Computers and Electronics in Agriculture 159 (2019) 28–33

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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

Original papers

Discrimination of *Camellia japonica* cultivars and chemometric models: An interlaboratory study

Clara Sousa[a], Cristina Quintelas[b], Catarina Augusto[a], Eugénio C. Ferreira[b], Ricardo N.M.J. Páscoa[a,*]

[a] LAQV/REQUIMTE, Departamento de Ciências Químicas, Faculdade de Farmácia, Universidade do Porto, Portugal
[b] CEB – Centre of Biological Engineering, University of Minho, Braga, Portugal

ARTICLE INFO

Keywords:
Camellia japonica
NIR spectroscopy
Interlaboratory comparison
PCA
PLS-DA

ABSTRACT

Camellia japonica is a valued plant since ancient times throughout the world mostly due to their ornamental flowers. It has a high number of cultivars, with very similar phenotypic and genotypic characteristics which are difficult to discriminate, being some of them often rare and with a high price at the market. Their discrimination is mostly done through visual inspection of the morphologic characteristics which is a hard and inefficient task. Spectroscopic techniques had already been used for taxonomic purposes at species and sub-species level with success and could be an alternative for accurate *C. japonica* cultivars discrimination. Despite the already recognized success of such techniques, most of the studies arises from a single laboratory and little is known about the robustness of these techniques regarding interlaboratory data transferability. In this context, the work developed herein presents a double aim: (I) to explore the ability of near infrared (NIR) spectroscopy and partial least square – discriminant analysis (PLS-DA) to discriminate *C. japonica* cultivars and (II) to evaluate data transferability between two independent laboratories (Lab A and Lab B). Air-dried leaves NIR spectra of 43 *C. japonica* plants (15 distinct cultivars) were acquired in both laboratories using two similar NIR instruments (same manufacturer and model). Spectra were further modelled by PLS-DA after exploratory analysis using principal component analysis (PCA): (I) individually for Lab A and Lab B; (II) using Lab A as calibration and Lab B as validation set and *vice-versa* and (III) with using Lab A and Lab B data together. The percentage of *C. japonica* cultivars discrimination for both laboratories was nearly the same (around 83%) indicating no significant differences between Lab A and Lab B analysis. However, the results were quite poor when spectra were modelled with data from a single laboratory and validated with the other (65.5 and 63.8%). When data were merged, 85.9% of correct cultivars assignments were obtained. The results herein obtained could benefit of including additional cultivars and plants; but demonstrated the ability of NIR spectroscopy for *C. japonica* cultivars discrimination. Regarding data transferability, even when dealing with a similar instrument, some issues arisen preventing easy and efficient spectral library transfers.

1. Introduction

Since ancient times that plants are used in the form of infusions, to treat several diseases due to their therapeutic and antioxidant properties, or as ornamental purposes (Kanth et al., 2014). *Camellia japonica* is one of these plants. It was initially found in Korea, Japan, China and Taiwan between altitudes of 300–1100 m (Kanth et al., 2014) but actually it is disseminated throughout the world. In fact, in the north of Spain (Galicia), around 2.5 million Camellia plants (majority of *C. japonica*) are produced and exported to other European countries (Salinero et al., 2012). These plants present vast phenotypic differences, more than 32,000 known cultivars (Vela et al., 2013), and possess a great economical value due to its beautiful flowers (Salinero et al., 2012). The study of its genetic structure revealed that these plants maintain high levels of genetic diversity and a high evolutionary potential which make them less susceptible to environmental stresses (Oh et al., 1996). This could explain with these species are also well adapted in many other countries. The common methods applied for the discrimination of *C. japonica* cultivars usually depend on the visual inspection of the morphologic characteristics of each plant by an expert.

* Corresponding author at: LAQV/REQUIMTE, Departamento de Ciências Químicas, Faculdade de Farmácia, Universidade do Porto, Rua Jorge Viterbo Ferreira, 228, 4050-313 Porto, Portugal.
E-mail address: rnpascoa@ff.up.pt (R.N.M.J. Páscoa).

https://doi.org/10.1016/j.compag.2019.02.025
Received 23 October 2018; Received in revised form 18 February 2019; Accepted 24 February 2019

However, this method is not very efficient. In this sense, it is essential to find an easy, low-cost and effective method for the discrimination of *C. japonica* cultivars. Near infrared (NIR) spectroscopy presents all these features and has already been successfully applied in combination with chemometric models for the discrimination of other plants, namely citrus species (Páscoa et al., 2018), hops varieties (Machado et al., 2018), Amazonian plant species (Lang et al., 2017), and grapevine varieties (Gutierrez et al., 2016). In all these works, around 90% of correct predictions were obtained considering air-dried powdered leaves (with exception of the latter work where the leaves were scanned directly on field) which demonstrate the suitability and efficiency of the technique. However, the robustness of such results was never assessed.

