**SSPT 2 - Figueres - Near-infrared spectroscopy and pattern recognition as screening methods for classification of commercial tobacco blends**

Authors: FIGUERES G.; ANDRIEU E.; BIESSE J.P.; VIDAL B.; DUMERY B.

# Near-infrared spectroscopy and pattern recognition as screening methods for classification of commercial tobacco blends.

Gilles FIGUERES, Elise ANDRIEU, Jean-Paul BIESSE, Bernard VIDAL, Bernard DUMERY
*ALTADIS Research Centre - 4, rue André Dessaux - 45404 Fleury les Aubrais France*

## Abstract

Group classification of tobacco blends is commonly performed using several different types of compositional data including tobacco compounds, additive or process components. But all of these wet chemistry methods are relatively time-consuming. A need exists for a fast and reliable procedure in order to determine the blend type. The purpose of this study was to assess the ability of near infrared reflectance spectroscopy (NIRS) to be a qualitative means to classify tobacco material based on its spectral features. Two hundred and seventy eight “blond” commercial cigarette products from European countries have been subjected to NIRS at a broad range of wavelengths (400 – 2500 nm). A Hierarchical Cluster Analysis (HCA) performed on chemical variables measured in tobacco blends has enabled us to define four distinctive groups among the commercial products. Different well-known supervised pattern recognition algorithms were applied to spectral data: Linear Discriminant Analysis either after a data compression step with a PCA (LDA/PCs) or on selected wavelengths by stepwise discriminant analysis (LDA/SW), Discriminant Partial Least Squares (DPLS) and Soft Independent Modelling of Class Analogy (SIMCA). The performance of the multivariate data models investigated here in combination with a variety of wavelength regions and data pre-treatment is evaluated by comparing the classification predictions with the predefined chemical categories. The LDA using factors scores calculated from near infrared region (1100 - 2500 nm) showed more accurate differentiations than those based on selected wavelengths or on the DPLS approach; whereas the SIMCA algorithm has a weak discrimination power. The work reported in this paper confirmed that near infrared spectroscopy coupled with an appropriate chemometric procedure can reveal the identity of a range of commercial blends with a high degree of confidence.

**Key words:** visible, near-infrared spectroscopy, classification, tobacco blends, PCA, LDA, DPLS, SIMCA.

## Introduction

Cigarette blends are made up of various grades of different tobacco types, expanded tobaccos, and reconstituted leaves with or without additives (casing or aromas). The blend composition is classically characterised by chemical analysis and sensory evaluation in order to survey the quality of industrial process, to ensure the identity of a blend, or to differentiate and discriminate blends of different origins. These analyses are relatively time consuming and require expensive analytical equipment. Near-infrared reflectance spectroscopy (NIRS) has been widely used for many years for the quantitative determination of many characteristics or blending ratios in food and agricultural products, including tobacco materials (1, 2, 3, 4) due to the speed of analysis, minimum sample preparation, relative low cost and non destructive method.

The reflectance spectrum of a tobacco blend is the summation of the spectra of its major chemical components and the absorption properties of the spectrum result from stretching and bending of bonds between a hydrogen and heavier atoms, oxygen, carbon and nitrogen. One way is to use these spectral characteristics for prediction of tobacco composition; this approach requires calibration to correlate the spectral response with known chemical concentrations.

Another way is to use NIRS as a qualitative tool in order to classify samples based on their spectral features. NIR spectral data combined with multivariate methods have been used for classification, discrimination or identification. Supervised pattern recognition methods applied to spectrometric data have been useful in classifying food and agricultural materials (5, 6, 7) without the need for chemical data. Until now only a few studies (2, 8) have been reported for the classification of tobacco mixtures by using the NIRS but in most instances it was employed to discriminate among a small number of them or a tobacco mixture laboratory preparation.

In this work we assessed the possibility of discriminating tobacco blends of cigarette brands sold in different European countries based on the NIR spectral information alone in combination with various pattern recognition methods. We compared the performance of these chemometric techniques to propose the most accurate method in order to know if near-infrared spectroscopy represents an alternative option for quality screening.

