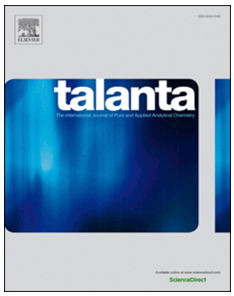
Accepted Manuscript



Development of a methodology to analyze leaves from *Prunus dulcis* varieties using near infrared spectroscopy

Sergio Borraz-Martínez, Ricard Boqué, Joan Simó, Mariàngela Mestre, Anna Gras

PII: S0039-9140(19)30599-5

DOI: https://doi.org/10.1016/j.talanta.2019.05.105

Reference: TAL 19990

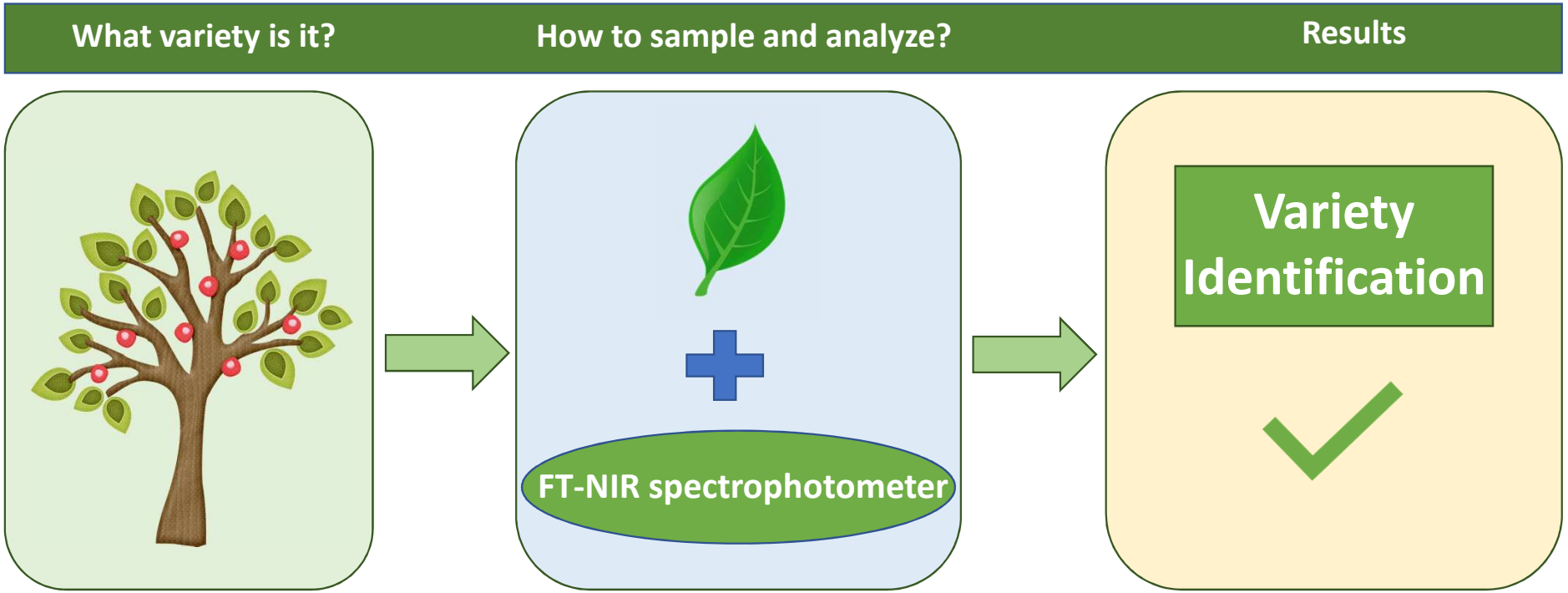
To appear in: *Talanta*

Received Date: 22 March 2019

Revised Date: 21 May 2019

Accepted Date: 27 May 2019

Please cite this article as: S. Borraz-Martínez, R. Boqué, J. Simó, Marià. Mestre, A. Gras, Development of a methodology to analyze leaves from *Prunus dulcis* varieties using near infrared spectroscopy, *Talanta* (2019), doi: https://doi.org/10.1016/j.talanta.2019.05.105.


What variety is it?
How to sample and analyze?
Results
FT-NIR spectrophotometer
Variety Identification

DEVELOPMENT OF A METHODOLOGY TO ANALYZE LEAVES FROM *PRUNUS DULCIS* VARIETIES USING NEAR INFRARED SPECTROSCOPY

Sergio Borraz-Martínez [a,c], Ricard Boqué [b], Joan Simó [a,d], Mariàngela Mestre [c], Anna Gras [a]

[a] *Universitat Politècnica de Catalunya, Department of Agri-Food Engineering and Biotechnology, Esteve Terrades 8, 08860 Castelldefels, Spain*
[b] *Universitat Rovira i Virgili, Department of Analytical Chemistry and Organic Chemistry, Campus Sescelades, 43007 Tarragona, Spain*
[c] *Agromillora Iberia S.L.U, Center of Initial Materials, Ctra. BV-2247 km. 3, 08770-Sant Sadurní d'Anoia*
[d] *Fundació Miquel Agustí, Esteve Terrades 8, 08860 Castelldefels, Spain*

Corresponding author: S. Borraz-Martínez: **sergio.borraz@upc.edu**

ORCID numbers of the authors: Sergio Borraz-Martínez: 0000-0002-5607-9462; Ricard Boqué: 0000-0001-7311-4824; Joan Simó: 0000-0002-2853-3381; Anna Gras: 0000-0003-0111-7246

Abstract

Near-infrared spectroscopy (NIRS) can be a faster and more economical alternative to traditional methods for screening varietal mixtures of nursery plants during the propagation process to ensure varietal purity and to avoid errors in the dispatch batches. The global objective of this work was to develop and optimize a NIR spectral collection method for construction of robust multivariate discrimination models. Three different varieties of *Prunus dulcis* (*Avijor*, *Guara*, and *Pentacebas*) of agricultural interest were used for this study. Sources of variation were investigated, including the position of the leaves on the trees, differences among trees of the same variety, and differences at the varietal level. Three types of processed samples were investigated. Fresh leaves, dried leaves, and dried leaves in powder form were included in each analysis. A study of spectral pre-treatment methods was also performed, and multivariate methods were applied to analyze the influence of different factors on classification. These included principal component analysis (PCA), partial least squares discriminant analysis (PLS-DA), and ANOVA simultaneous component analysis (ASCA). The results indicated that variety was the most important factor for classification. The spectral pre-treatment that provided the best results was a combination of standard normal variate (SNV), Savitzky-Golay first derivative, and mean-centering methods. With regard to the type of processed sample, the highest percentages of correct classifications were obtained with fresh and dried powdered leaves at both the training set and test set validation levels. This study represents the first step towards the consolidation of NIRS as a method to identify *Prunus dulcis* varieties.

Keywords: Optimization; Almond trees; Leaf analysis; Varietal purity; NIR; PLS-DA.

1. INTRODUCTION

Rapid discrimination between vegetal varieties is a key requirement for all nursery plant production. The huge diversity of vegetal materials necessitates the incorporation of new control systems along the nursery plant production chain to avoid mixing varieties and to ensure varietal purity in the dispatch batches.

Nowadays, the most extensively used methods for varietal identification are based on DNA analysis. These techniques include DNA amplification by the polymerase chain reaction (PCR) [1] followed by analysis of genetic variations, such as single nucleotide polymorphisms (SNPs) [2]. However, these biomolecular techniques are very expensive for routine analysis of a large number of samples. In this context, the use of spectroscopic analysis combined with chemometrics has recently increased. This combination comprises a rapid, accurate, and nondestructive methodology for the classification and authentication of agricultural products [3].

Near-infrared (NIR) spectroscopy has proved to be a powerful analytical tool and has been widely used in various sectors, including the petrochemical [4] and pharmaceutical industries [5]. It has also become a well-established technique for the quantitative and qualitative analysis of agricultural products [6]. Several recent studies have employed spectroscopic techniques for species discrimination [7,8], or differentiation of varietals within a species, such as tomato [9], rice [10] and lettuce [11]. For these reasons, NIRS can be considered a potential candidate for the differentiation of *Prunus dulcis* varieties.

Despite recent studies, there is a lack of knowledge regarding the best methodology for accurate sampling of leaves. Most of the published works on species discrimination do not consider factors derived from the nature of the samples, which are potential sources of variance. For example, mature trees have a heterogeneous canopy composed of leaves in different phenological stages. It is thus important to take the sampling procedure into account, especially when an analysis is performed with whole leaves. Improper sampling may generate invalid data, the use of which could lead to incorrect conclusions [12]. To perform correct sampling, it is important to recognize sources of variation and to control for factors from which variation originates. Therefore, it is necessary to first develop a sampling protocol and to select the best material for use.

Another analytically relevant aspect is the study of sample processing methods, which may considerably alter the vibrational spectrum of a sample compared to that collected with the sample in its native state. Due to economic and time constraints, it is generally best to avoid any type of sample processing. Moreover, modifying the native architecture of biological tissues can result in

the loss of information. Thus, performing analyses in vivo is preferred whenever possible [13]. Occasionally, however, sample processing is an indispensable step. In any case, the option that best accomplishes the objective of the study must be selected.

The aim of this work is to determine how sampling of vegetal material affects the collection of NIR spectra for the construction of a multivariate discriminant model for *Prunus dulcis* varietal classification. The specific objectives are to 1) determine whether there are differences among the analyzed regions of the leaves or between their upper and lower surfaces; 2) to determine whether differences exist due to the age of the leaves; 3) identify the best sampling procedure for varietal discrimination of almond trees; and 4) study pre-treatment of spectral data and to identify the pre-treatment that leads to the best classification model.

2. MATERIAL AND METHODS

2.1 Experimental design

2.1.1 Assay one

The first assay was performed to obtain information about the analyzed regions of fresh and dried leaves. Specifically, the NIR spectra differences resulting from including or excluding the primary veins of the leaves were examined (**Error! Reference source not found.**) together with analysis of the upper (adaxial) and lower (abaxial) leaf surfaces. Twenty samples from two varieties of almond trees, *Guara* and *Pentacebas*, were used for each experiment (Table 1). Results were evaluated by using PCA and PLS-DA models. This assay focused on aspects that affected only fresh and dried leaves. The information obtained in this assay was used for the development of the next two assays.

[Insert Fig.1]

2.1.2 Assay two

The second assay was designed to study the NIR spectra differences between young leaves and adult leaves and among samples from different trees of the same variety. The assay was performed on the *Guara* and *Pentacebas* varieties and on a third variety, *Avijor*. Four trees per variety were sampled, twelve in total. Twenty leaves were collected from each tree. Ten of the leaves were collected from the upper part of the branch (apex), which corresponded to young leaves, while the other ten were adult leaves that were collected from the lower part of the branch. Two hundred forty leaves were sampled in total (Table 1). Results were evaluated by using PCA and ASCA-ANOVA models.

2.1.3 Assay three

Three different leaf processing methodologies were studied in the third assay; one for fresh leaves, one for dried leaves, and the other for dried powdered leaves. The aim was to determine whether the water content and macrostructures of the leaves had any influence on discrimination results. It is important to note that every sample was processed with each of the three methods in order to increase the comparative robustness. We also identified the most suitable pre-treatment method for NIR spectral analysis. The applicability of NIRS for discriminating between *Prunus dulcis* varieties was evaluated by mean partial least squares discriminant analysis (PLS-DA). The available data were randomly divided into calibration (70%) and validation (30%) sets, but both sets contained the same proportion of each variety to prevent unbalanced representation of the almond tree classes. To improve the robustness of comparing results from the three sample processing methods, the same samples included in the three sample processing datasets were used for both cross validation and test set validation. All of the samples used for assay two were also used for this assay (Table 1).

[Insert Table 1]

2.2 Description of the sampling field

Vegetal material used in this study came from almond trees located at the mother plant field from the Center of Initial Materials of Agromillora Iberia, S.L.U. in Sant Sadurní d'Anoia (Catalonia, Spain). These trees are under a strict control in order to prevent the appearance of diseases and to ensure the sanitary quality of nursery plants. The use of molecular biology techniques to assess the traceability of the varieties was not necessary in this case because the almond trees were previously certificated by the company.

The samples were stored in a plastic bag after collection, assigned identifiers, and stored at 4 °C until analysis.

2.3 Sample pre-processing

Samples were analyzed either as fresh leaves without processing, as dried leaves, or as dried powdered leaves. To obtain dried leaves, fresh leaves were heated in an oven at 65 °C for 48 hours. A weight was placed on the leaves to keep them flat and to facilitate their posterior analysis. Once dried, the leaves were pulverized to a homogeneous powder with a grinder. Once samples were dried, they were stored in a desiccator with silica gel to prevent any influence from moisture. Only one leaf was used per experiment. Each sample was analyzed in the three ways. First, they were

analyzed in fresh, second in dried and finally in powdered. In all the experiments each sample was composed of one leaf only.

2.4 Acquisition of NIR spectra

Samples were scanned in reflectance mode using an Antaris II FT-NIR analyzer (Thermo Scientific, USA) equipped with an integrating sphere module. Samples were measured in the spectral range of 12000–3800 cm^{-1} (833–2630 nm). For each spectrum, 32 scans were averaged with a resolution of 4 cm^{-1}. Each sample was analyzed in triplicate. Fresh leaves and dried leaves were placed directly over the sphere and covered to prevent interference from environmental light. The powdered leaf samples were measured in a standard sample cup that came with the instrument. A background spectrum was collected every 20 minutes. All spectra were recorded as log(1/R), where R was the reflectance. Room temperature was maintained at ~25 °C, and the humidity remained constant throughout the spectral acquisition process.

2.5 Spectral data pre-treatment

This was an important step, because although different pre-treatments have been reported on extensively [14–16], there is still no clear consensus regarding the best pre-treatment or a guideline to follow. As can be seen in **Error! Reference source not found.**, the spectra contained very little noise. The raw spectra had to be corrected for additive and multiplicative effects that were probably due to light scattering.

[Insert Fig. 2]

A basic pre-treatment was performed in assays one and two, which consisted of the standard normal variate (SNV) method with mean centering. In the assay three, four different pre-treatments were applied and compared to identify the combination that provided the best results in the PLS-DA model. The combinations used were: SNV method with mean centering; SNV method with Savitzky-Golay (SG) first derivative and mean centering; and finally, SNV method followed by de-trending and mean centering. Spectral pre-treatments were performed using PLS_Toolbox (Eigenvector Research Incorporated, Manson, WA) with MATLAB R2017b (MathWorks, Natick, MA).

SNV is a normalization procedure for spectral light scattering correction. It is used to correct additive and multiplicative effects in the spectra due to particle size variation. SNV calculates the standard deviation of all the variables in a given sample spectrum. The entire data set is then normalized by this value, which yields a unit standard deviation (s = 1) for the sample spectrum

[17]. De-trending is sometimes used to remove constant, linear, or curved offsets and is often used in conjunction with SNV. With this method, the mean value or linear trend is subtracted from a vector or matrix. To achieve this, a polynomial of a given order is fitted to the entire data set, and the polynomial is simply subtracted. This algorithm fits all points in the baseline and the signal. [17]. SG first derivative was applied to remove baseline drift and to enhance small spectral differences. The SG derivative method includes a smoothing step, the Savitzky-Golay algorithm, which corrects for the increased noise due to application of the derivative. The SG derivatization algorithm requires selection of the filter width, which is the size of the window, the order of the polynomial, and the order of the derivative [18]. In this work, we selected a 15-point window and applied a second order polynomial. Mean centering is one of the most common pre-processing methods, in which the mean value of each column is calculated and subtracted from each individual value in the column. After mean centering, the mean of each column equals zero, and each row of mean-centered data reflects only how it differs from the average sample in the original data matrix [16].

2.6 Principal component analysis (PCA)

PCA captures the largest amount of variance in the data and reduces the dimensionality of the original dataset through calculation of a new set of variables called principal components (PCs). The PCs are linear combinations of the original variables. Samples and variables are projected onto the new PCs in the calculated PCA space. Samples are defined by their scores, and variables are defined by their loadings. Inspection of the scores and loading plots can lead to a better understanding of the different sources of variation in the data. As a data reduction technique, PCA is frequently the first step in the analysis of a high-dimensional data set. It can then be followed by classification, clustering, or other multivariate techniques [19].