The robustness of a technique can be evaluated through the comparison of the results obtained in different laboratories once each analytical method depends on different factors (e.g. analytical instruments, operators, laboratory conditions) (Yoon et al., 2000). In fact, little is known about the influence of the instrument, operators and laboratory conditions in results obtained in a given procedure. Bartl and collaborators in 1996 explored this topic through the analyses of aqueous solutions using different Fourier transform instruments (Bartl et al., 1996). Despite the data produced by different workers and instruments being different, the results were very similar. Nevertheless, these authors found that the collection of background spectra can be uneven and to transfer the data from one instrument to another can be very difficult. Yoon and collaborators reported the same problems about the transference of spectral libraries from one instrument to another because each instrument manufacturer tend to use their own file formats (Yoon et al., 2000). These authors pointed also some difficulties when transferring spectral libraries from one instrument to another of the same manufacturer but with different sampling accessories. In addition, interlaboratory studies including chemometric models were barely explored and, to the best of our knowledge, only one work was published exploring this topic. It used the determination of crude protein and water in forages (Ruisanchez et al., 2002). The results obtained showed some differences in terms of the outliers detected and the PLS parameters (e.g. root mean square error of prediction, slope). In this sense, this work proposes an interlaboratory study using the same NIR instrument (same manufacturer and model) present in two different laboratories exploring different chemometric tools, namely PCA as an exploratory tool and PLS-DA for the discrimination of *C. japonica* cultivars. Thus, the contribution of different operators and laboratory conditions will be evaluated as well as the ability of NIR spectroscopy to discriminate *C. japonica* cultivars.

2. Material and methods

2.1. Camellia japonica leaves

Leaves of 43 different plants of *C. japonica* belonging to 15 cultivars were collected in two distinct locations (Viveiro da Câmara Municipal do Porto, GPS: 41.155830, −8.558920 and Jardim Botânico do Porto, GPS: 41.153650, −8.642528). Details are presented in Table 1. Ten leaves per plant were collected twice within one month, making a total of 20 leaves per plant. After collection, fresh leaves were transported to the laboratory and air-dried at room temperature, avoiding daylight exposure, until no difference in mass. The air-dried leaves collected for each plant were milled through a coffee mill (MS 50, Taurus, Spain) and sieved. This was repeated for each one of the 43 *C. japonica* plants. The fine powder obtained for each plant was transferred to borosilicate flasks prior spectral acquisition and stored avoiding daylight exposure and at room temperature. Therefore, each flask contains all the leaves collected for each plant (20 leaves per plant).

2.2. Near infrared spectra collection

Near infrared spectra were collected in diffuse reflectance mode, in two Fourier transform near infrared spectrometers (FTLA 2000, ABB, Canada), of the same model present in two different laboratories (Laboratory A and Laboratory B), equipped with an indium-gallium-arsenide (InGaAs) detector, quartz halogen as light source and He-Ne laser. Both instruments were controlled through Bomen-Grams software (version 7, ABB, Canada). At laboratory A (Lab A), each flask was analysed in triplicate and the spectrum of each analysis was obtained within 10,000 and 4000 cm^{-1} with a resolution of 8 cm^{-1} and resulted from an average of 64 scans. Then, the borosilicate flaks were transported to the other laboratory (Laboratory B) in the day after avoiding temperature change and light exposure. At laboratory B (Lab B), the procedure for spectral acquisition performed at Lab A was repeated. The background was performed using the same reference material (Teflon) at both laboratories.

Table 1
Details about the *C. japonica* cultivars leaves included in this study.

Cultivar	N° of plants	Collecting local
Albino botti	1	VMP
	1	JBP
Alba plena	1	VMP
	1	JBP
Augusto leal gouveia pinto	2	VMP
	3	JBP
Bella milanese	2	JBP
Bella portuense	1	VMP
	2	JBP
Camurça	2	VMP
Colletti	2	VMP
	1	JBP
Conde do bonfim	3	JBP
Duchesse de nassau	3	JBP
Etoile polaire	2	JBP
Fimbria alba	2	VMP
Maria irene	2	JBP
Roi des belges	2	JBP
Saudade martins branco	4	VMP
Sophia	1	VMP
	2	JBP

VMP-Viveiro da Câmara Municipal do Porto; JBP – Jardim Botânico do Porto.