## Materials and Methods

Two hundred and seventy eight “blond” cigarette brands from different companies commercialised in twelve European countries were collected. These finished products represented different taste lines and covered four years from 2000 to 2003.

Tobacco samples were dried at 35 °C and ground through a 0.5 mm screen.

Reflectance spectra were recorded over the wavelength range 400 – 2500 nm (visible and near-infrared regions) at 2 nm intervals using a scanning monochromator NIRSystems 6500 (Foss NIRSystems Inc.). Duplicate acquisitions for each sample were collected using a 16, 32, 16 scan sequence. The average of the two raw spectra (log 1/reflectance) was pre-treated with Standard Normal Variate and Detrending mathematical transformations (SNV-D) and then derived once or twice in order to reduce the variation due to the physical effects and to enhance the spectral information.

Wet chemical analyses were performed to determine the main constituents of tobacco blends such as nitrogen compounds (total nitrogen, nitrate, amino acids, ammonium, alkaloids), carboxylic acids and volatile organic acids, polyphenols, sugars and amino sugars, inorganic elements and additives such as humectants… in total, 26 chemical criteria were used to characterise the tobacco blends.

### Chemometric analyses

*Linear Discriminant Analysis (LDA)*

LDA focuses on finding optimal boundaries between classes to permit the maximum separation among the different categories. But a data compression step is needed (9) in order to get a number of spectral variables lower than the sample number (condition of LDA application).Two data compression ways were tested: principal component analysis (PCA) and stepwise discriminant analysis (SW).

- A PCA was performed on centred and scaled spectral data and the score of each sample projected onto the factor space was calculated. The normalised scores of the PCA step were used in the LDA (LDA/PCs) based on the Mahalanobis distance between groups, to obtain a classification model. An unknown sample was assigned to the group with the nearest distance to the gravity centre of a group.
- The original spectral data were also reduced by a stepwise discriminant analysis. In this compression step the most explanatory wavelengths were selected step by step and used to calculate the discriminant modelling functions in the LDA procedure (LDA/SW).

PCA and LDA processing were carried out with Uniwin software coupled with StatGraphics Plus Version 5.1 (Manugistics Inc.).

*Discriminant Partial Least Squares analysis (DPLS)*

The DPLS exploits the potential of PLS regression against a coding variable to predict sample classifications (10). The class membership of each sample is described by an artificial "dummy" variable (Y): the dummy value "2" is assigned to a group with similar spectra and value "1" to the other groups. The comparison of each group to the others is performed in attributing successively the value "2" to each group. The DPLS model is then developed to relate the spectral data (X) and the assigned reference value (dummy variable Y). The DPLS regression method used the cross-validation for the estimation of the prediction error and the selection of the optimum number of calibration factors. A sample was classified into a group when its predicted value Y was between 1.5 and 2.5 (index of class belonging probability).

The computation was carried out using Winisi software version 1.5 (Foss NIRSytems Inc.)

*Soft Independent Modelling of Class Analogy (SIMCA)*

SIMCA uses the modelling properties of principal component analysis. Each class of observations is modelled separately by disjoint PC- models. After the separate modelling of each class, the models are used to predict a likely class membership for new observations. Based on the residual variation of each class, the distance of each observation to the model is calculated. A critical distance is computed for each class model. The closest model to the unknown sample is chosen by the value of sample to model distance with a certain level of significance. The critical distance corresponds to the 0.05 level and defines a 95% tolerance interval. The SIMCA method is performed using the SIMCA P software version 9.0 (Umetrics AB).

## Methodological approach

All the classification models described above were developed on a calibration or training sample set with known classes and the model performances were evaluated on a validation or test sample set.

For this study, the 278 finished products were separated into two sets: the training set contained 126 samples and the validation set was composed of 152 samples.

Model assessment was based on the correct classification rates by comparison of the classification predictions with the "true" classes. The "true" classes were defined from a hierarchical cluster analysis (HCA) performed on the 26 chemical criteria used to characterise the tobacco blends of the training set (126 samples).

