Heliyon 10 (2024) e27232

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Influence of sample preparation methods on FTIR spectra for taxonomic identification of tropical trees in the Atlantic forest

Douglas Cubas Pereira *, Breno Pupin, Laura de Simone Borma

National Institute for Space Research (INPE), São José dos Campos, 12227-010, Brazil

ARTICLE INFO

Keywords:
FTIR
Plant taxonomy
Multivariate analysis
Atlantic forest
Tropical trees

ABSTRACT

The Atlantic forest is one of the world's major tropical biomes due to its rich biodiversity. Its vast diversity of plant species poses challenges in floristic surveys. Fourier transform infrared spectroscopy (FTIR) enables rapid and residue-free data collection, providing diverse applications in organic sample analysis. FTIR spectra quality depends on the sample preparation methodology. However, no research on FTIR spectroscopy methodology for taxonomy has been conducted with tropical tree species. Hence, this study addresses the sample preparation influence on FTIR spectra for the taxonomic classification of 12 tree species collected in the Serra do Mar State Park (PESM) - Cunha Nucleus – São Paulo State, Brazil. Spectra were obtained from intact fresh (FL), intact dried (DL), and heat-dried ground (GL) leaves. The spectra were evaluated through chemometrics using Principal Component Analysis (PCA), Hierarchical Cluster Analysis (HCA), and Linear Discriminant Analysis (LDA) with validation by LDA-PCA. The results demonstrate that sample preparation directly influences tropical species FTIR spectra categorization capability. The best taxonomic classification result for all techniques, validated by LDA-PCA, was obtained from GL. FTIR spectra evaluation through PCA, HCA, and LDA allow for the observation of phylogenetic relationships among the species. FTIR spectroscopy proves to be a viable technique for taxonomic evaluation of tree species in floristic exploration of tropical biomes which can complement traditional tools used for taxonomic studies.

1. Introduction

The Atlantic forest is considered one of the main biodiversity hotspots due to its high number of species, high endemism, and extensive degradation [1,2]. The Serra do Mar State Park (PESM) is one of the last remnants of primary native forests in the Atlantic Forest biome in São Paulo state, encompassing submontane, montane, and high-montane forest formations, which harbor a great diversity of tree species [3]. Floristic surveys conducted in primary and secondary forest areas in PESM have revealed a diversity of 100–200 tree species per hectare and 562 species of higher plants in the altitudinal gradient from 10 to 1100 m [4–6].

Consequently, due to the high biodiversity present in the Atlantic Forest, taxonomic identification of tree species represents a challenge in fieldwork [3]. Taxonomic knowledge is essential to understand ecological species dynamics that can be reproduced in reforestation projects in Atlantic forest biome. This identification traditionally requires a deep knowledge of botany and usually involves the complete collection of leaves, flowers, and fruits of the species for accurate identification through exsiccates [3,6]. On the other hand, the use of chemical identification methodologies, such as spectroscopic techniques, has the advantage of being objective

* Corresponding author.
E-mail addresses: douglas.pereira@inpe.br (D. Cubas Pereira), breno0891@hotmail.com (B. Pupin), laura.borma@inpe.br (L. de Simone Borma).

https://doi.org/10.1016/j.heliyon.2024.e27232
Received 4 September 2023; Received in revised form 23 February 2024; Accepted 26 February 2024
Available online 27 February 2024
2405-8440/© 2024 Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

and quantitative, without requiring the complete collection of plant samples, only tissue fragments such as leaves are sufficient for the evaluation [7–9].

Among spectroscopic techniques, Fourier Transform Infrared Vibrational Spectroscopy with Attenuated Total Reflectance (FTIR-ATR) stands out as an effective tool for the analysis of plant organic material, aiming at sample classification [10–12]. FTIR is a non-invasive collection technique that does not generate laboratory waste, is fast and low-cost, allowing the characterization and quantification of organic functional groups that reflect the biochemical composition of plants [8–10].

Due to the high versatility of the technique, FTIR-based studies have been applied in various plant research. For example, taxonomic identification of species using leaf samples [7,10,13]), pollen [12,14,15], and roots [16], classification of invasive species hybrids [17], study of the influence of different land uses on grasses [18], evaluation of plant-environment-atmosphere interaction [14,19,20] influence of environmental pollution on pollen composition [21], change in plant biochemical composition after disease infection [22,23], influence of toxic soils on plants [24] evaluation of hydrogels use as plant substrate [25], and evaluation of species with pharmaceutical potential through *in situ* collection of spectra [8], among others.

The quality of the obtained FTIR spectrum depends on various factors, such as the field sample collection process, sample storage, sample preparation method, laboratory environment condition during spectra collection, spectral processing methods, and statistical analysis methodology of the data [9,27]. Most of the research on plant material using FTIR spectroscopy has been conducted with temperate region plants [7,10,12,13]. It is also important to consider that the distance between the field sites and laboratories may need the analysis of dry and ground samples [9].

Consequently, it is necessary to evaluate sample preparation methodologies for FTIR tropical species analysis. These areas present greater temperature range, biodiversity, and morphological diversity compared to temperate areas, which may have a significant impact on the FTIR collection and sampling process. Therefore, the present study aims to evaluate which sample preparation methodology allows for better taxonomic classification of twelve Atlantic Forest tree species FTIR spectra by chemometric techniques of Principal Component Analysis (PCA), Hierarchical Clustering Analysis (HCA), and Linear Discriminant Analysis (LDA).

2. Methodology

2.1. Species selection

Twelve native species were selected from the PESM in the municipality of Cunha, São Paulo state, Brazil (23°14′22.05″S; 45° 1′26.05″W) (Appendix A. Supplementary Material), namely: *Araucaria angustifolia* (*Aa*); *Campomanesia guaviroba* (*Cb*); *Guapira opposita* (*Go*); *Inga sessilis* (*Is*); *Myrsine gardneriana* (*Mg*); *Myrsine lineata* (*Ml*); *Myrsine umbellata* (*Mu*); *Nectandra lanceolata* (*Nl*); *Psychotria suterella* (*Ps*); *Schinus terebinthifolia* (*St*); *Senna multijuga* (*Sm*); *Sapium glandulosum (Sg)*. The selection done in the collection site was based on the list of species recommended for reforestation in the Atlantic Forest biome in the state of São Paulo and represent the most common species per family found in floristic surveys in the PESM [3–5,26]. *Araucaria angustifolia* was also evaluated because it is a representative threatened species found in the Cunha PESM nucleus [27].

2.2. Sample collection and storage

Leaf samples of the twelve species were collected in the field during the summer season between 10:00 a.m. and 2:00 p.m., as this period is ideal to avoid differences caused by the plants' circadian cycle [28]. The leaves were collected from the lower part of the trees canopy. Considering that leaf diseases can alter the FTIR spectroscopic signal and compromise the final results only healthy leaves were selected from the samples collected [22,23]. After collection, the fresh leaf samples were stored in plastic bags to prevent water loss during transportation to the laboratory, which occurred on the same day as collection. In the laboratory, the samples were kept in a refrigerator at 10 °C, they were not frozen prior to sampling, as has been done in research with temperate region species [9].

2.3. Sample preparation

For comparison purposes, three different types of sample preparation were performed: i) intact fresh leaves, *in natura* (FL); ii) intact oven dried leaves (DL); and iii) heat-dried ground leaves, using a mill (GL). FL stored for a maximum of 24 h after collection were directly evaluated on the spectrophotometer. The samples were not hydrated in the laboratory prior to FTIR spectroscopy analysis [9]. Considering that the objective was to observe the spectra characteristics of FL closer to field conditions, reckoning with the time between collection and laboratory analysis.

For the analysis of DL, the FL samples were placed in an oven at 60 °C for 72 h to remove moisture [9]. The same leaves specimens were evaluated in the category DL and after processing as the GL group. The spectra in DL were acquired from adjacent spots to the previous FL spectra spot measure to avoid post pressure impacts. During the drying step, excessively high temperatures were not used not to compromise the biochemical composition of the plants [29]. The spectra collection of FL and DL was performed on the abaxial side at three points per leaf [9,25]. Spectra were collected from 6 leaves per plant, totaling 18 FTIR spectra per species.

The selected DL were ground using an IKA A11 basic analytical mill, a pestle for complete pulverization of the samples, and sifted through a 10-mesh sieve. The finer the particle size of the analytical compound, the better the contact with the spectrophotometer sample holder and, consequently, the more intense the FTIR spectrum signal [9]. The analysis of GL was performed directly on the FTIR sample holder on the ATR as the FL and DL analysis where the same pressure was applied for all samples.

2.4. Preprocessing of FTIR spectra

The 216 analyzed spectra were obtained in mid-infrared absorbance with FTIR in the range of 4000 cm^{-1} to 450 cm^{-1}, with a resolution of 4 cm^{-1}, 32 scans at room temperature, and a data spacing of 2 cm^{-1} on the Bruker Optik GmbH Alpha II FTIR spectrophotometer equipped with a diamond crystal ATR (Appendix B. Supplementary Material). The spectra were preprocessed using the Bruker OPUS 8.5 software, including baseline correction by the rubber band method and spectral smoothing using the Savitsky-Golay algorithm (9 points) [30].

During preprocessing, the spectra were also normalized in the range of 1690 cm^{-1} to 1620 cm^{-1}, which showed the least variation among the different evaluated species. This range corresponds to Amide I, related to plants' structural proteins [10,11,31]. The spectra were analyzed focused on the fingerprint region of the samples (1770 cm^{-1} - 700 cm^{-1}).

The application of the first derivative in FTIR spectra is widely used in plant research [9,10]. However, it was observed that in the spectra of the tropical species evaluated in this study, there was amplification of noise in the fingerprint area. Therefore, the spectra were analyzed without that procedure.

2.5. Statistical analysis of FTIR spectra

One of the crucial statistical issues in multivariate analysis is the presence of collinearity throughout data. Although this factor is not prevalent for FTIR data, where the variables are measured independently, the preprocessing procedure can induce collinearity [8]. Therefore, previously to the PCA analysis, the data were assessed by calculating the Variance Inflation Factor (VIF) and mean centered using the function 'pca.fit_transform' on Python script with the Jupyter programming interface to observe if the dataset could be affected by multicollinearity.

PCA is a multivariate statistical method that allows the reduction of a data matrix by combining the original variables through the calculation of uncorrelated indices in order of importance based on variance [32]. The coefficients (loadings) determine how much each original variable contributes to each principal component extracted in PCA, and the combination of the original data weighted by the coefficients determines the scores [33]. The PCA scores result from the linear transformation of the source data that provide

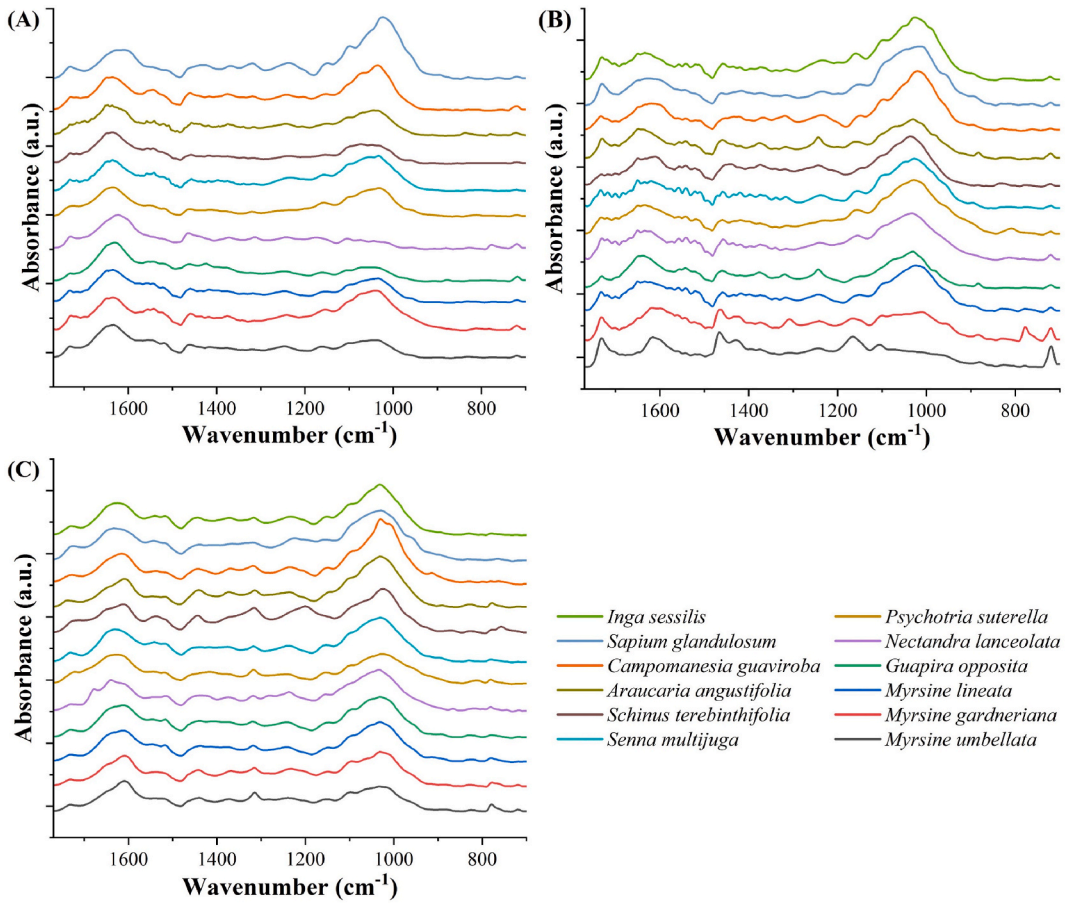

Fig. 1. Average FTIR spectra of plant samples fingerprint region (1770 cm^{-1} – 700 cm^{-1}): A = FL; B = DL and C = GL.

information about the distribution of the samples, aiding to understand existing patterns in the original data [33]. Therefore, PCA was conducted to observe patterns in the spectra data. The plotting of scores and loadings graphs were performed using the Spectroscopy Data PCA package (v1.30) in the software OriginPro 2022.

HCA is a multivariate data analysis technique that allows the use of variable values to establish a grouping scheme for objects (samples) into similar classes in order to observe the formation of groups [32]. Therefore, to observe the similarity relationships in the fingerprint region (1770 cm^{-1} - 700 cm^{-1}) of the FTIR spectra of the species, Hierarchical Cluster Analysis (HCA) was also performed in OriginPro 2022 by using Euclidean distance and the Ward method for each of the sample treatments [17].

LDA considers the classes or categories of the source data, aiming to discover a data projection that maximizes data variability, being suit to supervised evaluation [32]. The validation parameters of LDA-PCA allow us to observe the ability to classify a dataset. Accuracy measures the ability to correctly predict data in relation to the total data in the model, specificity predicts negative values, recall measures sensitivity to positive values, precision measures how accurate positive predictions are, and the F1-score combines recall with precision for a comprehensive analysis of possible false positives that influence the model [34].

Therefore, LDA was established based on the components obtained in PCA to reduce the dimensionality of the data, remove correlation, and classify the clusters based on classes [32]. The Cross-validation was conducted using the cross_val_score function in the scikit-learn library in Python where the dataset was split into 10 folds, by calculating the parameters of the model's ability to classify the data from all conducted sample treatments [35].

The first 4 principal components were used for validation with LDA-PCA as they were statistically sufficient to explain the total variance of the data without under/overfitting. The cross-validation presents better performance for classification of data than the single split tests that may induce bias, especially in small datasets [36,37]. In order to observe the predictive ability of taxonomic classification, the cross-validation test was conducted through the calculation of the quality parameters and plotting the confusion matrixes for the different samplings evaluated.

3. Results and discussion

3.1. Spectral peak analysis

There is good reproducibility of species-specific spectra in the three different sampling methods evaluated – FL, DL and GL (Fig. 1). The spectra analysis is focused in the FTIR fingerprint region (1770 cm^{-1} – 700 cm^{-1}) where the main defining molecules are found in plant samples [9–11]. Due to the sample conditions, it was not possible to collect FL spectra of *Senna multijuga* (*Sm*).

The main observed peaks in the three sampling methods are at 1640 cm^{-1} corresponding to the stretching of C═O and C═N bonds of the proteins in Amide I and the peak at 1030 cm^{-1} corresponding to the stretching of hydroxyls and carboxyls of polysaccharides [10,11,17,18,38,39].

The grinding process enables a better definition of bands, especially in the region of 1200 cm^{-1} to 1000 cm^{-1}, region which presents discriminant saccharides among different plant species as well as internal structural proteins and lipids of the plant cell wall [8,10–12,40].

The band at 1440 cm^{-1} also appears with better definition in GL samples, as also observed by Holden [28]. The grinding procedure promotes greater homogenization, reflecting the total biochemical content of the leaves more accurately, compared to FL and DL, where especially the composition of the leaf outer surface is reflected in the FTIR spectrum [11,20,27].

There is a slight variation in peak position among the samples from different treatments; however, no significant alteration in the main peaks is observed. This result is expected since the samples were not subjected to chemical extraction, which can significantly

Table 1
Frequency Assignments Peaks - Fingerprint Region (1770-700 cm^{-1}).

Wavenumber (cm^{-1})	Vibration Assignment	Approximate Components	References
1730	υ of the C═O ester group	Pectin, Polysaccharides (Triglycerides)	[9,10,12,17,21,38]
1670[(a)]	υ of β-turns in the C═O bond	Proteins of Amide I	[17]
1640	υ of the C═O and C–N bonds	Amide I/Proteins	[10,17,18,38,39]
1540	υ of the C═N and N–H bonds	Amide II/Lipids and particularly Proteins	[8,11,12,17]
1460	δ_{ass} of the C–H bond in -CH2 and -CH3 groups	Amide III/Cell Wall Proteins	[11,17,21]
1440	υ of carbon bonds in aromatic rings	Lipids/Fatty Acids/Proteins	[17,27]
1370	δ_s of the C–H bond in -CH2 and -CH3 groups	Cell Wall Proteins/Lipids	[9,11,17]
1320	δ of the –CH bond	Hemicellulose/Cellulose	[11]
1240	υ of the C═N, N–H, and C═O bonds	Amide IV/Proteins/Hemicellulose	[7,21]
1150	υ_{ass} of the C–O and C–N bonds in the –COOH group	Chlorophyll/Cell Wall Polysaccharides	[11,40]
1100	υ of the C–*O*–C bond in esters and the C–N bond	Polysaccharides/Cutin	[9,10,12,39]
1030	υ of the O–H and C–OH bonds	Polysaccharides/Glucomannan	[10,11]
830[(b)]	δ_{oop} of C–H bonds in rings	Polyphenols	[8,38]
730[(c)]	δ_{ass} of the -CH2 bond in-plane "rocking"	Lignin (Rings)	[38]

Caption: υ: Stretching. υ_s: Symmetric stretching. υ_{ass}: Asymmetric stretching. δ: Bending. δ_s: Symmetric bending. δ_{ass}: Asymmetric bending. δ_{oop}: Out-of-plane bending. (a) present in *N. lanceolata*; (b) present in *P. suterella, M. gardneriana, S. terebinthifolia*, and *A. angustifolia*; (c) present in *M. umbelata, P. suterella, M. gardneriana, S. terebinthifolia, S. glandulosum*, and *A. angustifolia*.

alter the position and intensity of FTIR bands [41]. The tentative main assignments of the spectra collected for the twelve tropical trees are gathered in Table 1.

There are studies related to some of the species here evaluated concerning their biochemical composition, especially their applications (Appendix C. Supplementary Material). However, some species here analyzed require further studies to improve their composition knowledge in which FTIR spectroscopy can be used as an important tool to describe biomolecular components (Table 1).

3.2. PCA and HCA

The preliminary analysis of Variance Inflation Factor (VIF) do not reveal evidence of collinearity in the data, which allows to proceed with PCA and HCA. Fig. 2 presents the scores and loadings representations of the PCA and the HCA for each of the three sampling methods evaluated. The score plots represent 89,8%, 85,1% and 79,2% of the data total variance for FL (Fig. 2A), DL (Fig. 2D) and GL (Fig. 2G), respectively.

The loadings demonstrate that the polysaccharides, especially glucomannan (1150 cm^{-1} - 1030 cm^{-1}) are the main influence in the PC1 and PC2 in all sampling methods (Fig. 2B, E and 2H) which suggests a reflection of the samples wax cuticle composition particularly on FL as also observed by Chen et al. [8] in the analysis of the medicinal plant *Lonicera japonica.* The polysaccharide pectin (1730 cm^{-1}) is influential in the PC2 especially in the intact samples which also suggests it may constitute a major leaf epidermis compound [28].

Some species presents morphological attributes that may affect the contact between sample and spectrophotometer. That was observed in the species *Aa* and *Nl*, both species present coriaceous leaves with a thicker wax cuticle [42]. Which is a possible factor for the higher dispersion observed in the HCA of FL and DL (Fig. 2C, F). The *Aa*, *Sg* and *Myrsine* species FL samples coupling in the HCA suggest a possible chemical similarity among the volatile compounds present in the leaf epidermis of these species [27,43,44]. However, further studies by FTIR in FL are needed to observe the capacity to identify cuticle compounds in the next steps of this research.

The FL and DL samples cannot be separated by species with as observed by the high dispersion in the PCA and HCA (Fig. 2A, C, 2D,

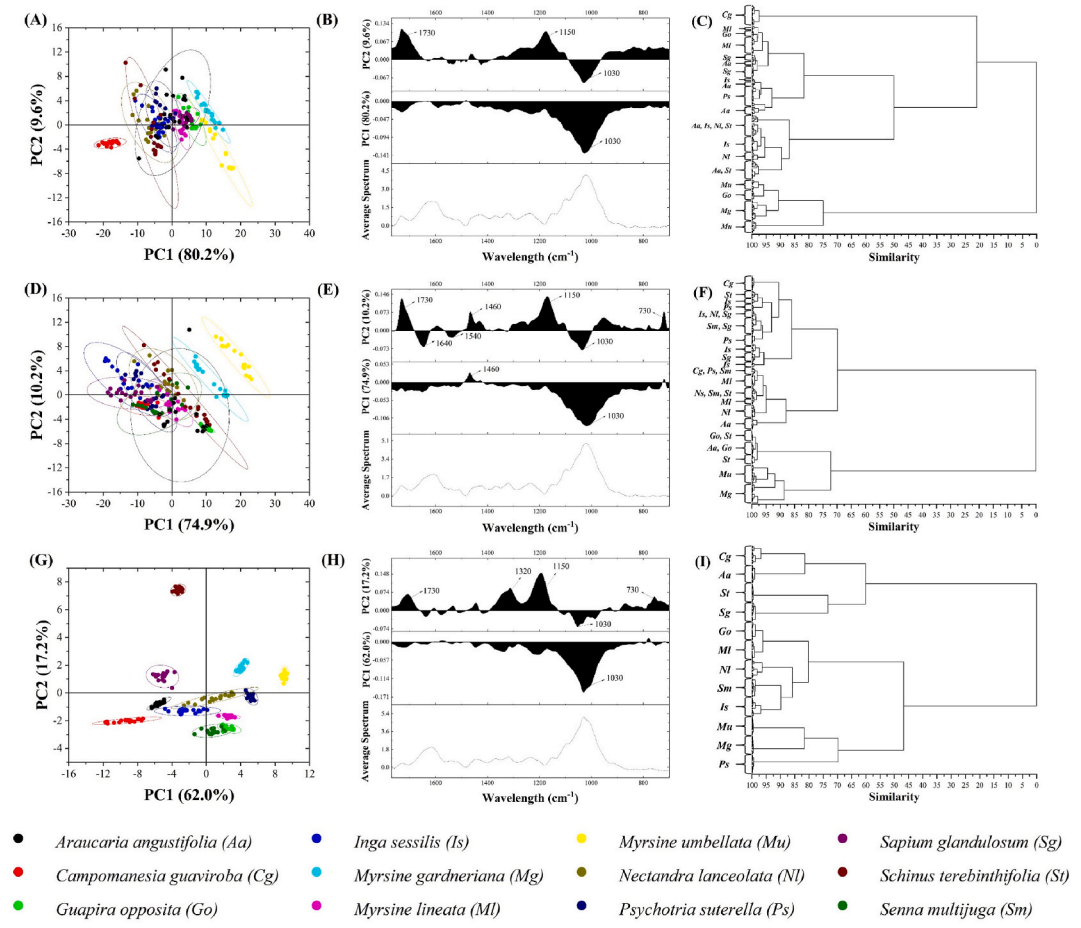

Fig. 2. Score plot and loadings from PCA with 95% confidence ellipse and HCA graph plot of the fingerprint region (1770 cm^{-1} – 700 cm^{-1}) of FL (A, B, C), DL (D, E, F) and GL (G, H, I).

2F), which corroborates the premise that plant leaf spectra cannot be explained just by the cuticle composition and structure alone [30]. Also, other factors such as the distance of the field areas from laboratory, storage, and leaf morphology can affect the intact leaves FTIR spectra quality [11]. The HCA of FL and DL also indicates low capability of classification of intact leaves spectra where only the species *Cb*, *Mu* and *Mg* presents a grouping tendency.

Better results of classification with *in natura* samples were obtained through *in loco* spectra collection with a portable FTIR spectrophotometer [8]. However, especially in the context of tropical biomes, there are difficulties to collect data *in loco*, as observed

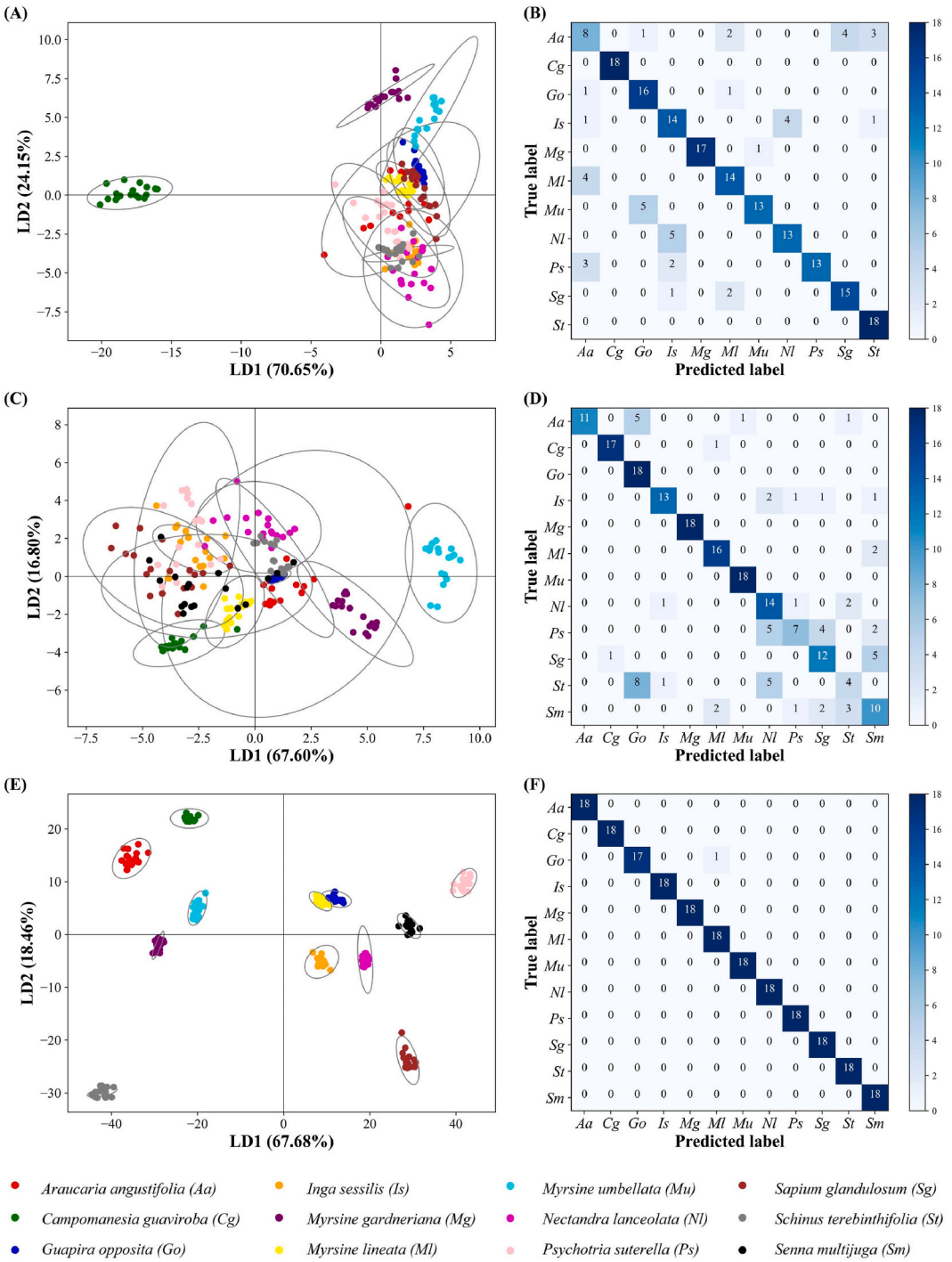

Fig. 3. Score plot graph LDA-PCA with 95% confidence ellipse and LDA-PCA prediction confusion matrixes (B, D, F) - Fingerprint Region (1770 cm^{-1} – 700 cm^{-1}): FL (A, B); DL (C, D); GL (E, F).

during this research, which demands processing methods such as sample grinding to a better evaluation of samples.

Thus, in the present analysis the GL samples separate in homogeneous groups by the PCA and HCA accordingly to the species (Fig. 2G and I). The PCA can successfully separate samples which have some linear correlation between the features [45]. The clades in phylogenetic classification establish groups that share the same evolutionary ancestor [46]. PCA scores of GL indicate a tendency of separation between species representing the Rosids (*Cb*) and the Asterids (*Mu*) clades along the PC1 axis [47].

The HCA can classify GL by species clusters with similarity bigger than 95%. Also, clusters of phylogenetically close species are formed: by family: Myrsinaceae species *Mu* and *Mg* (82,5%) and Fabaceae species *Sm* and *Is* (90%), and by clade: Asterids families: *Ml* (Myrsinaceae) and *Go* (Nyctaginaceae) (95%), *Ps* (Rubiaceae) with *Mu* and *Mg* (Myrsinaceae) (70%) and *Mu*, *Mg*, *Ps*, *Ml* and *Go* (47.5%); Rosids families: *St* (Anacardiaceae) and *Sg* (Euphorbiaceae) (72.5%) and *Cg* (Myrtaceae) with *Sg* (Euphorbiaceae) and *St* (Anacardiaceae) (60%) [47]. These results support the applicability of FTIR in taxonomic researches with tropical plants.

The similarity observed among phylogenetically distant species, especially *Nl* and *Aa* with the remaining species may be related with similar internal compounds shared by those species that became more prominent after the grinding process [28]. Although, future studies are needed to clarify the FTIR capability to evaluate biomolecular tropical plant compounds.

3.3. LDA-PCA validation

The combination of LDA with the scores obtained from PCA allows evaluating the ability to classify samples by group (label) in a supervised manner [33]. Therefore, the LDA-PCA is applied in the different method evaluated (Fig. 3). Also, the quality parameters for the prediction per species and per method are disposed in Fig. 4.

The LDA-PCA first two components respectively explain 94,8%, 84,4% and 86,1% of FL (Fig. 3A), DL (Fig. 3C) and GL (Fig. 3E). The prediction by the confusion matrix is conducted with the first four PCs to avoid under/overfitting. As also observed in the PCA/HCA,

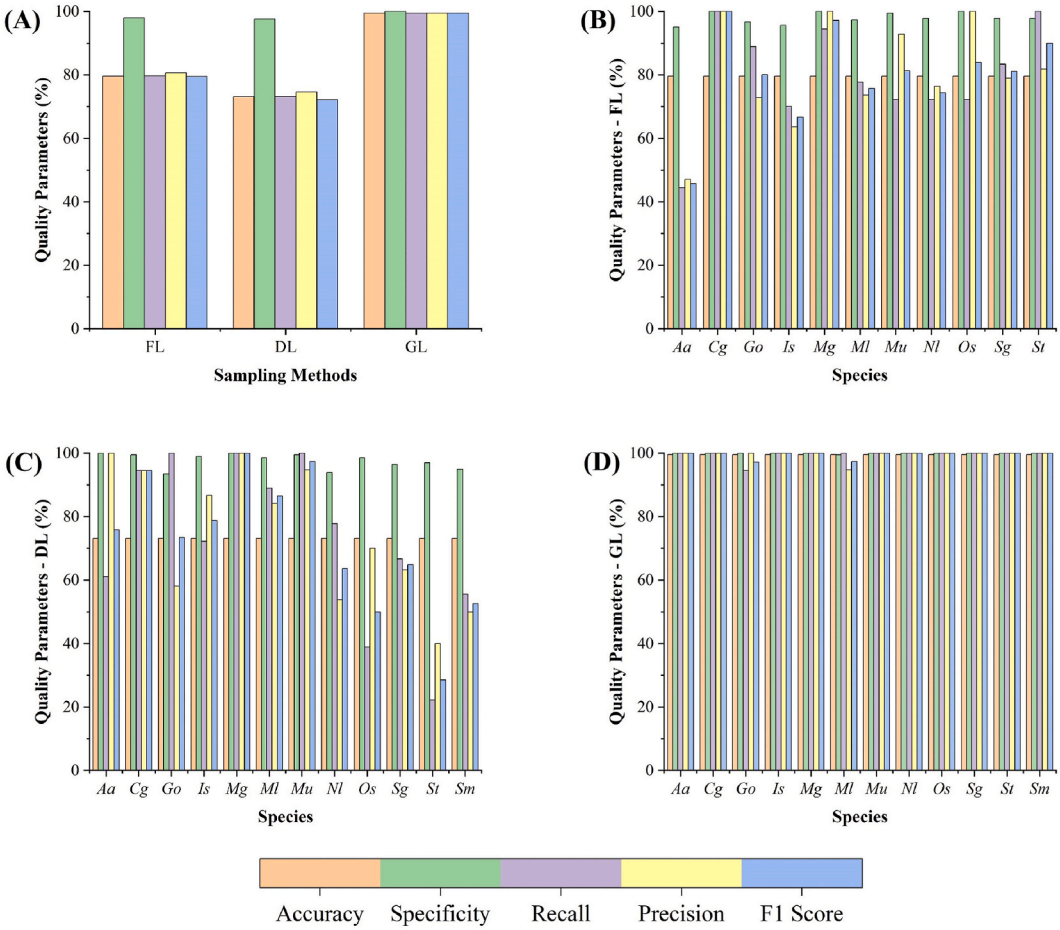

Fig. 4. LDA-PCA classification quality parameters for the three different sampling methods (A) and for each species: FL (B); DL (C) and GL (D) – 18 spectra for each species on the dataset cross-validation. **Caption:** *Araucaria angustifolia* (*Aa*); *Campomanesia guaviroba* (*Cb*); *Guapira opposita* (*Go*); *Inga sessilis* (*Is*); *Myrsine gardneriana* (*Mg*); *Myrsine lineata* (*Ml*); *Myrsine umbellata* (*Mu*); *Nectandra lanceolata* (*Nl*); *Psychotria suterella* (*Ps*); *Schinus terebinthifolia* (*St*); *Senna multijuga* (*Sm*); *Sapium glandulosum* (*Sg*).

LDA-PCA cannot classify the FL and DL all samples by species (Fig. 3A and C). However, some species FL presented good classification on the LDA, especially *Cg*, that species have membranaceous leaves abaxially composed with tector trichomes which grant humidity retention [48]. That morphological adaptation might be an important factor playing for the better classification of *Cg* FL and DL. For FL only *Cg* and *St*, and for DL only *Go, Mg* and *Mu* were 100% correctly predicted by the LDA-PCA validation (Fig. 3B, D – Fig. 4B, C).

Factors such as leaf composition and morphology may be altered by the drying process [29] and consequently alter the FTIR signal. The reduction in prediction from FL to DL affected particularly species with membranaceous leaves [42,49] exceptionally *St* and *Ps* (Fig. 3B and D). Both species also have biochemical similarities such as the presence of volatile monoterpenes on the wax cuticle that might had been lost in DL and affected the prediction [50,51]. Therefore, it is necessary to consider the morphological and biochemical aspects of the species during the leaves processing, more research is required to better comprehend the relation of biochemical composition in the classification of FTIR spectra.

However, the LDA-PCA of GL demonstrate a good capacity to separate the samples by species (Fig. 3E), as indicated by the quality parameters (Fig. 4A). LDA-PCA provides a better separation of the species *Aa* and *Nl* from the other phylogenetically distant species that overlapped in the PCA (Fig. 2G vs Fig. 3E). The overlap of confidence ellipse only occurs between *Go* and *Ml*, both species from families of the Asterid clade [47].

The Asterid clade families represented in the samples have a closer phylogenetical relation [47]. The LDA-PCA components evidences that on this study through the Asterid species similarity in LD2 (*Mg, Mu, Ml, Go* and *Ps*). Rubiaceae species *Ps* is more distant from the Asterids of Myrsinaceae and Nycatginaceae in LD2 which is consistent with the clade phylogeny [47,52].

Despite the broader representation of Rosid families in the samples, a proximity in close related species is observed between Fabaceae *Sm* and *Is* in LD2 and between the Euphorbiaceae *Sg* and the Fabaceae *Sm* in LD1, the proximity of these families is also confirmed by the genetic sequencing of species from these orders [52]. The *St* separated farer from the other Rosid species, which is consistent, considering its order Sapindales is phylogenetically more distant from the other Rosid orders represented in the samples [46]. The order Myrtales has a closer relationship with the order Sapindales according to studies of genetic sequencing of species belonging to these orders [53]. The LDA-PCA here also indicates that the Myrtales *Cg* is closer to the Sapindales *St* in relation to LD1.

Therefore, the best classification is achieved with GL. In Fig. 3F, it is observed that only one sample of *Go* was misclassified as *Ml*, two species belonging to closely related families (Myrsinaceae and Nyctaginaceae) within the Asterid clade [47,54]. This treatment was the only one to achieve 100% in all parameters for the classification of the evaluated tropical tree species (Fig. 4D). Holden [28] also observed that analysis of GL reduces variation between samples, resulting in higher accuracy, specificity, and sensitivity of results in the validation through LDA-PCA, as also demonstrated in the present analysis.

This analysis highlights the importance of defining sample preparation method for FTIR spectra of tropical plant species collection. The need for a treatment to homogenize the samples is observed. Heat-dry grinding proves to be an efficient method for the classification of FTIR spectra of species, as validated by LDA-PCA analysis. Traditional phylogenetic studies using genetic sequencing are costly, and there are many gaps in understanding the classification of many species [52,53]. Hence, FTIR being a rapid technique, it conveys potential as a complementary technique for phylogenetic studies when combined with multivariate statistics of PCA, HCA and LDA-PCA cross-validation [9,10,17].

Despite the longer time of processing, the analysis of ground samples in future studies is proving to be reliable for the analysis of plant samples FTIR spectra where the distance between lab and field is unavoidable. Finally, the grinding allows insights of the internal composition of plants, that can be important for futures studies in the characterization of tropical trees biochemistry for a broad variety of applications as botanic, pharmacology and environmental studies.

The results here obtained suggests that the taxonomic classification of intact *in natura* leaves FTIR spectra are limited in natural field conditions especially by the lapse time between the collection of samples and spectra acquisition. FL evaluation will be important to consider how the environment influences the plants, for that reason future analysis in the field with portable FTIR spectrophotometer may be a solution for the distance limitation improving the results.

Finally, considering that the leaves were collected during the summer season, new studies examining the seasonal impact over tropical plant FTIR spectra are needed. The FTIR precise analysis capability may proportionate the evaluation of individual plant specimen responses to degradation which can substantiate the use of the FTIR spectroscopy technique in environmental studies, especially in forest recovery projects. These studies are the main future targets in the ongoing research project.

4. Conclusion

Sample collection, transportation, and storage were identified as limiting factors for obtaining classifiable FTIR spectra of tropical trees. The FTIR spectra of the intact leaves (FL, DL) does not contain enough information to provide the addressed tropical trees taxonomic classification by PCA, HCA and LDA. Otherwise, the analysis indicates that heat-dried ground (GL) leaves spectra can be taxonomically classified as validated by LDA-PCA combined. The multivariate analysis suggests that FTIR spectroscopy provides information on the phylogenetic relationship of tropical tree species. Combining FTIR with other techniques used for plant taxonomic studies may contribute to a better understanding of tropical species classification, especially in floristic surveys of reforestation projects in endangered tropical biomes as the Atlantic forest.

Data availability statement

All data here evaluated are available by requesting the corresponding author.

CRediT authorship contribution statement

Douglas Cubas Pereira: Writing – review & editing, Writing – original draft, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Breno Pupin:** Writing – review & editing, Data curation. **Laura de Simone Borma:** Writing – review & editing, Supervision, Project administration, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Financial support by the Coordination of Superior Level Staff Improvement - Brazil (CAPES) – Finance Code 001. The authors are thankful to the Graduate Program in Earth System Science (PGCST) at National Institute for Space Research (INPE) for assisting the research to continue, to Dr Luciana Maria Ferrer for helping in text review, and to Dr. Kelly Cristina Tonello for the support at Cunha Nucleus where the samples were collected.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2024.e27232.

References

[1] N. Myers, et al., Biodiversity hotspots for conservation priorities, Nature 403 (6772) (2000) 853–858, https://doi.org/10.1038/35002501.
[2] R.C. Forzza, et al., New Brazilian floristic list highlights conservation challenges, Bioscience 62 (1) (2012) 39–45, https://doi.org/10.1525/bio.2012.62.1.8.
[3] C.A. Joly, et al., Florística e fitossociologia em parcelas permanentes da Mata Atlântica do sudeste do Brasil ao longo de um gradiente altitudinal, Biota Neotropica 12 (2012) 125–145, https://doi.org/10.1590/S1676-06032012000100012.
[4] O.T. DE AGUIAR, et al., Flora Fanerogâmica de um trecho da floresta densa Secundária No Parque Estadual Da Serra do Mar-Núcleo Cunha/Indaiá: Cunha (Sp), Revista do Instituto Florestal 13 (1) (2001) 1–18, https://doi.org/10.24278/2178-5031.2001131625.
[5] A.L.C. Rochelle, R. Cielo-Filho, F.R. Martins, Florística e estrutura de um trecho de floresta ombrófila densa atlântica submontana no Parque Estadual da Serra do Mar, em Ubatuba/SP, Brasil, Biota neotropica 11 (2011) 337–346, https://doi.org/10.1590/S1676-06032011000200032.
[6] N.M. Marchiori, et al., Tree community composition and aboveground biomass in a secondary Atlantic forest, Serra do Mar state park, São Paulo, Brazil, Cerne 22 (2016) 501–514, https://doi.org/10.1590/01047760201622042242.
[7] S.T. Gorgulu, M. Dogan, F. Severcan, The characterization and differentiation of higher plants by Fourier transform infrared spectroscopy, Appl. Spectrosc. 61 (3) (2007) 300–308, https://doi.org/10.1366/000370207780220903.
[8] J. Chen, et al., Rapid and automatic chemical identification of the medicinal flower buds of *Lonicera* plants by the benchtop and hand-held Fourier transform infrared spectroscopy, Spectrochim. Acta Mol. Biomol. Spectrosc. 182 (2017) 81–86, https://doi.org/10.1016/j.saa.2017.03.070.
[9] T. Durak, J. Depciuch, Effect of plant sample preparation and measuring methods on ATR-FTIR spectra results, Environ. Exp. Bot. 169 (2020) 103915, https://doi.org/10.1016/j.envexpbot.2019.103915.
[10] C.A. Holden, et al., Know your enemy: application of ATR-FTIR spectroscopy to invasive species control, PLoS One 17 (1) (2022) e0261742, https://doi.org/10.1371/journal.pone.0261742.
[11] S. Türker-Kaya, C.W. Huck, A review of mid-infrared and near-infrared imaging: principles, concepts and applications in plant tissue analysis, Molecules 22 (1) (2017) 168, https://doi.org/10.3390/molecules22010168.
[12] J. Depciuch, et al., Identification of birch pollen species using FTIR spectroscopy, Aerobiologia 34 (2018) 525–538, https://doi.org/10.1007/s10453-018-9528-4.
[13] S.W. Kim, et al., Taxonomic discrimination of flowering plants by multivariate analysis of Fourier transform infrared spectroscopy data, Plant Cell Rep. 23 (2004) 246–250, https://doi.org/10.1007/s00299-004-0811-1.
[14] B. Zimmermann, A. Kohler, Infrared spectroscopy of pollen identifies plant species and genus as well as environmental conditions, PLoS One 9 (4) (2014) e95417, https://doi.org/10.1371/journal.pone.0095417.
[15] A. Kendel, B. Zimmermann, Chemical analysis of pollen by FT-Raman and FTIR spectroscopies, Front. Plant Sci. 11 (2020) 352, https://doi.org/10.3389/fpls.2020.00352.
[16] B. Rewald, C. Meinen, Plant roots and spectroscopic methods–analyzing species, biomass and vitality, Front. Plant Sci. 4 (2013) 393, https://doi.org/10.3389/fpls.2013.00393.
[17] C.A. Holden, et al., Regional differences in clonal Japanese knotweed revealed by chemometrics-linked attenuated total reflection Fourier-transform infrared spectroscopy, BMC Plant Biol. 21 (1) (2021) 1–20, https://doi.org/10.1186/s12870-021-03293-y.
[18] R. Rana, et al., Leaf Attenuated Total Reflection Fourier Transform Infrared (ATR-FTIR) biochemical profile of grassland plant species related to land-use intensity, Ecol. Indicat. 84 (2018) 803–810, https://doi.org/10.1016/j.ecolind.2017.09.047.
[19] R. Lahlali, et al., ATR–FTIR spectroscopy reveals involvement of lipids and proteins of intact pea pollen grains to heat stress tolerance, Frontiers in plant science 5 (2014) 747, https://doi.org/10.3389/fpls.2014.00747.
[20] M. Bağcioğlu, et al., Monitoring of plant–environment interactions by high-throughput FTIR spectroscopy of pollen, Methods Ecol. Evol. 8 (7) (2017) 870–880, https://doi.org/10.1111/2041-210X.12697.
[21] J. Depciuch, et al., FTIR analysis of molecular composition changes in hazel pollen from unpolluted and urbanized areas, Aerobiologia 33 (2017) 1–12, https://doi.org/10.1007/s10453-016-9445-3.
[22] S.A. Hawkins, et al., Comparison of FTIR spectra between huanglongbing (citrus greening) and other citrus maladies, J. Agric. Food Chem. 58 (10) (2010) 6007–6010, https://doi.org/10.1021/jf904249f.
[23] Y.Y. Chow, A. Ting, Y. Su, Influence of fungal infection on plant tissues: FTIR detects compositional changes to plant cell walls, Fungal Ecology 37 (2019) 38–47, https://doi.org/10.1016/j.funeco.2018.10.004.

[24] S. Palacio, et al., Gypsophile chemistry unveiled: Fourier transform infrared (FTIR) spectroscopy provides new insight into plant adaptations to gypsum soils, PLoS One 9 (9) (2014) e107285, https://doi.org/10.1371/journal.pone.0107285.
[25] D.C. Pereira, B. Pupin, K.K. Sakane, Avaliação do uso de hidrogel no desenvolvimento da *Rapanea ferruginea* com restrição hídrica por espectroscopia vibracional no infravermelho médio com transformada de Fourier (FTIR-UATR), Revista Ambiente & Água 16 (2021), https://doi.org/10.4136/ambi-agua.2744.
[26] L.M. Barbosa, et al., Lista de espécies indicadas para restauração ecológica para diversas regiões do estado de São Paulo, São Paulo: Instituto de Botânica, 2017, pp. 7–344. Available in: https://www.infraestruturameioambiente.sp.gov.br/institutodebotanica/wp-content/uploads/sites/235/2019/10/lista-especies-rad-2019.pdf.
[27] R.M. Peralta, et al., Biological activities and chemical constituents of Araucaria angustifolia: an effort to recover a species threatened by extinction, Trends Food Sci. Technol. 54 (2016) 85–93.
[28] C.A. Holden, ATR-FTIR Spectroscopy-Linked Chemometrics: A Novel Approach to the Analysis and Control of the Invasive Species Japanese Knotweed, 2023. Doctorate Thesis. Lancaster University (United Kingdom). Available in: https://search.proquest.com/openview/28e74bf922d7c89cf93051c9a867e7a9/1?pq-origsite=gscholar&cbl=2026366&diss=y.
[29] A.K. Babu, et al., Review of leaf drying: mechanism and influencing parameters, drying methods, nutrient preservation, and mathematical models, Renewable and sustainable energy reviews 90 (2018) 536–556, https://doi.org/10.1016/j.rser.2018.04.002.
[30] A. Götz, et al., Apparent penetration depth in attenuated total reflection Fourier-transform infrared (ATR-FTIR) spectroscopy of *Allium cepa* L. epidermis and cuticle, Spectrochim. Acta Mol. Biomol. Spectrosc. 224 (2020) 117460, https://doi.org/10.1016/j.saa.2019.117460.
[31] C. Berthomieu, R. Hienerwadel, