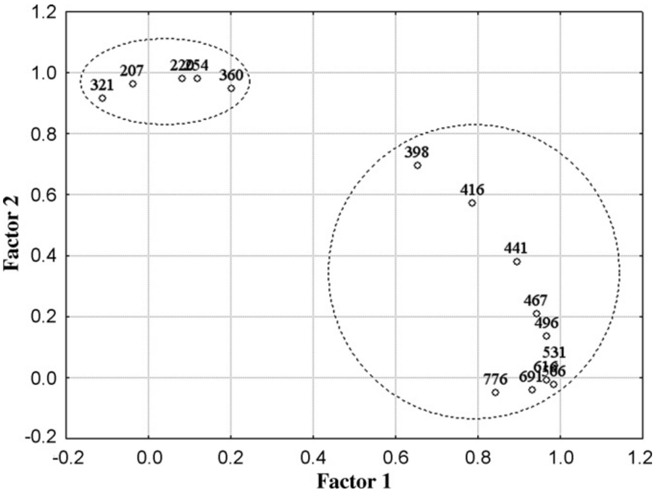

Fig. 3. Varimax rotated factor loading graph for the selected variables.

Table 1
Factor loadings and coefficients of the extracted components after Varimax rotation.

Variable	PC1		PC2	
	Loadings	Coefficients	Loadings	Coefficients
NM207	−0.0376	−0.0441	*0.9634*	0.1859
NM220	0.0829	−0.0291	*0.9795*	0.1840
NM254	0.1191	−0.0244	*0.9799*	0.1826
NM321	−0.1102	−0.0516	*0.9139*	0.1794
NM360	0.2021	−0.0122	*0.9465*	0.1729
NM398	0.6533	0.0570	0.6935	0.1061
NM416	0.7875	0.0795	0.5705	0.0771
NM441	*0.8953*	0.1013	0.3801	0.0363
NM467	*0.9427*	0.1144	0.2097	0.0017
NM496	*0.9678*	0.1208	0.1349	−0.0136
NM531	*0.9873*	0.1273	0.0369	−0.0331
NM566	*0.9853*	0.1295	−0.0237	−0.0447
NM616	*0.9668*	0.1265	−0.0090	−0.0411
NM691	*0.9321*	0.1233	−0.0416	−0.0459
NM776	0.8442	0.1121	−0.0493	−0.0438

Values in italics denote important factor loadings (>0.8).

but correlated therefore allowing to reduce the dimensionality of the original data.

3.4. Discriminant analysis and sample classification

In order to select the variables that can better classify samples based on their UV–Vis spectrum profile a standard discriminant model was employed. Although forward or backward DA modes could be used to provide additional data reduction the final outcome was not deemed satisfactory so they were not further considered. Besides the 13 variables inserted into the standard DA mode are a significant simplification to the variables present in the original dataset (approx. 700). The overall discriminatory power of the model, calculated by means of the Wilk's λ (defined as the ratio of the determinant of the within-groups variance – covariance matrix to the determinant of the total variance – covariance matrix) was 0.0285 ($F_{52,118} = 3.4252$) and was found to be statistically significant at the $p = 0.05$ level with a high average variable redundancy of 96.8% which can be attributed to the predictable pattern of the absorbance spectrum (Table 2).

The four discriminant functions (canonical roots) calculated for the model (Table 3) show that the first three functions account for almost 96.5% of the total variance. The post hoc classification of the training data set (Fig. 4a) showed that discriminant function root (1) was responsible for the separation of B–C–W from G–J, whereas discriminant function root (2) further aided the separation of B from C–W. Root (3) could be used for improving the separation between C and W as well as between G and J (Fig. 4b). The use of root 4 (Fig. 4c) did not provide any information with regard to sample discrimination. More details can be obtained from the classification matrix (Table 4). The diagonal of the matrix contains the correct classifications. An overall 95.7% correct classification was achieved with brands B, C, G and J being the most successful in their classification (100%). For brand W, two samples were misclassified as belonging to brand C. The determination of squared mahalanobis distances from group centroids showed that samples W2 and W7 were misclassified as belonging to group C although the absolute difference of the distances from group centroids was less than one suggesting that misclassification is marginally within the experimental error. Unfortunately, the exact qualitative or quantitative composition of each sample was not known in order to shed more light on the origin of these differences. Nevertheless, it was known that brands G, C and W use a blend composed of the same tobacco species, while C and W are produced from the same manufacturer and imported from the same country which justifies the observed overlap in their classification. On the other hand, brands B and J contain a single tobacco species (and different from each other).