The various samples were aggregated into homogeneous classes of tobacco blends according to their chemical composition. The mode of aggregation chosen was the Ward algorithm (StatGraphics Plus software).

The dendrogram (Figure 1) summarises the HCA results of the training set and indicates four main groups.

The samples of group A (n = 24) are high in volatile nitrogen compounds, sugars and process markers and contain more additives than the other groups. The samples of group B (n = 26), which was close to the group C (n = 45) but with less polyphenol compounds, amino acids or carboxylic acids, contain high levels of alkaloids and sugars.. The last group D (n = 31) very distinct from the three others was characterised by high contents of sugars and polyphenols but was poor in nitrogen compounds, and had little amount or no additives.

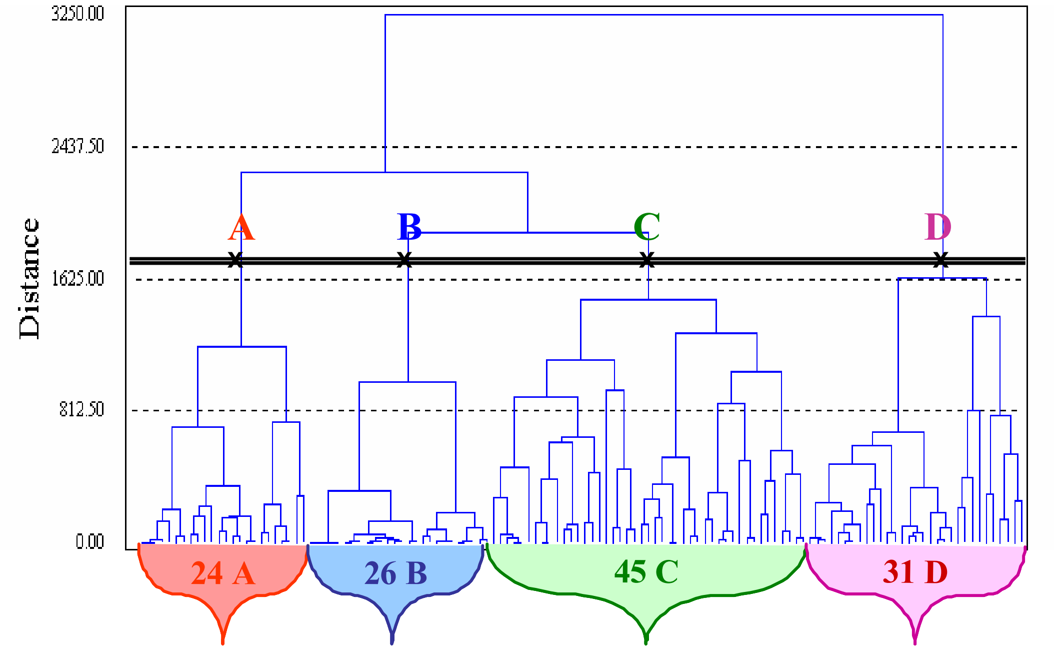

**Fig. 1**. Cluster tree of 126 tobacco blends from training set based on chemical variables.

The spectral classification of cigarette blends developed with different chemometric methods was compared with the chemical classification. Classification computations were carried out on the whole spectrum data (400 to 2500 nm called "visible + nir" region) or on the near-infrared part of spectrum (1100 to 2500 nm called "nir" region).

## Results and Discussion

The first LDA discrimination was performed after a data compression by PCA. An investigation was conducted to find the optimal PC number in order to obtain the most efficient LDA: in the "visible+nir" region, 30, 40 60 and 80 PCs and in the "nir" region, 10, 20 and 30 PCs were tested respectively. Figure 2 shows the percentage of training set samples correctly classified according to the PCs number: using 20 PCs in the near infrared segment and using 30 PCs in the "visible+nir" segment, tobacco blends were classified correctly in 100% of the cases. The discriminant scores plot generated using the first two discriminant factors revealed a good separation within four groups. The graphical representation of Figure 3 corresponds to the LDA applied to the nir spectral segment: separation of the four groups along axis 1 and along axis 2 for groups A and B.

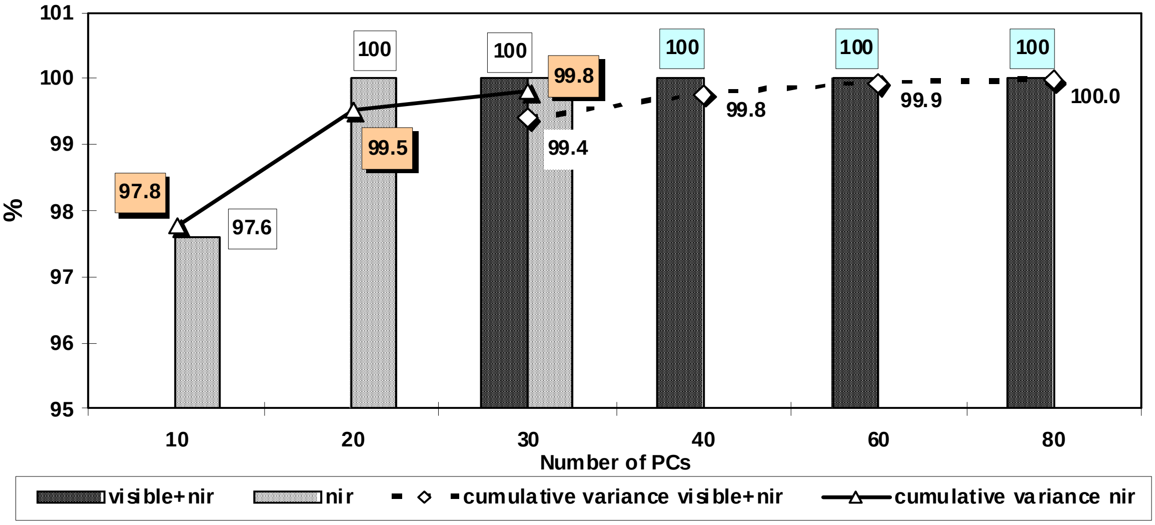

**Fig. 2.** Percentage of correctly classified samples according to PCs number and percentage of cumulative variance for the samples of training set.

Both models (20 and 30PCs) were used to test the validation set samples both on the "nir" and the "visible+nir" regions. The test samples were projected as supplementary observations onto the factor space defined by training spectra and their scores were calculated.

The 20 PCs (nir region) and 30 PCs (visible +nir) models classified correctly 144 (94.7%) and 143 (94.1%) out of a total of 152 test samples, respectively.

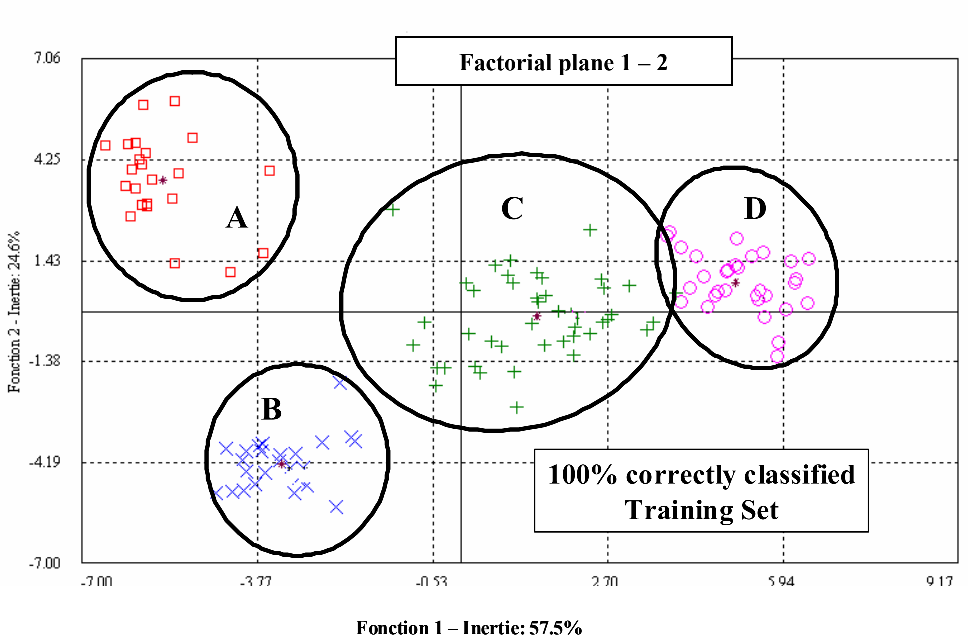

**Fig. 3.** Discriminant score plot (LDA/20 PCs; nir spectral segment).

In order to prevent the multicolinearity problem caused by wavelengths, a stepwise LDA was also performed to select the most effective variables among the original 206 wavelengths (only 138 in the nir segment): after averaging the spectrum on every five data points, 71 and 57 wavelengths were selected in “visible+nir” and “nir” regions, respectively. Then, a LDA was computed based on these selected variables. With the LDA/SW method, the correct discrimination percentage attained 100% in the training set for both spectral regions. In the validation set, 12 samples were misclassified (7.9%) in the “visible+nir” region, among which 10 samples belong to group C and in the “nir” region, 15 samples (12 from group C) were misclassified (9.9%). Most of the misclassified samples from group C were recognised as samples belonging to group B (the closest group to C).

The DPLS regression models were developed after first or second derivative of spectra using two various wavelength regions (vis+nir or nir alone)

Table 1 shows calibration statistics from the discriminant analysis. The results indicate that the DPLS models accounted for 84 to 90% (coefficient of determination $R^2$) of the variability for tobacco blend classification with standard errors very close around a value of 0.2. The wavelength region of 400-2500 nm gave the best calibration statistics with a first derivative of spectral data (99.3% correctly classified); only one blend out of 126 samples was misclassified.

In the validation set, the prediction rate varied from 82% (second derivative of the near infrared spectrum) to 91.4% (first derivative of the whole spectrum). For this last prediction, only 13 samples were not well predicted: 8 not correctly identified but assigned to the closest group and 5 samples as missing samples.

**Table 1.** Calibration and prediction statistics using DPLS algorithm.

| wavelength range (nm) | derivative order of log1/R | R2 | SECV | % of samples correctly classified in training set | optimal PC ı for moc |
|---|---|---|---|---|---|
| 400-2500 | 1st | 0.90 | 0.21 | 99.3 | 14 |
| | 2nd | 0.87 | 0.21 | 96.0 | 11 |
| 1100-2500 | 1st | 0.84 | 0.20 | 98.6 | 9 |
| | 2nd | 0.85 | 0.19 | 97.4 | 9 |

Finally a SIMCA class model was created after a local PC-analysis of each homogeneous class of observations. The optimal PC number was determined for each class by cross-validation. In this study 4 local PCA-models were constructed pertaining to each separate class of 126 tobacco blends from the training set. In SIMCA, the closest model to the sample was chosen by the value of sample-to-model distance with a certain level of significance, 5% in this report. The results from a classification analysis may be presented in a plot called the Coomans' plot: class distances for two classes are plotted against each other in a scatter plot. The Figure 4 shows the distribution of calibration blends between groups A and C. The separation of the modelling samples was fairly satisfactory: 97 and 96% of samples were correctly classified in both spectral segments, "visible +nir" and "nir", respectively.

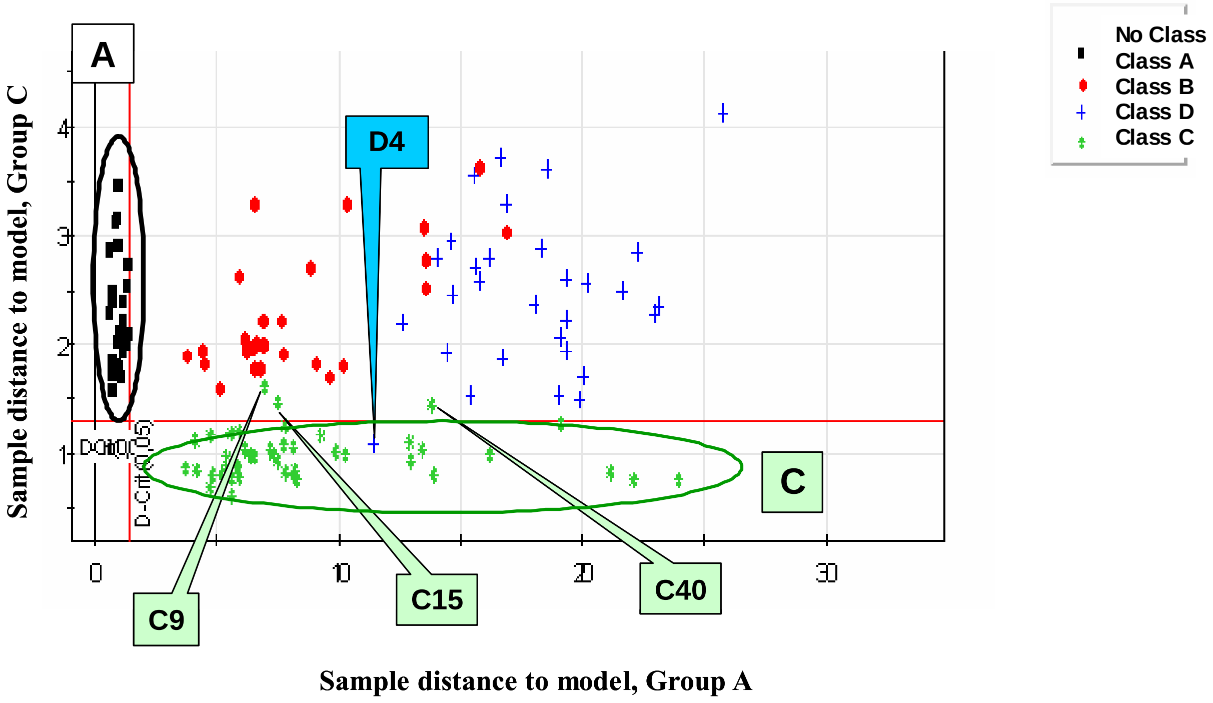

**Fig. 4.** Coomans plot of discrimination results between groups A and C using transformed nir spectra (SNV-D 1st derivative) of the calibration set.

The following Coomans plot (Figure 5) illustrates that the classification of test blends was not satisfactory, in fact only 36 and 31% of samples were correctly classified, in the visible+nir and the nir regions, respectively.