2.7 Partial Least Squares Discriminant Analysis (PLS-DA)

PLS-DA is a classification technique widely used in research studies concerning both varietal classification and authentication of geographical origin [10,20]. PLS-DA is based on the PLS regression algorithm, which searches for linear combinations of the original variables (latent variables) that display maximum covariance with the Y-variables (classes). A discriminator, or threshold, is created that separates the different classes [21]. This technique allows determination of whether or not a given sample belongs in a specific predefined class [22]. The optimal number of factors or latent variables (LVs) for the PLS-DA models was estimated with a cross-validation procedure, and the number yielding the minimum classification error was selected. Venetian blinds

cross validation was used for the calibration with a data split of 10 and one sample per blind (thickness).

2.8 ASCA-ANOVA

Designed experiments with a single dependent variable are typically analyzed with ANOVA [23]. Problems occur when hundreds or thousands of variables are measured simultaneously, which is the case in spectroscopic analysis. ANOVA is thus not useful for analyzing multivariate data. Multivariate ANOVA (MANOVA) [24], the natural multivariate extension of ANOVA, breaks down when the number of measurements is smaller than the number of variables [25].

ANOVA-simultaneous component analysis (ASCA) [26] is a method used to determine which factors in a fixed-effect experimental design are significant relative to the residual error. ASCA allows an ANOVA-like analysis, even when there are more variables than samples. Two matrices are used to perform the procedure. The X-matrix contains the experimental data, while the F-matrix represents the experimental design. PCA of each factor in the effect (X) matrix reduces the number of variables to a smaller number of principal components. In this way, the parameter estimation functionality of ANOVA is merged with PCA, and the presence of more variables than samples is no longer problematic [27]. Due to the hierarchy of factors analyzed in the present study, a nested design referred to as multi-level simultaneous component analysis (MLSCA) [28] was applied. Hence, the leaf age factor was nested within the tree factor, which in turn was nested within the variety factor.

3. RESULTS AND DISCUSSION

3.1 Assay one

3.1.1 Comparison of leaf midvein and lamina

Whether differences exist within the same leaf is a question that frequently arises. For this reason, spectra were collected in different areas of healthy leaves. The two regions of the leaves used for comparison are shown in Fig. 1. PCA was performed with two of the almond tree varieties, *Guara* and *Pentacebas*, to identify possible differences between the measurement areas on fresh and dried samples. These results are shown in Fig. 2.

[Insert Fig. 3]

Differences when including or not the primary vein were detected. The data clouds with and without midvein form separate clusters in both kinds of pre-processed samples. This cluster

separation can be observed in both varieties, although the separation is clearer for the Pentacebas variety. Differences were detected whether or not the primary vein was included. The data clouds with and without the midvein formed separate clusters for both processed sample types. This cluster separation was observed with both varieties, although the separation was more pronounced in the results from the *Pentacebas* variety. Considering the macrostructures and compositions of the analysis regions were not equivalent, which was reflected in their spectral signatures, these differences were justifiable. When the primary vein was scanned, the reflectance spectra of both the primary vein and the laminar regions located on either side of the primary vein were collected. Taking into account that secondary veins were present in the laminar regions, identifying differences between these regions indicated the primary vein had a profound influence on the spectra.

The apical region and a region adjacent to the leaf margin showed more damage and decay than the central region of the leaves. Consequently, the central region was usually more stable. The leaf size could make it difficult to completely exclude the primary vein during measurement of the laminar region. Collecting spectra in the central region, including the primary vein, could therefore provide a standardized measure.

3.1.2 Comparison of adaxial and abaxial surfaces

Differences between the upper and lower surfaces of the leaves were also investigated. These results are shown in **Error! Reference source not found.**. In both fresh and dried samples, results of PCA revealed differences between the spectra obtained from the upper and lower leaf surfaces. However, this difference was not as clear in fresh leaves of the *Pentacebas* variety. The upper and lower surfaces of leaves in all plants are different. In addition, the stomas are usually present on the abaxial surface together with trichomes and others surface features. The differences between these two surfaces could be the cause for separation of their spectra in the PCA plots.

[Insert Fig. 4]

A PLS-DA model was built to determine which surfaces were most suitable for discriminating between two almond tree varieties using fresh or dried samples. The classification results are shown in Table 2. The PLS-DA model had a classification score of 100% for both types of processed samples when the upper leaf surface was analyzed. Perfect discrimination was obtained using the lower leaf surface as well. Based on these results, the differences identified by PCA did not affect the discrimination results with either surface.

[Insert Table 2]

3.2 Assay two

3.2.1 Variability between trees of the same variety

Differences among trees of the same variety are important to consider when building a classification model. This source of variation determines the number of trees of each variety that must be sampled for development of the final model. If the variance is very large, it could affect the model's discrimination capability. The PCA results from assay 2 are shown in Fig 4.

[Insert Fig. 5]

No differences were identified among the four trees studied within each variety. This was the case for fresh, dried, and dried powdered leaves. This was remarkable, because if significant differences were found, it would have been more difficult to build a good classification model. Also noteworthy was that the same results were obtained with samples processed with the three different methods, and with samples of different varieties. Such similar behavior in all cases is a positive indicator when creating a classification model. A more exhaustive study of the variability between trees was performed using the ASCA-ANOVA method, which is discussed in section 3.2.3.

3.2.2 Variability between leaves of the same variety

Since differences among almond trees of the same variety were not detected at the PCA level, we decided to include all samples of the same variety in a single PCA model. This made it easier to study the variability among samples within each variety while increasing the robustness of the model with more samples. The results of PCA modelling are shown in **Error! Reference source not found.**.

[Insert Fig. 6]

Two clusters could be distinguished using only the first two principal components. This separation was very clear in some cases, such as the dried processed samples of the *Pentacebas* variety, for which the two clusters were completely separated (Fig. 6f). The results of all of the PCA models were similar, regardless of the sample processing method or the variety studied. However, overlap between the two data clusters was observed in some cases, such as dried samples of the *Avijor*

variety (Fig. 6d). The overlap could be explained by the presence of leaves in a phenological stadium intermediate between young and adult. It was possible to observe the progressive growth of the leaves, although this was not the goal of the assay. In any case, the results indicated there were differences between young and adult leaves at the spectral level. This difference should be considered at the time of sampling.

3.2.3 ASCA-ANOVA analysis

To study variability between *Prunus dulcis* varieties more deeply, an ASCA-ANOVA model was constructed for young and adult leaves from trees of the same variety. The modelling results are shown in **Error! Not a valid bookmark self-reference.**. The raw spectra pre-treatment used to develop the model, SNV with mean centering, was the same as that used for the PCA models.

[Insert Table 3]

Tree variety was the most influential factor for variance among fresh and dried powdered leaves and accounted for 30.26% and 24.99%, respectively, of the total effect in these samples. Despite explaining 19.25% of the effect for dried leaves, tree variety was not the factor that accounted for the majority of variance. For two of the three processing methods, the variety factor had the greatest effect, which indicated that differences between varieties were important. The tree factor explained little of the variance for the three processing methods, which was in agreement with the PCA results shown in Fig 4. This indicated strong homogeneity between trees of the same variety, an aspect that could be key for effective discrimination between varieties. For fresh and dried powdered leaves, the age (young/adult) factor explained a higher percentage of variance than the tree factor, but it accounted for less of the variance than tree variety. In dried powdered leaves, the difference between the age and tree factors was not large. The age factor accounted for 6.68% of the variance, while the tree factor explained 1.87%. The difference was more notable for fresh leaves, as the age factor accounted for 19.11% of the explained variance. The age factor was most significant for dried leaves, accounting for 24.18% of the explained variance. Therefore, the age factor had a greater effect in non-powder samples. These results also correlated with the results of the PCA (**Error! Reference source not found.**), in which differences due to leaf age were observed, but overlap of the cluster regions was detected.

All of the variance not explained by the studied factors accumulated in the residual term. In the three types of processed samples, the residual accounted for a high percentage of the variance.

Fresh leaves had a lower residual than either the dried or dried powdered leaves. It was thought that the main source of uncontrolled variance was the physiological state of the leaves, which included damage to the leaves and climatologic agents. The combination of these abiotic factors with biotic factors influences plant physiology [29,30]. It is important to note that the leaves used in this study came from trees located in an outdoor field.

3.3 Assay three

3.3.1 Spectral pre-treatment study

[Insert Table 4]

Error! Reference source not found. shows the results of the PLS-DA modelling using different spectral pre-treatments. The best classification results for the three types of samples were obtained with the SNV pre-treatment and application of the SG first derivative and mean centering. This was curious, because although modelling was performed for one material (almond tree leaves), the samples analyzed were completely different in terms of their macrostructures and dry compositions. With this spectral pre-treatment, 100% classification accuracy was achieved for at least one variety with each sample processing method. Results were even more remarkable with dried powdered leaves, for which 100% accuracy was attained in the test set validation for all three varieties. The lowest accuracy obtained with this spectral pre-treatment was 97.5% at both the cross-validation and test set validation levels. No relevant differences between the other two spectral pre-treatments were observed, so de-trending did not appear to have a significant effect. It is important to note that in the case of fresh leaves, similar results were obtained with the three different spectral pre-treatments.

3.3.2 Sample processing study

Each sample processing method had its advantages and disadvantages. Fresh leaves did not require any processing, so measurement was faster and easier than it was with the other types of samples. However, the water content of the leaves was a disadvantage, because it generated wide bands in the NIR spectra. This could make discrimination between varieties more difficult. Samples can be dehydrated to circumvent the effects of water, but this process is time-consuming (48 h), so it is not the best option if rapid identification is required.

To evaluate which of the processed samples was the most suitable for varietal classification, the advantages and disadvantages of each were considered together with the PLS-DA classification

results obtained with SNV spectral pre-treatment and application of the SG first derivative and mean centering (Table 4).

The results obtained with the three types of sample processing at the calibration level could be considered quite good, although those obtained with fresh leaves were less stellar. The dried powdered leaves provided a higher percentage of correct classifications. For the test set validation, high percentages of correct classifications were obtained with all varieties and processed sample types. The results provided by the dried leaves were not as good as those obtained with the other two processed sample types, although the *Pentacebas* variety was correctly classified in 100% of the test set validations. Fresh leaves provided almost perfect classification, and nearly 100% correct classification was attained with dried powdered leaves. Taking only the PLS-DA results into account, the best sample processing method was drying and powdering the leaves. Considering the methodological aspects, using fresh leaves was the fastest and easiest option. The biggest drawback of fresh leaves was their water content, but this did not seem to hinder discrimination between the varieties studied.

In the ASCA-ANOVA model performed in assay two (Table 3), the strongest effect on dried leaves was contributed by the leaf age factor. The age factor accounted for more variability than even the tree variety factor, which could be problematic. Fresh leaves exhibited more favorable behavior in the ASCA-ANOVA model. Results of the ASCA-ANOVA model with dried powdered leaves were similar to those obtained with fresh leaves, but the residual was higher.

4. CONCLUSIONS AND PERSPECTIVES

In this study, we defined a methodology for construction of a classification model that could discriminate between *Prunus dulcis* varieties using NIRS. We also identified the most important sampling and analysis aspects. In assay one, differences were seen in the PCA whether or not the midvein was included. The central leaf region provided more useful information for discriminating between almond tree varieties, because it contained both the primary vein and the laminar tissues. We also attempted to determine which surface of the leaves, adaxial or abaxial, was the most suitable for analysis. Despite the spectral differences observed, the comparison made using the PLSDA model indicated this was not an important aspect.

In assay two, no notable differences were detected between trees of the same variety, which indicated that trees within each variety were quite homogeneous. Differences were observed at the PCA level between young and adult leaves, which indicated age was important to consider during the sampling process.

The best results from the PLS-DA models in assay three were obtained with dried powdered leaves when SNV was used for spectral pre-treatment with application of the SG first derivative (15-point window, second order) and mean centering. However, fresh leaves appeared to be the easiest and most suitable samples for laboratory or industrial analysis. These results indicated that both fresh leaves and dried powdered leaves could be useful for discriminating between *Prunus dulcis* varieties using NIR spectroscopy.

All the information gathered in the present study will be used to build a classification model that includes more *Prunus dulcis* varieties. The potential of NIR spectroscopy for the classification of almond tree varieties and its implementation as a quality control tool in the nursery plant industry will be studied.

Acknowledgment

The authors thank Thermo Scientific for temporary assignment of the NIR equipment.

Funding

This work was supported by Generalitat de Catalunya through a grant from Program of Industrial Doctorates (DI-COF 2017) and by The Spanish Ministry of Economy and Competitiveness (project AGL 2015-70106-R).

REFERENCES

[1] K. Mullis, F. Faloona, S. Scharf, R. Saiki, G. Horn, H. Erlich, Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction., Cold Spring Harb. Symp. Quant. Biol. 51 Pt 1 (1986) 263–73. doi:10.1101/SQB.1986.051.01.032.

[2] P.K. Gupta, J.K. Roy, M. Prasad, Single nucleotide polymorphism a new paradigm for molecular marker technology and DNA polymorphism detection with emphasis on their use in plant.pdf, Curr. Sci. 80 (2001) 524–535. doi:10.2307/24104242.

[3] M. Makky, P. Soni, In situ quality assessment of intact oil palm fresh fruit bunches using rapid portable non-contact and non-destructive approach, J. Food Eng. 120 (2014) 248–259. doi:10.1016/J.JFOODENG.2013.08.011.

[4] M. V. Reboucas, J.B. dos Santos, D. Domingos, A.R.C.G. Massa, Near-infrared spectroscopic prediction of chemical composition of a series of petrochemical process streams for aromatics production, Vib. Spectrosc. 52 (2010) 97–102. doi:10.1016/J.VIBSPEC.2009.09.006.

[5] M. Verstraeten, D. Van Hauwermeiren, M. Hellings, E. Hermans, J. Geens, C. Vervaet, I. Nopens, T. De Beer, Model-based NIR spectroscopy implementation for in-line assay monitoring during a pharmaceutical suspension manufacturing process, Int. J. Pharm. 546 (2018) 247–254. doi:10.1016/j.ijpharm.2018.05.043.

[6] Z. Seregély, T. Deák, G.D. Bisztray, Distinguishing melon genotypes using NIR

spectroscopy, Chemom. Intell. Lab. Syst. 72 (2004) 195–203. doi:10.1016/j.chemolab.2004.01.013.

[7] Q. Fan, Y. Wang, P. Sun, S. Liu, Y. Li, Discrimination of Ephedra plants with diffuse reflectance FT-NIRS and multivariate analysis, Talanta. (2010). doi:10.1016/j.talanta.2009.09.018.

[8] F.M. Durgante, N. Higuchi, A. Almeida, A. Vicentini, Species Spectral Signature: Discriminating closely related plant species in the Amazon with Near-Infrared Leaf-Spectroscopy, For. Ecol. Manage. 291 (2013) 240–248. doi:10.1016/j.foreco.2012.10.045.

[9] Y.N. Shao, C.Q. Xie, L.J. Jiang, J.H. Shi, J.J. Zhu, Y. He, Discrimination of tomatoes bred by spaceflight mutagenesis using visible/near infrared spectroscopy and chemometrics, Spectrochim. Acta - Part A Mol. Biomol. Spectrosc. (2015). doi:10.1016/j.saa.2015.01.018.

[10] L. Zhang, S.S. Wang, Y.F. Wang, J.R. Pan, C. Zhu, Discrimination of Transgenic Rice Based on Near Infrared Reflectance Spectroscopy and Partial Least Squares Regression Discriminant Analysis, Rice Sci. (2015). doi:10.1016/j.rsci.2015.09.004.

[11] L. de Oliveira Moura, D. de Carvalho Lopes, A.J. Steidle Neto, L. de Castro Louback Ferraz, L. de Almeida Carlos, L.M. Martins, Evaluation of Techniques for Automatic Classification of Lettuce Based on Spectral Reflectance, Food Anal. Methods. 9 (2016) 1799–1806. doi:10.1007/s12161-015-0366-5.

[12] G.D. Batten, Plant analysis using near infrared reflectance spectroscopy: the potential and the limitations, Aust. J. Exp. Agric. 38 (1998) 697–706. doi:10.1071/EA97146.

[13] P. Skolik, M.R. McAinsh, F.L. Martin, Biospectroscopy for Plant and Crop Science, in: C.S. João Lopes (Ed.), Compr. Anal. Chem., Elsevier, 2018: pp. 15–49. doi:10.1016/BS.COAC.2018.03.001.

[14] A. Candolfi, R. De Maesschalck, D. Jouan-Rimbaud, P.A. Hailey, D.L. Massart, The influence of data pre-processing in the pattern recognition of excipients near-infrared spectra, J. Pharm. Biomed. Anal. 21 (1999) 115–132. doi:10.1080/13554790490495140.

[15] L. Xu, Y.P. Zhou, L.J. Tang, H.L. Wu, J.H. Jiang, G.L. Shen, R.Q. Yu, Ensemble preprocessing of near-infrared (NIR) spectra for multivariate calibration, Anal. Chim. Acta. 616 (2008) 138–143. doi:10.1016/j.aca.2008.04.031.