Table 2
Standard discriminant analysis of the selected variables.

Variable	Partial λ	Tolerance	Redundancy (%)
NM207	0.8209	0.0145	98.6
NM220	0.6744	0.0105	98.9
NM254	0.8555	0.011	98.9
NM321	0.5148	0.0465	95.4
NM360	0.7361	0.0545	94.6
NM441	0.8767	0.0115	98.9
NM467	0.8089	0.0112	98.8
NM496	0.9707	0.0173	98.3
NM531	0.7838	0.0209	97.9
NM566	0.9701	0.0194	98.1
NM616	0.9635	0.0398	96.0
NM691	0.8593	0.0513	94.9
NM776	0.7917	0.1193	88.1

Table 3
Canonical root coefficients.

Coefficient	Root 1	Root 2	Root 3	root 4
b_{207}	71.46	26.29	3.85	83.19
b_{220}	−10.51	−158.47	21.66	−114.99
b_{254}	−87.31	44.79	−36.79	37.00
b_{321}	27.34	84.16	−35.24	14.12
b_{360}	14.04	−13.88	63.23	−27.08
b_{441}	25.32	−59.71	−72.38	−17.36
b_{467}	−34.76	96.97	72.23	−17.90
b_{496}	11.35	13.18	13.49	38.66
b_{531}	−29.40	−45.82	−44.33	14.11
b_{566}	15.51	12.07	−3.47	−9.31
b_{616}	−7.00	−4.27	12.13	−12.02
b_{691}	13.53	9.06	11.20	17.83
b_{776}	−5.15	−6.91	−9.86	−8.51
b_0	−32.57	50.25	−3.41	−7.50
Eigenvalue	4.998	1.480	0.860	0.267
% Cummulative variance	0.657	0.852	0.965	1.000

The DA was completed with the calculation of the so-called classification functions. These functions allow the post hoc classification of the items in the training data set or the classification of new items by calculating the relevant components scores and then entering the results into the classification functions. The function yielding the highest result would indicate the group which best fits the new sample. In this work, five classification functions were calculated (one for each group) and were of the form:

$$f(G_i) = b_0 + \sum_{i=1}^{n} b_i PC_i$$

where i is the number of groups, b_0 is the constant inherent to each group, n is the number of parameters used to classify a set of data into a given group, b_i is the weight coefficient assigned by DA to a given selected parameter (PC_i). The calculated coefficients are given in Table 5 and the post hoc classification of the training data set by means of these functions was found to be 95.7% as previously discussed.

Table 4
Classification matrix.

		Predicted classification					
		B	C	G	J	W	% Correct
Observed classification	B	9					100
	C		10				100
	G			10			100
	J				8		100
	W		2			8	80
	Total	9	12	10	8	8	95.7

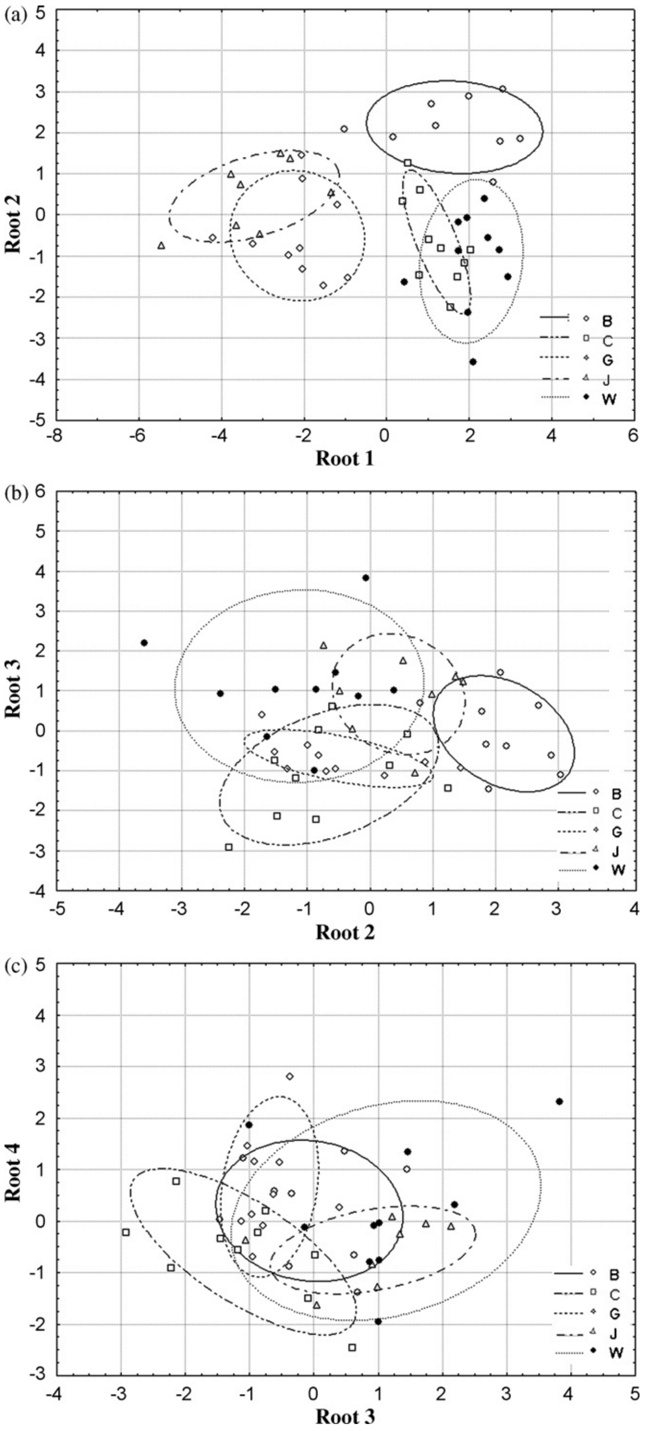

Fig. 4. Canonical score graph of the discriminant functions for the post hoc classification of the data. (◯) B, (□) C, (◊) G, (△) J, (●) W.

3.5. Classification trees

Table 5
Classification functions.

Coefficient	B	C	G	J	W
b_{207}	2402.1	2227.6	2096.7	1951.7	2350.0
b_{220}	10619.5	11134.7	11011.0	11045.8	11157.2
b_{254}	−9108.4	−9186.5	−8854.0	−8824.5	−9333.0
b_{321}	−4105.2	−4326.8	−4403.2	−4425.0	−4410.7
b_{360}	1515.2	1504.2	1446.6	1553.4	1641.2
b_{441}	712.2	955.7	810.1	631.4	830.8
b_{467}	−2165.3	−2480.5	−2341.9	−2074.8	−2409.7
b_{496}	−541.2	−627.4	−609.6	−634.4	−563.3
b_{531}	545.4	720.4	812.4	710.7	630.4
b_{566}	29.8	0.3	−63.3	−62.1	−7.5
b_{616}	−525.5	−513.4	−500.6	−463.2	−500.0
b_{691}	237.7	180.9	163.6	154.3	227.0
b_{776}	163.3	201.7	203.4	196.7	172.0
b_0	−2674.1	−2788.5	−2683.3	−2600.6	−2854.8

The resolution among these samples was then pursued with the aid of classification trees. Since classification and regression trees (C&RT) is a nonparametric technique, it does not make any assumptions on the data and they can be used straightforwardly to extract information form the raw data (A_g in this work). To avoid excessively tress and the formation of many nodes that could significantly complicate the interpretation various methods and stopping rules were examined. The best results, that provided a good compromise between classification capacity and the complexity of the tree, were obtained when all 15 variables determined by *K*-means cluster analysis on the variables over the objects were deployed in discriminant linear combination splits for ordered variables with prune on deviance as the stopping rule. The graph of Fig. 5 shows that classification is accomplished at 5 terminal and 3 interim nodes. The tree structure of Table 6 gathers sample classification for each node. As we can observe all samples were perfectly classified suggesting that C&RT can be a powerful tool to the classification of tobacco samples based on the UV–Vis spectrum of their hydrophilic components.

3.6. Cross validation

The results from the validation study are shown in Table 7. As we can observe, samples belonging to the same batch as the calibration samples were classified very effectively although some deviations were observed which can be attributed to instrumental and exper-

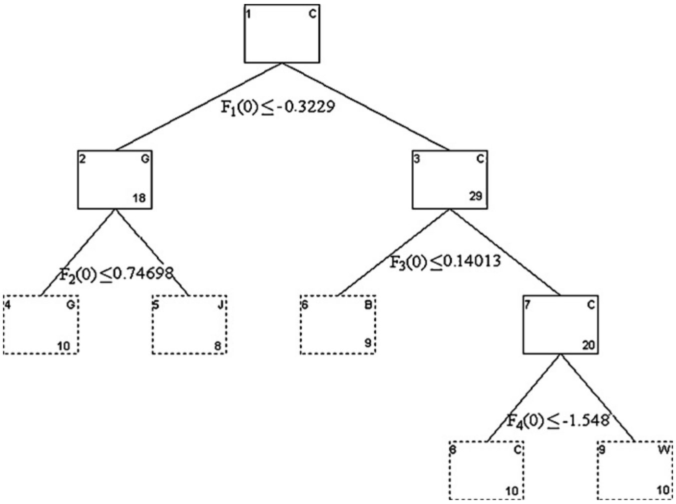

Fig. 5. Classification tree developed using all variables. Values in the upper left corner stand for the node number, letters in the up right corner represent the sample code and values in the lower left corner the number of samples classified to each node. Straight line boxes are interim nodes and dot line boxes terminal nodes. Linear combination split conditions: $F_1(0) = 0.32 + 0.64A_{207} - 0.56A_{220} - 1.78A_{254} + 1.56A_{321} + 0.14A_{360} - 22.85A_{398} + 61.12A_{416} - 23.19A_{441} - 30.73A_{467} + 16.52A_{496} - 54.92A_{531} + 52.66A_{566} - 17.68A_{616} + 30.86Ab_{691} - 23.42A_{776}$; $F_2(0) = -0.75 + 0.54A_{207} - 4.42A_{220} - 2.01A_{254} + 7.05A_{321} - 12.30A_{360} + 37.22A_{398} + 3.35A_{416} - 47.07A_{441} + 150.65A_{467} - 21.28A_{496} + 0.94A_{531} - 208.72A_{566} + 134.13A_{616} - 52.73A_{691} - 53.84A_{776}$; $F_3(0) = -0.14 - 3.41A_{207} + 7.71A_{220} - 3.31A_{254} - 11.62A_{321} + 17.29A_{360} - 46.25A_{398} - 0.16A_{416} + 124.61A_{441} - 122.36A_{467} - 55.08A_{496} + 27.19A_{531} + 54.89A_{566} - 110.71A_{616} - 170.03\ A_{691} + 86.75A_{776}$; $F_4(0) = 1.55 + 1.78A_{207} - 0.39A_{220} - 4.69A_{254} + 5.12A_{321} - 8.65A_{360} - 33.88A_{398} + 218.35A_{416} - 357.04A_{441} + 193.31A_{467} + 27.29A_{496} - 225.25A_{531} + 59.97A_{566} - 157.01A_{616} + 408.75A_{691} - 142.31A_{776}$.

Table 6
Classification tree structure.

Node No.	Left branch	Right branch	B	C	G	J	W	Predicted class	% Correct
1	2	3	9	10	10	8	10	C	21.3
2	4	5	0	0	10	8	0	G	55.5
3	6	7	9	10	0	0	10	C	52.6
4			0	0	10	0	0	G	100
5			0	0	0	8	0	J	100
6			9	0	0	0	0	B	100
7	8	9	0	10	0	0	10	C	50.0
8			0	10	0	0	0	C	100
9			0	0	0	0	10	W	100

Table 7
Cross validation of discriminant and C&RT (in brackets) models using two different data sets.

Brand	Number of samples	Predicted classification DA (C&RT)					Accuracy (%)
		B	C	G	J	W	
		Samples from the same batch					
B	7	7(7)	0(0)	0(0)	0(0)	0(0)	100(100)
C	5	0(0)	4(5)	1(0)	0(0)	0(0)	80(100)
G	5	0(0)	0(0)	5(5)	0(0)	0(0)	100(100)
J	6	0(0)	0(0)	0(0)	6(6)	0(0)	100(100)
W	5	0(0)	1(0)	0(0)	0(0)	4(5)	80(100)
Total	28	7(7)	5(5)	6(5)	6(6)	4(5)	90(100)
		Samples from different batch					
B	5	5(5)	0(0)	0(0)	0(0)	0(0)	100(100)
C	5	0(0)	4(4)	0(0)	0(0)	1(1)	80(80)
G	5	0(0)	1(1)	3(4)	0(0)	1(0)	60(80)
J	5	0(0)	1(0)	0(0)	4(5)	0(0)	80(100)
W	5	0(0)	1(2)	1(0)	0(0)	3(3)	60(60)
Total	25	5(5)	7(7)	4(4)	4(5)	5(4)	76(84)

imental errors due to the fact that samples were analyzed with a 3 months interval. On the other hand, samples belonging to different batch were misclassified and especially those with similar composition reflecting changes in the composition of the samples (e.g. harvesting conditions, storage, production processes, etc.) that cannot be accounted for by the models unless appropriate calibration sets are available that can maximize feature extraction and account for the specific characteristics of the samples. Nevertheless, the discrimination power of the model was not completely deteriorated which may be ascribed to its high efficiency as well as to the mild extraction conditions (i.e. water) reflecting changes only in their hydrophilic fraction.

The results of the cross validation study reflect the fact that classification techniques are by definition data-driven, therefore appropriate training sets are necessary to select variables that lead to a meaningful feature detection.

4. Conclusions

The application of UV–Vis spectrophotometry as an alternative or complementary approach to tobacco classification was assessed. The method was focused on the discrimination of commercial tobacco products (cigarettes) with different as well as similar qualitative composition. From the results it was concluded that tobacco samples could be perfectly discriminated based on the normalized UV–Vis spectrum of their aqueous extracts and applying multivariate chemometric techniques like linear discriminant analysis and classification and regression trees. Based on the results obtained, the proposed protocol was deemed satisfactory for supporting exclusionary hierarchy purposes for assessing the authenticity of a sample or as corroborative evidence in forensic examinations, since it can discriminate among samples with different composition and provide evidence on the composition of samples with similar composition. Most importantly, the method employs a simple extraction procedure and a series of multivariate chemometrics readily available in most commercial software thus alleviating the need for sophisticated statistical procedures and algorithm programming. However, since the classification procedure is data-driven, representative training and validation samples must be available that can be used to select variables that lead to a meaningful feature extraction.

References

[1] DG Joint Research Centre, Institute for Health and Consumer Protection, Tobacco, Cigarettes and Cigarette Smoke – An Overview, Ispra, Italy (2007).

[2] L.-F. Huang, K.-J. Zhong, X.-J. Sun, M.-J. Wu, K.-L. Huang, Y.-Z. Liang, F.-Q. Guo, Y.-W. Li, Comparative analysis of the volatile components in cut tobacco from different locations with gas chromatography–mass spectrometry (GC–MS) and combined chemometric methods, Anal. Chim. Acta 575 (2006) 236–245.

[3] G. Pieraccini, S. Furlanetto, S. Orlandini, G. Bartolucci, I. Giannini, S. Pinzauti, G. Moneti, Identification and determination of mainstream and sidestream smoke components in different brands and types of cigarettes by means of solid-phase microextraction–gas chromatography–mass spectrometry, J. Chromatogr. A 1180 (2008) 138–150.

[4] G. Xiang, L. Yang, X. Zhang, H. Yang, Z. Ren, M. Miao, A comparison of three methods of extraction for the determination of polyphenols and organic acids in tobacco by UPLC–MS–MS, Chromatographia 70 (2009) 1007–1010.

[5] M.J. Chang, J.D. Naworal, K. Walker, C.T. Connell, Investigations on the direct introduction of cigarette smoke for trace elements analysis by inductively coupled plasma mass spectrometry, Spectrochim. Acta B 58 (2003) 1979–1996.

[6] C.C. Crispino, K.G. Fernandes, M.Y. Kamogawa, J.A. Nóbrega, A.A. Nogueira, M.M.C. Ferreira, Multivariate classification of cigarettes according to their elemental content determined by inductively coupled plasma optical emission spectrometry, Anal. Sci. 23 (2007) 435–438.

[7] A. Wojtowicz, R. Bassilakis, W.W. Smith, Y.G. Chen, R.M. Carangelo, Modeling the evolution of volatile species during tobacco pyrolysis, J. Anal. Appl. Pyrolysis 66 (2003) 235–261.

[8] L.-J. Ni, L.-G. Zhang, J. Xie, J.-Q. Luo, Pattern recognition of Chinese flue-cured tobaccos by an improved and simplified K-nearest neighbors classification algorithm on near infrared spectra, Anal. Chim. Acta 633 (2009) 43–50.

[9] E.D.T. Moreira, M.J.C. Pontes, R.K.H. Galvao, M.C.U. Araujo, Near infrared reflectance spectrometry classification of cigarettes using the succes-

sive projections algorithm for variable selection, Talanta 79 (2009) 1260–1264.

[10] M. Hana, W.F. McClure, T.B. Whitaker, M.W. White, D.R. Bahler, Applying artificial neural networks: Part II. Using near infrared data to classify tobacco types and identify native grown tobacco, J. Near Infrared Spectrosc. 5 (1997) 19–25.

[11] X. Liu, H.-C. Chen, T.-A. Liu, Y.-L. Li, Z.-R. Lu, W.-C. Lu, Application of PCA-SVR to NIR prediction model for tobacco chemical composition, Spectrosc. Spectral Anal. 27 (2007) 2460–2463.

[12] Y. Zhang, Q. Cong, Y. Xie, J. Yang, B. Zhao, Quantitative analysis of routine chemical constituents in tobacco by near-infrared spectroscopy and support vector machine, Spectrochim. Acta A 71 (2008) 1408–1413.

[13] Y. Shao, Y. He, Y. Wang, A new approach to discriminate varieties of tobacco using vis/near infrared spectra, Eur. Food Res. Technol. 224 (2007) 591–596.

[14] C. Tan, M. Li, X. Qin, Study of the feasibility of distinguishing cigarettes of different brands using an Adaboost algorithm and near-infrared spectroscopy, Anal. Bioanal. Chem. 389 (2007) 667–674.

[15] C. Tan, X. Qin, M. Li, Comparison of chemometric methods for brand classification of cigarettes by near-infrared spectroscopy, Vib. Spectrosc. 51 (2009) 276–282.

[16] H.B. Pfaue, Analysis of water in food by near infrared spectroscopy, Food Chem. 82 (2003) 107–115.

[17] M. Blanco, I. Vilarroya, NIR spectroscopy: a rapid-response analytical tool, Trends Anal. Chem. 21 (2002) 240–250.

[18] N.C. Thanasoulias, E.T. Piliouris, M-S.E. Kotti, N.P. Evmiridis, Application of multivariate chemometrics in forensic soil discrimination based on the UV–Vis spectrum of the acid fraction of humus, Forensic Sci. Int. 130 (2002) 73–82.

[19] N.C. Thanasoulias, N.A. Parisis, N.P. Evmiridis, Multivariate chemometrics for the forensic discrimination of blue ball-point pen inks based on their vis spectra, Forensic Sci. Int. 138 (2003) 75–84.

[20] S.A. Bellomarino, X.A. Conlan, R.M. Parker, N.W. Barnett, M.J. Adams, Geographical classification of some Australian wines by discriminant analysis using HPLC with UV and chemiluminescence detection, Talanta 80 (2009) 833–838.

[21] G.M. Hadad, A. El-Gindy, W.M.M. Mahmoud, HPLC and chemometrics-assisted UV-spectroscopy methods for the simultaneous determination of ambroxol and doxycycline in capsule, Spectrochim. Acta A 70 (2008) 655–663.

[22] A. Rodgman, T.A. Perfetti, The Chemical Components of Tobacco and Tobacco Smoke, CRC Press, Boca Raton, Fl, USA, 2009.

[23] J.H. Zar, Biostatistical Analysis, Fourth ed., Prentice-Hall, Englewood Cliffs, NJ, USA, 1999.

[24] C.Z. Katsaounos, D.L. Giokas, I.D. Leonardos, M.I. Karayannis, Speciation of phosphorus fractionation in river sediments by explanatory data analysis, Water Res. 41 (2007) 406–418.

[25] D.L. Massart, B.G.M. Vandeginste, S.N. Deming, Y. Michotte, L. Kaufman, Chemometrics: A Textbook, Elsevier, Amsterdam, The Netherlands, 1988.