2.3. Data analysis

The spectra collected at both laboratories were modelled PCA and PLS-DA. PCA was used to detect outliers and the formation of clusters (Naes et al., 2004). PLS-DA was used as a supervised model to develop discrimination models (Barker and Rayens, 2003). In this case, all samples were classified considering each *C. japonica* cultivar as a different class. Data was divided into two data sets (70% for calibration and 30% for validation) in a random mode but avoiding unbalanced classes in the calibration and validation sets and guaranteeing that all *C. japonica* cultivars were included in both sets. Therefore, from the three spectra obtained for each sample, two spectra were used to calibrate and one spectrum to validate. The leave-one-sample-out procedure was used to estimate the optimum number of latent variables (LVs) using only the calibration set. The optimum number of LVs was selected based on a compromise between the highest percentage of correct predictions and the lowest number of LVs (when the variation in the percentage of correct predictions between two consecutive LVs was lower than 5%, the lowest number of LVs was selected). Aiming to identify the best spectral region, NIR spectra were divided in five spectral regions: R1 from 4956 to 4030 cm^{-1}, R2 from 5376 to 4960 cm^{-1}, R3 from 6076 to 5380 cm^{-1}, R4 from 7466 to 6080 cm^{-1} and R5 from 10,000 to 7470 cm^{-1}. All these regions were tested individually and in all possible combinations, which give a total of 31 possible combinations. Several pre-processing techniques, namely standard normal variate (SNV) and a Savitzky-Golay (SG) filter (x,y,z; where x is the filter width, y is the polynomial order, and z is the

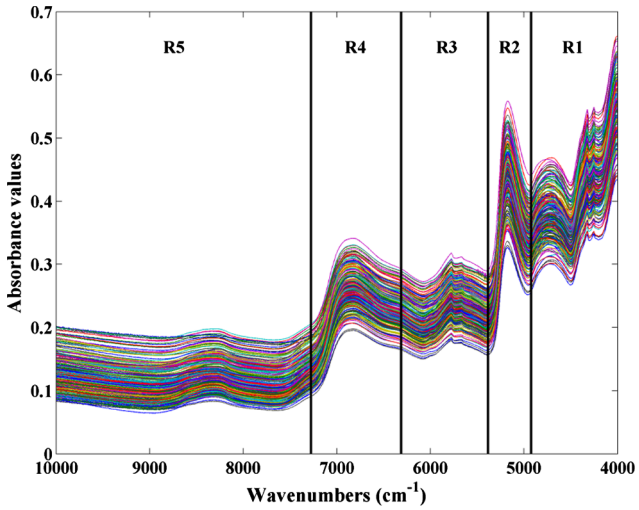

Fig. 1a. *Camellia japonica* air-dried leaves raw spectra collected at Lab A.

derivative used) were tested as well, individually and in all possible combinations to find the best pre-processing technique (creating a total of 8 possible combinations). After finding the best spectral region and re-processing technique, the validation set was projected in the PLS-DA calibration model to assess the percentage of correct predictions for each *C. japonica* cultivar. The predictions of PLS-DA models were expressed in the form of confusion matrices, where the sum of the diagonal elements gives the total percentage of correct predictions (Pascoa et al., 2016). All the calculations were performed in Matlab version R2009b (MathWorks, Natick, USA) and PLS Toolbox version 5.5.1 (Eigenvector Research Incorporated, Manson, USA).

3. Results and discussion

3.1. Unsupervised analysis

Spectra of air-dried *C. japonica* leaves obtained in reflectance mode are depicted in Figs. 1a and 1b, respectively for Lab A and Lab B. Globally, there was no naked-eye differences in the spectra obtained at both laboratories. With the objective of highlighting some possible differences between the spectra obtained at both laboratories, it was decided to plot the average spectra of each laboratory (Fig. 1c). Regarding the average of spectra (Fig. 1c), it can be noticed that the spectra baseline from Lab A has a slightly higher absorbance and that

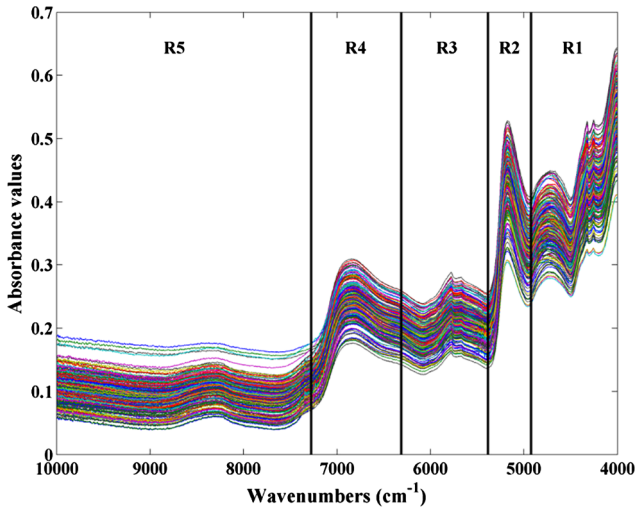

Fig. 1b. *Camellia japonica* air-dried leaves raw spectra collected at Lab B.

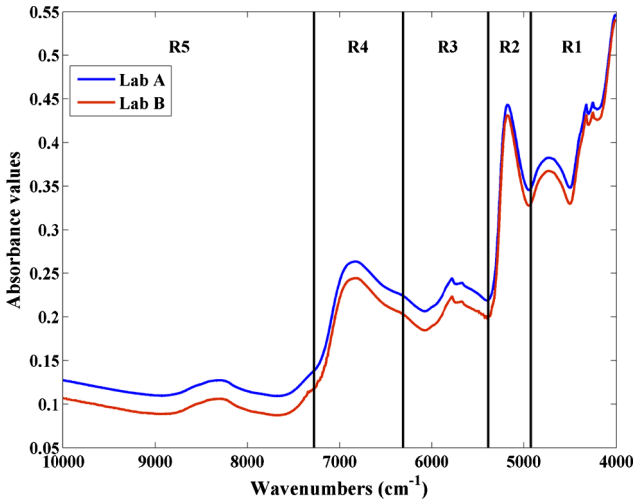

Fig. 1c. Mean spectra of *Camellia japonica* air-dried leaves collected at both laboratories.

the beginning of region R3 of Lab B spectra possess higher noise. The herein obtained differences in spectra baseline of both laboratories were already denoted in other previous studies (Bartl et al., 1996).

The PCA (through Q residuals and Hotelling's T^2 statistics) revealed the presence of some outliers in the spectra of both laboratories, which were further removed. Moreover, the PCA showed no significant differences between the scores obtained at both laboratories (data not shown) pointing to quite similar spectral data.

3.2. Supervised analysis

3.2.1. Optimization of spectral region and pre-processing technique

PLS-DA was applied in data sets from Lab A and Lab B to find the best spectral region and pre-processing technique. The number of LVs was tested until a maximum of 20, using different spectral regions and pre-processing techniques as referred previously. A total of 4960 (31 spectral regions combinations × 8 pre-processing techniques combinations × 20 LVs) PLS-DA models were developed for each laboratory data set. As the number of PLS-DA models developed is vast, only the results obtained for the best pre-processing techniques were shown (Table 2). The best spectral regions and pre-processing techniques were selected based on the highest result of correct predictions obtained. The best spectral regions were R1 and R3, and the combination of both these regions yielded the best results for both laboratories. This was consistent for the majority of the pre-processing techniques tested. Regarding the best pre-processing technique, SG filter (x,2,1) combined with SNV produced the best results (Table 2). Most of the PLS-DA models developed needed more than 15 LVs, however each PLS-DA model developed included 15 different classes (*C. japonica* cultivars).

3.2.2. Discrimination of Camellia japonica cultivars

After the selection of the best spectral region (R1 + R3 for both laboratories) and the best spectral pre-processing technique [Lab A: SG(17,2,1) + SNV and Lab B: SG(19,2,1) + SNV] the discrimination of *C. japonica* cultivars were assessed with spectral data from both laboratories independently.

Regarding Lab A, the optimum developed PLS-DA model with 17 LVs yielded a percentage of correct predictions of 83.4%. The confusion matrix obtained for the PLS-DA model is shown in Table 3. The confusion matrix besides giving the percentage of correct predictions also shows the best and worst *C. japonica* cultivars predicted. From its analysis, it is clear that not all *C. japonica* cultivars were well classified. Indeed, "Bella milanese" cultivar was misclassified with "Bella portuense" and "Roi des belges" cultivars; "Bella portuense" cultivar was

Table 2
Percentage of correct predictions of the best PLS-DA models obtained at Lab A and Lab B with the validation data set.

Spectral region	Pre-processing technique							
	SG (15,2,1) + SNV		SG (17,2,1) + SNV		SG (19,2,1) + SNV		SNV	
	Lab A	Lab B	Lab A	Lab B	Lab A	Lab B	Lab A	Lab B
R1	80.0% (16 LV)	75.7% (13 LV)	81.4% (16 LV)	77.3% (14 LV)	80.8% (16 LV)	77.2% (13 LV)	80.7% (16 LV)	80.0% (16 LV)
R2	57.2% (16 LV)	50.7% (9 LV)	58.9% (15 LV)	50.8% (11 LV)	58.3% (15 LV)	51.1% (11 LV)	62.2% (14 LV)	51.5% (9 LV)
R3	78.5% (16 LV)	69.0% (12 LV)	77.8% (15 LV)	69.2% (14 LV)	77.9% (16 LV)	69.3% (13 LV)	78.7% (16 LV)	79.1% (15 LV)
R4	63.4% (17 LV)	48.5% (11 LV)	63.0% (15 LV)	49.0% (12 LV)	62.1% (16 LV)	50.3% (11 LV)	68.1% (14 LV)	53.7% (12 LV)
R5	54.7% (19 LV)	36.3% (14 LV)	64.3% (14 LV)	38.1% (11 LV)	57.4% (17 LV)	40.7% (14 LV)	66.2% (16 LV)	49.1% (9 LV)
R1 + R3	81.1% (16 LV)	81.0% (15 LV)	83.4% (17 LV)	81.7% (15 LV)	79.6% (15 LV)	82.7% (15 LV)	82.1% (16 LV)	80.8% (16 LV)

misclassified with other cultivars and “Duchesse de Nassau” cultivar was misclassified with “Bella milanese” cultivar. On the other hand, “Conde do bonfim” and “Saudade martins branco” cultivars were well classified with almost 100% of correct predictions.

The best PLS-DA model for Lab B (Table 2) was achieved with 15 LVs yielded a percentage of correct cultivars predictions of 82.7%. The corresponding confusion matrix is shown in Table 4. The results were quite similar to those obtained for Lab A. In this case, the worst prediction involved “Bella portuense” cultivar that was misclassified with other cultivars and “Bella milanese” cultivar was misclassified with “Roi des belges” cultivar. The cultivars that were better predicted were “Colletti” and “Conde do bonfim”.

These results indicate that there are no significant differences between the results obtained at Lab A and Lab B. Therefore, the use of different laboratory conditions and operators seems to have no impact on the results obtained. This is in agreement with the conclusions pointed by Bart and collaborators (Bartl et al., 1996) where it was mentioned that the results obtained by different workers were very similar.

Moreover, the results obtained in the best PLS-DA models for both laboratories (approximately 80% of correct predictions) demonstrate the capacity of NIR spectroscopy to discriminate *C. japonica* cultivars. These results were slightly lower than the results obtained in similar works referred in the introduction section (Gutierrez et al., 2016; Lang et al., 2017; Machado et al., 2018; Páscoa et al., 2018) but in this work several cultivars from different locations were included. In this sense, it is known that different types of soil have a significant impact over the leaves of the plant (Pascoa et al., 2016) and this could somewhat explain the lower percentage of correct predictions.

Regarding the confusion matrices obtained (Tables 3 and 4) there are some congruencies among the best and the worst predicted cultivars for both laboratories. Despite the absence of any genotypic data regarding the cultivars similarity, the congruency among laboratories allow us to hypothesize that the misclassifications are probably among cultivars with higher genotypic and /or phenotypic similarity.

3.2.3. Assessing data transferability between both laboratories

With the objective of assessing data transferability between Lab A and Lab B two different strategies were tested: (I) using data from Lab A as the calibration set and data from Lab B as the validation set and *vice-versa;* and (II) mixing all data together (Lab A + Lab B) and randomly select a calibration and a validation set but ensuring that both calibration and validation sets possess data from both laboratories.

3.2.3.1. Calibration and validation sets from different laboratories. In this analysis, the best pre-processing techniques found previously for Lab A and Lab B sets were maintained but different regions and LVs were tested. The best results obtained are shown in Table 5. It seems that no significant differences arise from using Lab A data to calibrate and Lab B to validate *or vice-versa*. In addition, the results indicate that data sets from both laboratories should be quite different once the percentage of correct cultivars predictions was lower when compared with the previous results. If both data sets were similar, the percentage of correct predictions should also be similar. Even discarding region R3, for possessing higher noise (Lab B data), the results were worse than those obtained independently for each laboratory. The results herein

Table 3
Confusion matrix for the best PLS-DA discrimination model based on the NIR spectroscopy technique applied to leaves samples at Lab A considering both regions R1 and R3 with SG (17,2,1) followed by SNV (83.4% overall correct prediction rate and 17 LVs, n = 252).

Predicted cultivars	Real cultivars														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	3.2	0.6	0.6	0	0.5	0	0	0	0	0.4	0	0	0	0	0
2	0.4	4.0	0	0	0.4	0	0	0.1	0	0.2	0	0	0	0	0.1
3	0.2	0.3	10.5	0.1	0.5	0	0	0	0.2	0	0	0	0.1	0.1	0
4	0	0	0	3.3	0	0	0	0	1.6	0	0	0	0.2	0	0
5	0.6	0.1	0.1	1.2	3.5	0	0	0	0.3	0.4	0	0.5	0.1	0	0.2
6	0	0.4	0	0	0.1	6.3	0	0	0	0	0	0	0	0.1	0
7	0	0	0	0	0	0	6.9	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	10.2	0	0	0	0	0.2	0	0
9	0	0	0	0.3	0.4	0.5	0.5	0	3.3	0.1	0	0.1	0	0	0.1
10	0.1	0	0	0	0	0	0	0	0	3.3	0	0	0	0	0
11	0	0	0	0	0.1	0	0	0	0	0	6.8	0	0	0	0
12	0	0	0	0.3	0.2	0	0	0	0	0	0	4.7	0	0	0
13	0.2	0	0.6	0.9	0	0	0	0	0.3	0	0.4	0	2.7	0	0.1
14	0	0.3	0	0.1	0.1	0	0	0	0	0	0	0	0	9.9	0
15	0	0	0	0.1	0.2	0	0	0	0	0	0	0	0.1	0	4.9

Legend: 1 – “Albino botti”; 2 – “Alba plena”; 3 – “Augusto leal gouveia pinto”; 4 – “Bella milanese”; 5 – “Bella portuense”; 6 – “Camurça”; 7 – “Colletti”; 8 – “Conde do bonfim”; 9 – “Duchesse de nassau”; 10 – “Etoile polaire”; 11 – “Fimbria alba”; 12 – “Maria irene”; 13 – “Roi des belges”; 14 – “Saudade martins branco”; 15 – “Sophia”; Values are in %.

Table 4
Confusion matrix for the best PLS-DA discrimination model based on the NIR spectroscopy technique applied to leaves samples at Lab B considering both regions R1 and R3 with SG (19,2,1) followed by SNV (83.4% overall correct prediction rate and 17 LVs, n = 252).

Predicted cultivars	Real cultivars														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	2.5	0	0.2	0	1.0	0.7	0	0	0	0.8	0	0	0	0	0
2	0.3	4.8	0	0	0.1	0	0	0	0	0	0	0	0	0	0
3	0.1	0.8	10.2	0	0	0	0	0	0	0	0.4	0	0.1	0.3	0
4	0	0	0.1	3.4	0.2	0	0	0	0.9	0	0	0	0.5	0	0
5	0.2	0	0.1	1.0	3.3	0	0	0	0.2	0.6	0	0.1	0	0.7	0.7
6	0.1	0.6	0	0	0	5.8	0	0	0.1	0	0.2	0	0	0.1	0
7	0	0	0	0	0	0	6.9	0	0	0	0	0	0	0	0
8	0	0	0	0	0.1	0	0	9.9	0	0	0	0.3	0	0	0
9	0	0	0.4	0.1	0.2	0.2	0	0	4.0	0	0	0.2	0	0	0
10	0.3	0	0	0	0.1	0	0	0	0	3.1	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	6.9	0	0	0	0
12	0	0	0	0	0.3	0	0	0	0	0.4	0	4.4	0	0	0
13	0	0	1.1	0.8	0.1	0	0	0	0.2	0	0	0	3.0	0	0
14	0	0	0	0	0.1	0.1	0	0	0	0	0	0	0	10.2	0
15	0	0	0	0	0.5	0.1	0	0.1	0.4	0	0	0	0	0	4.1

Legend: 1 – “Albino botti”; 2 – “Alba plena”; 3 – “Augusto leal gouveia pinto”; 4 – “Bella milanese”; 5 – “Bella portuense”; 6 – “Camurça”; 7 – “Colletti”; 8 – “Conde do bonfim”; 9 – “Duchesse de nassau”; 10 – “Etoile polaire”; 11 – “Fimbria alba”; 12 – “Maria irene”; 13 – “Roi des belges”; 14 – “Saudade martins branco”; 15 – “Sophia”; Values are in %.

Table 5
Percentage of correct predictions of PLS-DA models obtained with calibration and validation sets from different laboratories.

Spectral regions	Calibration data: Lab A	Calibration data: Lab B
	Validation data: Lab B	Validation data: Lab A
R1	59.2% (6 LV)	66.1% (10 LV)
R3	40.8% (4 LV)	54.6% (10 LV)
R1 + R3	65.5% (6 LV)	63.8% (7 LV)

obtained reflect the same problems reported by other authors (Bartl et al., 1996; Yoon et al., 2000) being in this case even more problematic since the instruments used belong to the same manufacturer, model and sampling accessories. An explanation to this can be found in (Bouveresse and Massart, 1996) where the authors refer that differences in the age of the instruments can be responsible for the differences found in instrumental response. The results of this study point that the possibility of building an universal NIR spectral library as suggested by some authors (Bouveresse and Massart, 1996) could be a very complex task as advised by others (Yoon et al., 2000).

3.2.3.2. Calibration and validation with all data. Another interesting approach was to combine all data obtained at both laboratories in just one matrix and randomly select a calibration and validation data sets. In this case, the best pre-processing techniques found previously at Section 3.2.1 were tested, as well as the best spectral regions (R1, R3 and R1 combined with R3) and LVs (up to 20). The combination of regions R1 and R3 with the application of SG (19,2,1) followed by SNV and 18 LVs yielded the best PLS-DA model with 85.9% of correct predictions (Table 6). This result is in agreement with the results obtained in Section 3.2.2 in terms of the best processing technique, spectral regions and percentage of correct predictions. However, it does give any indication if the spectra obtained at Lab A and Lab B laboratory can be considered similar or different. The confusion matrix is shown in Table 7. In this case, the worst predictions involved “Bella portuense” cultivar, which was misclassified with “Albino botti” cultivar, and the best predictions were obtained with “Conde do bonfim” and “Saudade martins branco” cultivars. This is in agreement with the results found when using only the spectra obtained at Lab A.

4. Conclusions

The results obtained in this work (around 85% of correct predictions for the validation sets of both laboratories) demonstrate that NIR spectroscopy coupled to PLS-DA can be a reliable tool to discriminate *C. japonica* cultivars. Despite the results obtained in Lab A and Lab B being similar (in terms of correct predictions), when the PLS-DA models calibrations were performed with spectra obtained in one laboratory and validated with the spectra obtained in the other, the results get worst. These findings indicate that even using an instrument of the same manufacturer and model it will be difficult to transfer the spectral library. This reinforces the idea that to build a universal spectral library using NIR technique can be quite challenging and several precautions should be previously considered (e.g. the instrument manufacturer, the sampling mode as well as the maintenance and the age of the instruments).

Acknowledgements

This work received financial support from the European Union (FEDER funds POCI/01/0145/FEDER/007265 and POCI/01/0145/FEDER/016735) and National Funds (FCT/MEC, Fundação para a

Table 6
Percentage of correct predictions of PLS-DA models obtained using all data.

Spectral regions	Pre-processing techniques			
	SNV + SG (15,2,1)	SNV + SG (17,2,1)	SNV + SG (19,2,1)	SNV
R1	78.1% (12 LV)	78.1% (12 LV)	82.1% (18 LV)	81.0% (13 LV)
R3	71.1% (12 LV)	71.0% (12 LV)	70.8% (12 LV)	73.0% (12 LV)
R1 + R3	83.2% (15 LV)	83.5% (15 LV)	85.9% (18 LV)	85.2% (17 LV)

Table 7
Confusion matrix for the PLS-DA discrimination model based on the NIR spectroscopy technique applied to leaves samples using all the data combined (Lab A and Lab B) considering both regions R1 and R3 with SG (19,2,1) followed by SNV (85.9% overall correct prediction rate and 18 LVs, n = 504).

Predicted cultivars	Real cultivars														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	3.0	0.7	0.4	0	1.0	0.1	0	0	0	0.2	0	0	0	0	0
2	0.1	4.9	0	0	0.3	0	0	0	0	0.2	0	0	0	0	0
3	0.3	0.7	10.1	0	0.2	0	0.1	0	0.3	0	0.1	0	0	0	0
4	0	0	0.2	4.1	0.1	0	0	0	0.8	0	0	0	0.3	0	0
5	0.3	0	0	1.0	3.9	0	0	0	0.4	0.1	0	0.9	0	0.1	0.5
6	0	0.2	0	0	0	7.1	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	7.2	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0
9	0.1	0	0	0	0.4	0.1	0.1	0	4.2	0	0	0.5	0	0	0
10	0.2	0	0	0	0	0	0	0	0	3.4	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	7.3	0	0	0	0
12	0.1	0	0	0	0.5	0	0	0	0	0	0.3	4.5	0	0	0
13	0	0	0.4	0.8	0.2	0	0	0	0	0	0.1	0	3.9	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	9.1	0
15	0	0	0	0	0.1	0	0	0	0	0	0	0	0.1	0	3.4

Legend: 1 – “Albino botti”; 2 – “Alba plena”; 3 – “Augusto leal gouveia pinto”; 4 – “Bella milanese”; 5 – “Bella portuense”; 6 – “Camurça”; 7 – “Colletti”; 8 – “Conde do bonfim”; 9 – “Duchesse de nassau”; 10 – “Etoile polaire”; 11 – “Fimbria alba”; 12 – “Maria irene”; 13 – “Roi des belges”; 14 – “Saudade martins branco”; 15 – “Sophia”; Values are in %.

Ciência e Tecnologia and Ministério da Educação e Ciência) under the Partnership Agreement PT2020 UID/QUI/50006/2013 and through project PTDC/AGR-PRO/6817/2014. Clara Sousa was funded through the NORTE-01-0145-FEDER-000024—“New Technologies for three Health Challenges of Modern Societies: Diabetes, Drug Abuse and Kidney Diseases”. Ricardo Páscoa and Cristina Quintelas also thank FCT for the financial support.

References

Barker, M., Rayens, W., 2003. Partial least squares for discrimination. J. Chemometr. 17, 166–173.
Bartl, F., Delgadillo, I., Davies, A.N., Huvenne, J.P., Meurens, M., Volka, K., Wilson, R.H., 1996. An interlaboratory comparison of sample presentation methods for the analysis of aqueous solutions using Fourier transform infrared spectroscopy. Fresenius J. Anal. Chem. 354, 1–5.
Bouveresse, E., Massart, D.L., 1996. Standardisation of near-infrared spectrometric instruments: a review. Vib. Spectrosc. 11, 3–15.
Gutierrez, S., Tardaguila, J., Fernandez-Novales, J., Diago, M.P., 2016. Data mining and NIR spectroscopy in viticulture: applications for plant phenotyping under. Field Cond. Sens. 16.
Kanth, B.K., Lee, K.Y., Lee, G.J., 2014. Antioxidant and radical-scavenging activities of petal extracts of Camellia japonica ecotypes. Hortic. Environ. Biotechnol. 55, 335–341.
Lang, C., Almeida, D.R.A., Costa, F.R.C., 2017. Discrimination of taxonomic identity at species, genus and family levels using Fourier Transformed Near-Infrared Spectroscopy (FT-NIR). For. Ecol. Manage. 406, 219–227.
Machado, J.C., Faria, M.A., Ferreira, I., Pascoa, R., Lopes, J.A., 2018. Varietal discrimination of hop pellets by near and mid infrared spectroscopy. Talanta 180, 69–75.
Naes, T., Isaksson, T., Fearn, T., Davies, T., 2004. Interpreting PCR and PLS solutions. User-Friend. Guide Multivariate Calib. Class. 1, 39–54.
Oh, G.S., Kang, S.S., Chung, M.G., 1996. Temporal genetic structure in Camellia japonica (Theaceae). Genes Genet. Syst. 71, 9–13.
Pascoa, R., Lopo, M., dos Santos, C.A.T., Graca, A.R., Lopes, J.A., 2016. Exploratory study on vineyards soil mapping by visible/near-infrared spectroscopy of grapevine leaves. Comput. Electron. Agric. 127, 15–25.
Páscoa, R.N., Moreira, S., Lopes, J.A., Sousa, C., 2018. Citrus species and hybrids depicted by near-and mid-infrared spectroscopy. J. Sci. Food Agric.
Ruisanchez, I., Rius, F., Maspoch, S., Coello, J., Azzouz, T., Tauler, R., Sarabia, L., Ortiz, M.C., Fernandez, J.A., Massart, D., Puigdomenech, A., Garciia, C., 2002. Preliminary results of an interlaboratory study of chemometric software and methods on NIR data. Predicting the content of crude protein and water in forages. Chemometrics Intell. Lab. Syst. 63, 93–105.
Salinero, C., Feas, X., Mansilla, J.P., Seijas, J.A., Vazquez-Tato, M.P., Vela, P., Sainz, M.J., 2012. H-1-nuclear magnetic resonance analysis of the triacylglyceride composition of cold-pressed oil from camellia japonica. Molecules 17, 6716–6727.
Vela, P., Salinero, C., Sainz, M.J., 2013. Phenological growth stages of Camellia japonica. Ann. Appl. Biol. 162, 182–190.
Yoon, W.L., Jee, R.D., Moffat, A.C., 2000. An interlaboratory trial to study the transferability of a spectral library for the identification of solvents using near-infrared spectroscopy. Analyst 125, 1817–1822.