[16] Å. Rinnan, F. van den Berg, S.B. Engelsen, Review of the most common pre-processing techniques for near-infrared spectra, TrAC - Trends Anal. Chem. 28 (2009) 1201–1222. doi:10.1016/j.trac.2009.07.007.

[17] R.J. Barnes, M.S. Dhanoa, S.J. Lister, Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra, Appl. Spectrosc. 43 (1989) 772–777. doi:10.1366/0003702894202201.

[18] A. Savitzky, M.J.E. Golay, Smoothing and Differentiation of Data by Simplified Least Squares Procedures., Anal. Chem. 36 (1964) 1627–1639. doi:10.1021/ac60214a047.

[19] R.S.K. Barry M. Wise, Neal B. Gallager, Rasmus Bro, Jeremy M. Shaver, Willem Windig, Chemometrics Tutorial for PLS _ Toolbox and Solo, 2006. doi:10.1016/j.cplett.2004.08.130.

[20] P. Wang, Z. Yu, Species authentication and geographical origin discrimination of herbal medicines by near infrared spectroscopy: A review, J. Pharm. Anal. 5 (2015) 277–284. doi:10.1016/j.jpha.2015.04.001.

[21] R.G. Brereton, G.R. Lloyd, Partial least squares discriminant analysis: Taking the magic away, J. Chemom. 28 (2014) 213–225. doi:10.1002/cem.2609.

[22] D. Ballabio, V. Consonni, Classification tools in chemistry. Part 1: linear models. PLS-DA, Anal. Methods. 5 (2013) 3790. doi:10.1039/c3ay40582f.

[23] L. St»hle, S. Wold, Analysis of variance (ANOVA), Chemom. Intell. Lab. Syst. 6 (1989) 259–272. doi:10.1016/0169-7439(89)80095-4.

[24] L. St»hle, S. Wold, Multivariate analysis of variance (MANOVA), Chemom. Intell. Lab. Syst. 9 (1990) 127–141. doi:10.1016/0169-7439(90)80094-M.

[25] G. Zwanenburg, H.C.J. Hoefsloot, J.A. Westerhuis, J.J. Jansen, A.K. Smilde, ANOVA–principal component analysis and ANOVA–simultaneous component analysis: A comparison, J. Chemom. 25 (2011) 561–567. doi:10.1002/cem.1400.

[26] A.K. Smilde, J.J. Jansen, H.C.J. Hoefsloot, R.J.A.N. Lamers, J. van der Greef, M.E. Timmerman, ANOVA-simultaneous component analysis (ASCA): A new tool for analyzing designed metabolomics data, Bioinformatics. 21 (2005) 3043–3048. doi:10.1093/bioinformatics/bti476.

[27] J.J. Jansen, H.C.J. Hoefsloot, J. Van Der Greef, M.E. Timmerman, J.A. Westerhuis, A.K. Smilde, ASCA: Analysis of multivariate data obtained from an experimental design, J. Chemom. 19 (2005) 469–481. doi:10.1002/cem.952.

[28] O.E. de Noord, E.H. Theobald, Multilevel component analysis and multilevel PLS of chemical process data, J. Chemom. 19 (2005) 301–307. doi:10.1002/cem.933.

[29] N. Suzuki, R.M. Rivero, V. Shulaev, E. Blumwald, R. Mittler, Abiotic and biotic stress combinations, New Phytol. 203 (2014) 32–43. doi:10.1111/nph.12797.

[30] J.C.M.S. Moura, C.A.V. Bonine, J. de Oliveira Fernandes Viana, M.C. Dornelas, P. Mazzafera, Abiotic and Biotic Stresses and Changes in the Lignin Content and Composition in Plants, J. Integr. Plant Biol. 52 (2010) 360–376. doi:10.1111/j.1744-7909.2010.00892.x.

FIGURE CAPTIONS

Fig. 1. Image of an almond leaf showing the two studied regions.

Fig. 2. Mean raw spectra from the three processed sample types. Fresh leaf (green dashed line); dried powdered leaf (blue solid line); and dried leaf (red dotted line).

Fig. 2. PCA results from the Guara and Pentacebas varieties with and without inclusion of the midvein. The presence of the midvein is indicated by red diamonds, and absence of the midvein is indicated by green squares. a) Dried leaf of the Guara variety; b) dried leaf of the Pentacebas variety; c) fresh leaf of the Guara variety; d) fresh leaf of the Pentacebas variety.

Fig. 3. PCA results showing the differences between the adaxial (red diamonds) and abaxial (green squares) leaf surfaces. a) Dried leaf of the Guara variety; b) dried leaf of the Pentacebas variety; c) fresh leaf of the Guara variety; d) fresh leaf of the Pentacebas variety.

Fig 4. PCA results from the study of differences between trees of the same variety. Each tree is represented by a different symbol (triangle, circle, diamond, and square). a) Fresh leaf of the Avijor variety; b) fresh leaf of the Guara variety; c) fresh leaf of the Pentacebas variety; d) dried leaf of the Avijor variety; e) dried leaf of the Guara variety; f) dried leaf of the Pentacebas variety; g) dried powdered leaf of the Avijor variety; h) dried powdered leaf of the Guara variety; i) dried powdered leaf of the Pentacebas variety.

Fig. 6. PCA results from the study of differences between young (yellow circles) and adult (pink stars) leaves. a) Fresh leaf of *Avijor* variety; b) fresh leaf of *Guara* variety; c) fresh leaf of *Pentacebas* variety; d) dried leaf of *Avijor* variety; e) dried leaf of *Guara* variety; f) dried leaf of *Pentacebas* variety; g) dried powdered leaf of *Avijor* variety; h) dried powdered leaf of *Guara* variety; i) dried-powdered leaf of *Pentacebas* variety.

534 **TABLES**

535 **Table 1**. Summary of the samples used in the study.

	Varieties	Fresh samples	Dried samples	Dried-powdered samples
		Number of samples		
Assay one	*Guara*	10	10	not used
	Pentacebas	10	10	not used
Assay two	*Avijor*	80	80	80
	Guara	80	80	80
	Pentacebas	80	80	80
Assay three	*Avijor*	80	80	80
	Guara	80	80	80
	Pentacebas	80	80	80

543

	Real class	Data set	Fresh samples		Dried samples	
			Assigned class		Assigned class	
			Guara	*Pentacebas*	*Guara*	*Pentacebas*
Adaxial	*Guara*	Cross-validation	100 %	100 %	100 %	100 %
	Pentacebas		100 %	100 %	100 %	100 %
Abaxial	*Guara*	Cross-validation	100 %	100 %	100 %	100 %
	Pentacebas		100 %	100 %	100 %	100 %

544 **Table 2**. PLS-DA results from the comparison of adaxial and abaxial leaf surfaces.

545

546 **Table 3**. Results of ASCA-ANOVA modelling to study variance of the factors.

	Fresh leaves		Dried leaves		Dried-powdered leaves	
Factor	*Principal components*	*Effect %*	*Principal components*	*Effect %*	*Principal components*	*Effect %*
Variety	2	30.26	2	19.25	2	24.99
Tree	3	1.83	3	4.69	3	1.87
Young / adult	1	19.11	1	24.18	1	6.68
Residual	6	48.80	3	51.88	3	66.46

547

548

549

550

551

552

553 **Table 4.** PLS-DA model results of the spectra pre-treatment and study of the types of pre-processed
554 samples.

Dried-powdered leaves				
Real class	Data set	Assigned class		
		SNV + Mean center	*SNV + 1st derivative + Mean center*	*SNV + De-trending + Mean center*
Avijor	Cross-validation	87.4 %	99.2 %	86.6 %
	Test set validation	97.5 %	100 %	97.5 %
Guara	Cross-validation	89.9 %	99.2 %	89.1 %
	Test set validation	96.6 %	100 %	96.6 %
Pentacebas	Cross-validation	97.5 %	100 %	97.5 %
	Test set validation	99.2 %	100 %	99.2 %
Dried leaves				
Real class	Data set	Assigned class		
		SNV + Mean center	*SNV + 1st derivative + Mean center*	*SNV + De-trending + Mean center*
Avijor	Cross-validation	97.5 %	99.2 %	95.0 %
	Test set validation	95.0 %	98.3 %	93.3 %
Guara	Cross-validation	95.0 %	100 %	93.3 %
	Test set validation	92.5 %	97.5 %	91.7 %
Pentacebas	Cross-validation	97.5 %	99.2 %	93.3 %
	Test set validation	97.5 %	100 %	98.3 %
Fresh leaves				
Real class	Data set	Assigned class		
		SNV + Mean center	*SNV + 1st derivative + Mean center*	*SNV + De-trending + Mean center*
Avijor	Cross-validation	100 %	97.5 %	100 %
	Test set	99.2 %	100 %	99.2 %

	validation			
Guara	Cross-validation	99.2 %	97.5 %	99.2 %
	Test set validation	98.3 %	99.2 %	98.3 %
Pentacebas	Cross-validation	99.2 %	100 %	99.2 %
	Test set validation	99.2 %	99.2 %	99.2 %

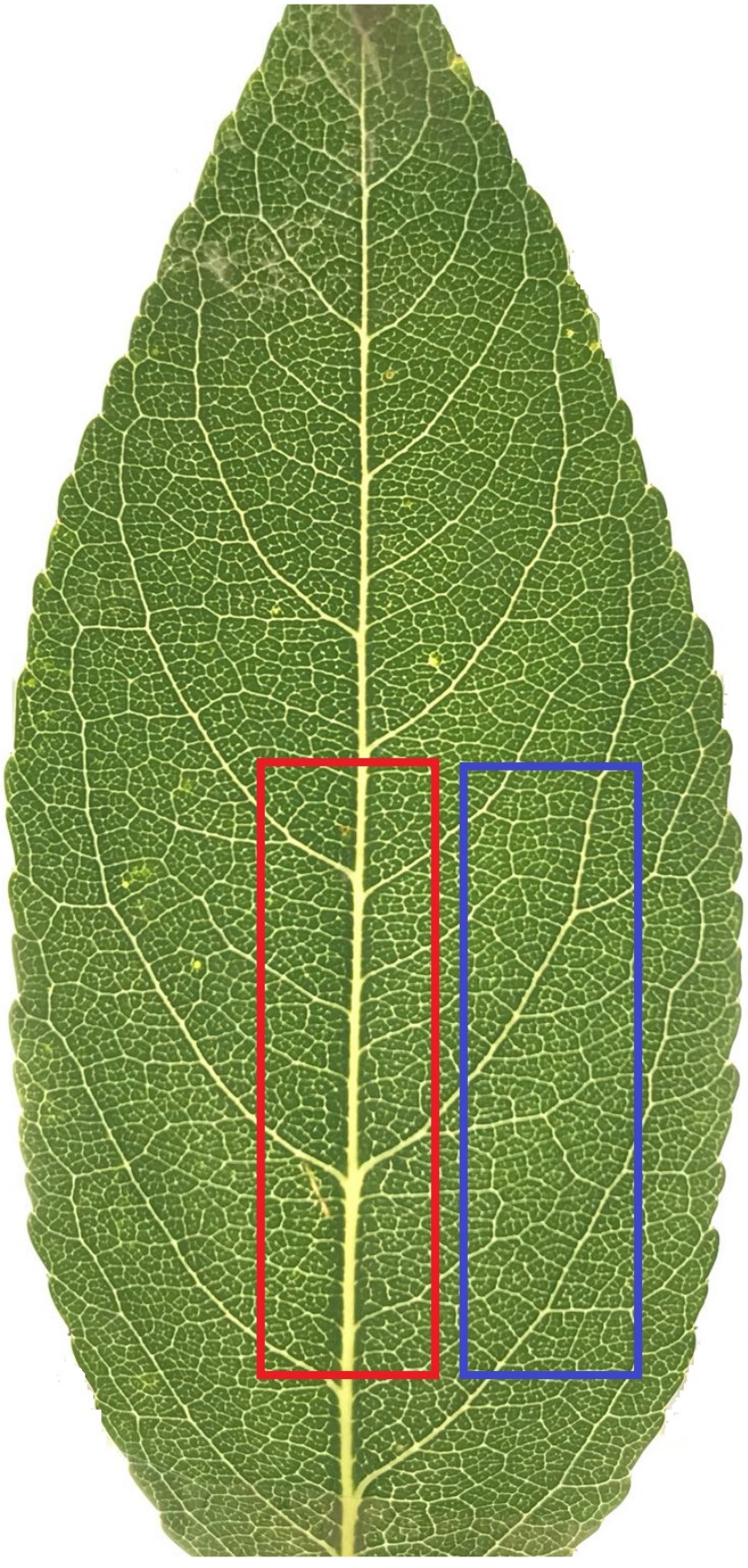
555

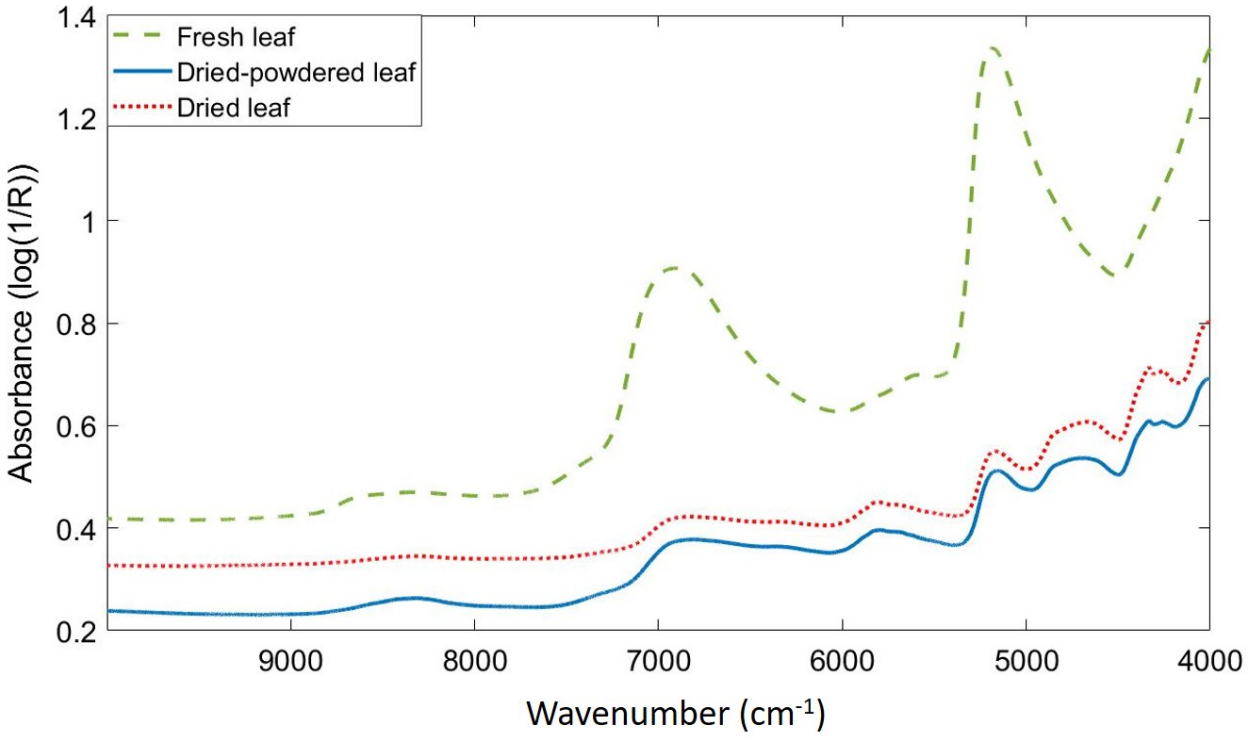
556

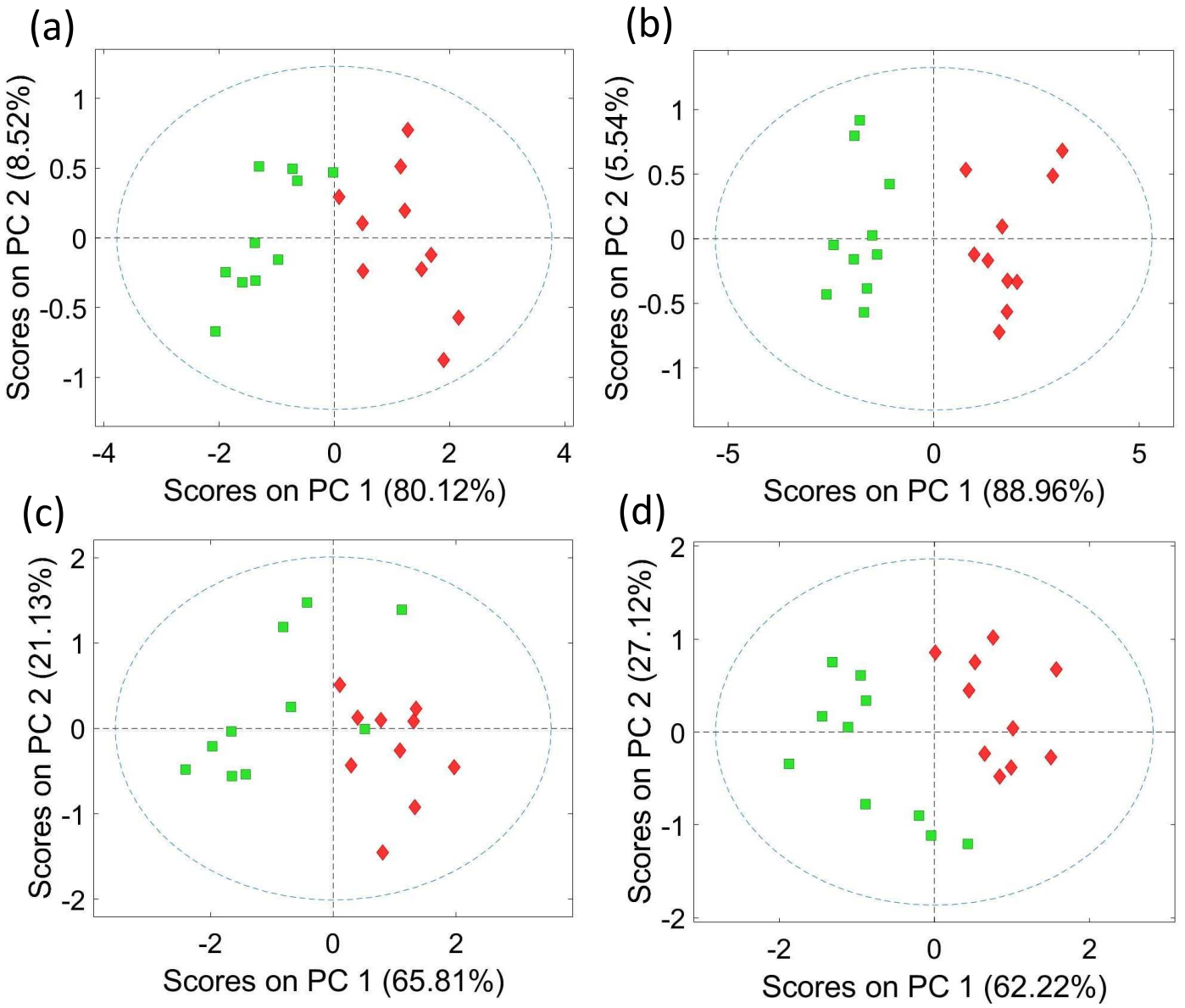
557

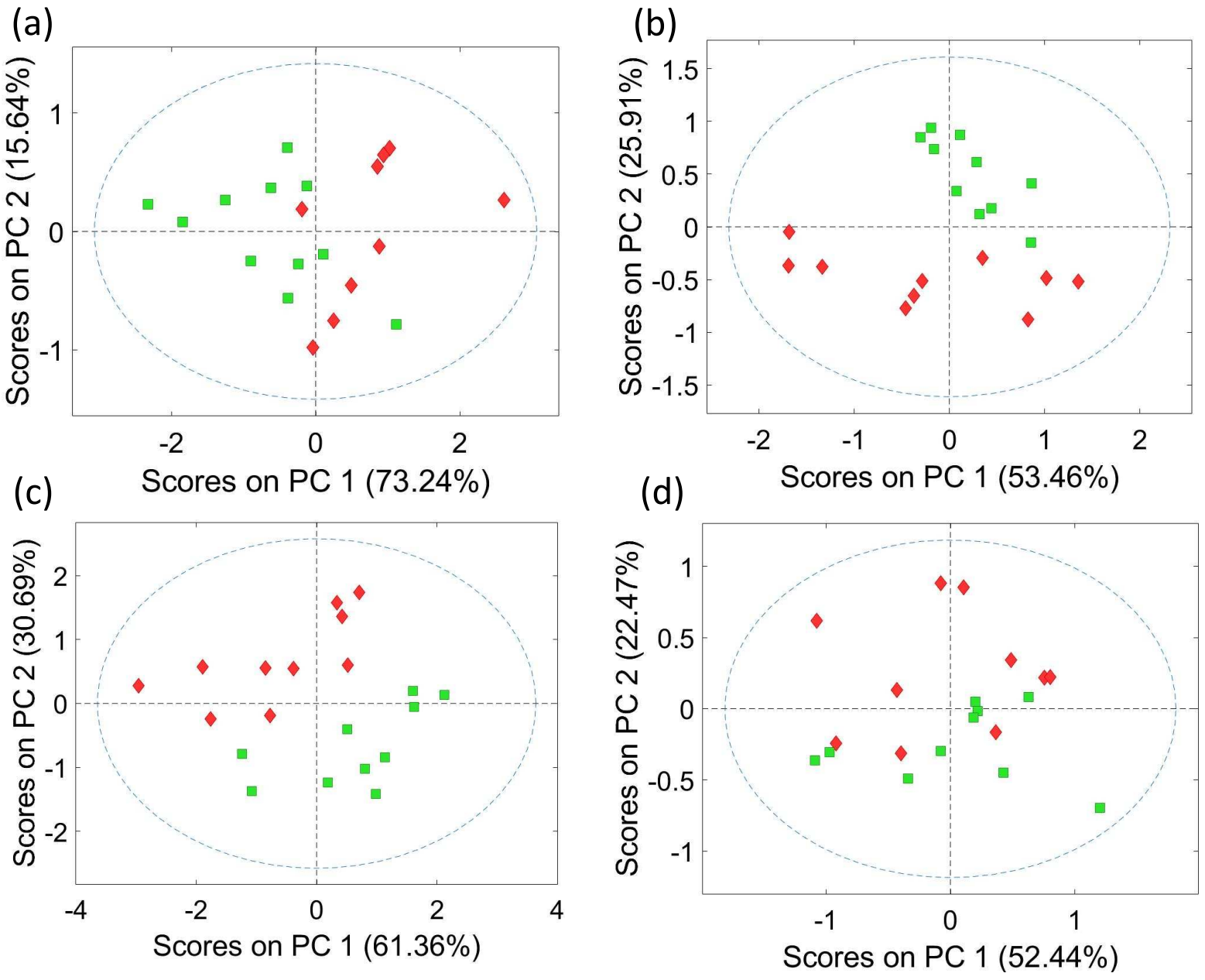
558

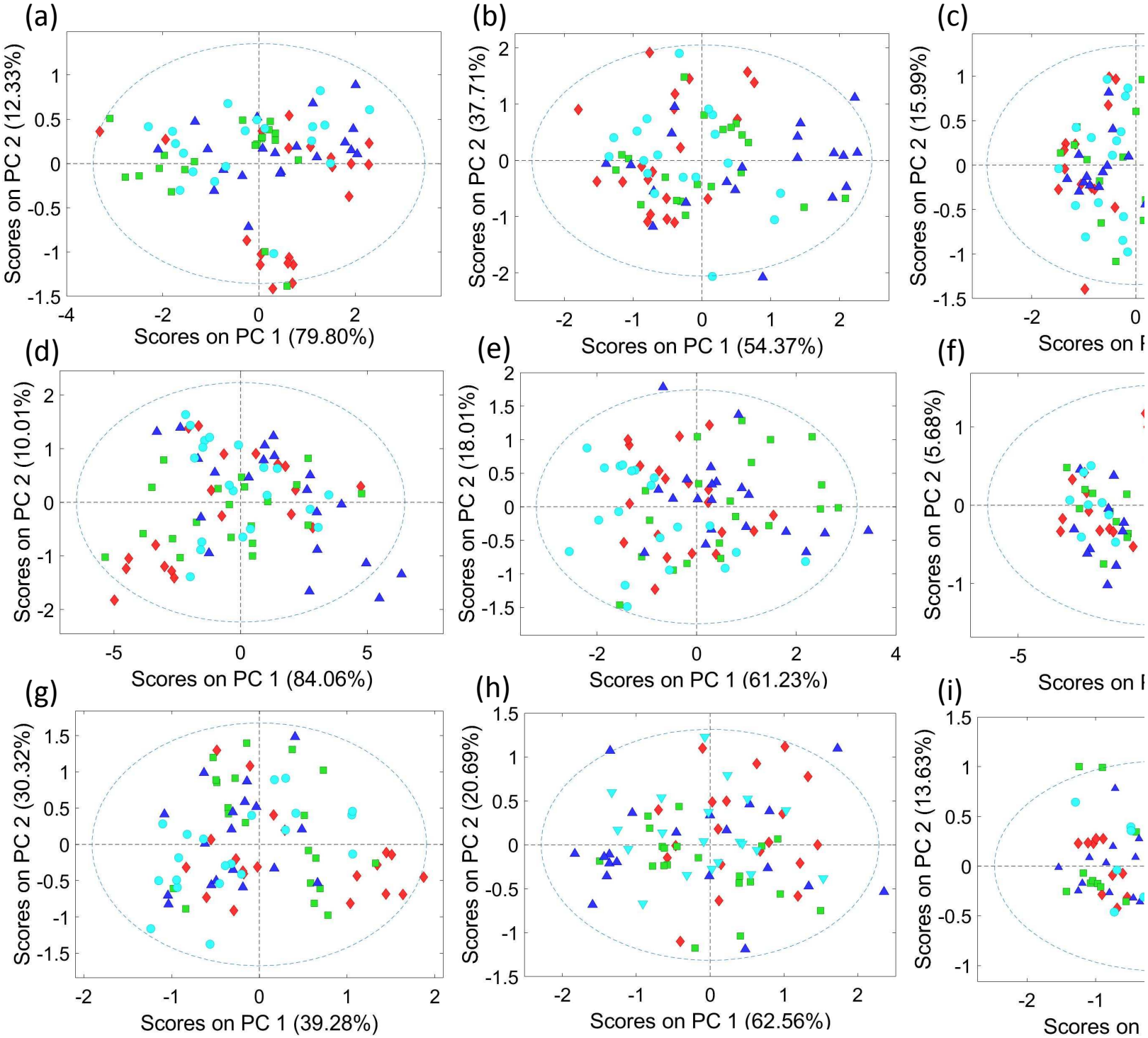
559

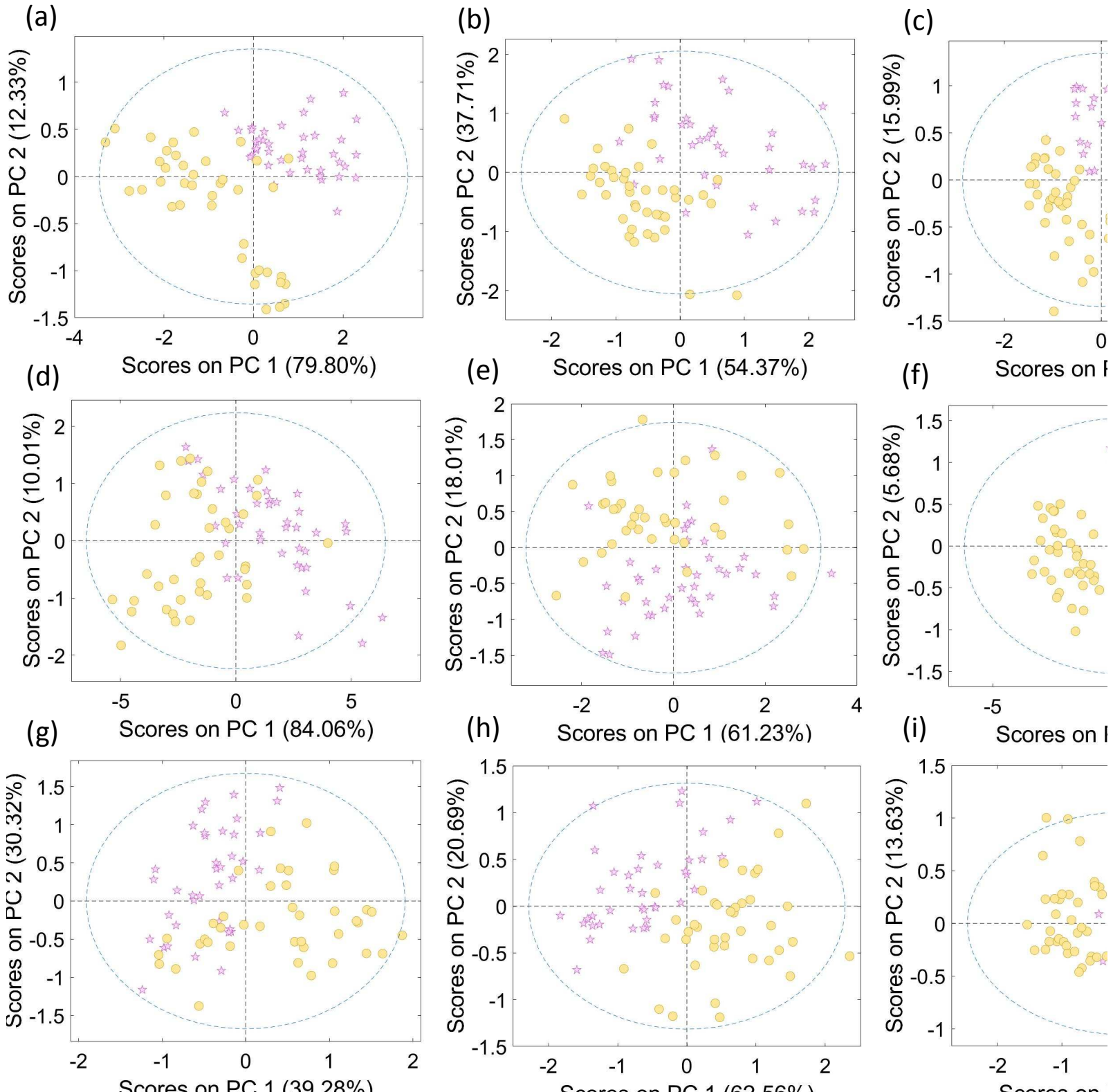





(a)
Scores on PC 2 (8.52%)
Scores on PC 1 (80.12%)
(b)
Scores on PC 2 (5.54%)
Scores on PC 1 (88.96%)
(c)
Scores on PC 2 (21.13%)
Scores on PC 1 (65.81%)
(d)
Scores on PC 2 (27.12%)
Scores on PC 1 (62.22%)


(a)
Scores on PC 2 (15.64%)
Scores on PC 1 (73.24%)
(b)
Scores on PC 2 (25.91%)
Scores on PC 1 (53.46%)
(c)
Scores on PC 2 (30.69%)
Scores on PC 1 (61.36%)
(d)
Scores on PC 2 (22.47%)
Scores on PC 1 (52.44%)


(a)
Scores on PC 2 (12.33%)
Scores on PC 1 (79.80%)
(b)
Scores on PC 2 (37.71%)
Scores on PC 1 (54.37%)
(c)
Scores on PC 2 (15.99%)
(d)
Scores on PC 2 (10.01%)
Scores on PC 1 (84.06%)
(e)
Scores on PC 2 (18.01%)
Scores on PC 1 (61.23%)
(f)
Scores on PC 2 (5.68%)
(g)
Scores on PC 2 (30.32%)
Scores on PC 1 (39.28%)
(h)
Scores on PC 2 (20.69%)
Scores on PC 1 (62.56%)
(i)
Scores on PC 2 (13.63%)


(a)
Scores on PC 2 (12.33%)
Scores on PC 1 (79.80%)
(b)
Scores on PC 2 (37.71%)
Scores on PC 1 (54.37%)
(c)
Scores on PC 2 (15.99%)
(d)
Scores on PC 2 (10.01%)
Scores on PC 1 (84.06%)
(e)
Scores on PC 2 (18.01%)
Scores on PC 1 (61.23%)
(f)
Scores on PC 2 (5.68%)
(g)
Scores on PC 2 (30.32%)
(h)
Scores on PC 2 (20.69%)
(i)
Scores on PC 2 (13.63%)

Highlights

- NIRS was used for discriminating between three *Prunus dulcis* varieties.
- Several spectral pre-treatment strategies were investigated.
- A combination of SNV, SG first derivative, and mean centering methods was optimal.
- Tree variety and leaf age were the most important classification factors for PLS-DA.
- NIRS is a rapid and economical method for nursery plant classification.