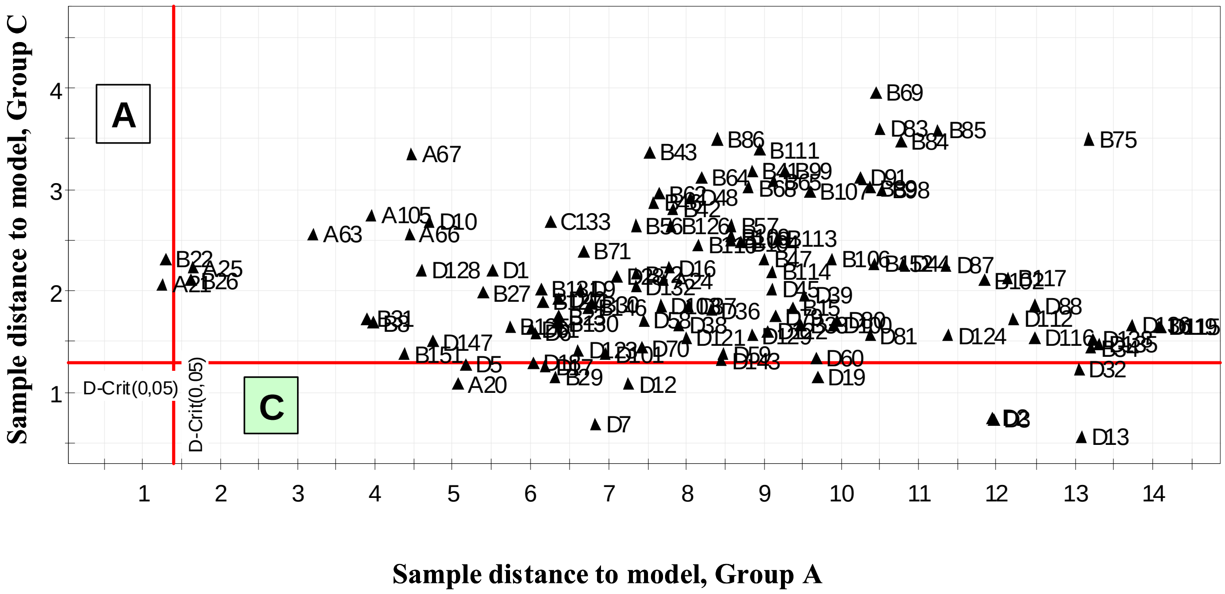

**Fig. 5.** Coomans plot of discrimination results between groups A and C using transformed nir spectra (SNV-D 1st derivative) of the validation set.

The main results found with the different statistical supervised methods are summarised in Table 2.

**Table 2.** Comparison of method efficiency on training and validation sets (% correctly classified).

| | | | LDA | | | | DPLS | | SIMCA |
|---|---|---|---|---|---|---|---|---|---|
| | | | LDA / PCs | | LDA / SW | | 1st derivative | 2nd derivative | |
| Visible+Nir | % correctly classified | TRAINING SET n = 126 | 30 PCs | 100% | 71λ | 100% | 99.30% | 96.0% | 96.9% |
| | | VALIDATION SET n = 152 | | 94.1% (143/152) | | 92.1% (140/152) | 91.4% (139/152) | 86.2% (131/152) | 35.9% |
| Nir | | TRAINING SET n = 126 | 20 PCs | 100% | 57λ | 100% | 98.60% | 97.40% | 95.8% |
| | | VALIDATION SET n = 152 | | 94.7% (144/152) | | 90.1% (137/152) | 88.8% (135/152) | 82.2% (125/152) | 30.7% |

## Conclusion

This study has shown the potential of visible and near infrared spectroscopy as an effective screening tool on the basis of spectral information to classify tobacco blends from cigarette products. Four supervised pattern recognition methods in combination with different mathematical pre-treatment of signal data and spectral regions have been compared. The linear discriminant analysis based on principal component scores on the first derivative nir region is the most appropriate method to classify blends from spectral data with a high prediction rate of 95%. However, we could note that the DPLS algorithm was the most convenient with almost 92% of correctly classified samples in the visible and near infrared spectral segment combined with the first derivative data. The SIMCA approach exhibited the poorest classification performance and seemed not suitable to discriminate the unknown tobacco blends of finished products.

Finally this qualitative approach of near infrared spectral data using supervised multivariate techniques could be a valuable and fast identification tool of tobacco blend types.

## References

[1] D.A. Burns, E.W. Ciurczak, Handbook of NIR analysis, (1992).

[2] S.C. Lo, 49th TCRC, Lexington, KY, USA, 1995.

[3] AMC. Hovell, MFP. Barido, 50th TCRC, VA, USA, 1996.

[4] J.M. Johnson, E.L. Butler, R.D. Stevens, 53rd TSRC, Montreal, Quebec, Canada, 1999.

[5] K. IIzuka, T. Aishima, J. Food Sc. 62 (1997) 101.

[6] M.D. Atkinson, A. P. Jervis, R.S. Sangha, Can. J. For. Res. 27 (1997) 1896.

[7] Y. Roggo, L. Duponchel. J.-P. Huvenne, Anal. Chim. Acta 477 (2003) 187.

[8] L.M. Dominguez, S.K. Seymour, in Making Light Work: Advances in Near Infrared Spectroscopy, Ed by I. Murray and I.A. Cowe.VCH, Weinheim, 179 (1992).

[9] A. Siriex, G. Downey, J. Near Infrared Spectrosc. 1 (1993) 187.

[10] D. Cozzolino, H E. Smyth, M. Gishen, J.Agri.Food Chem. 51 (2003) 7703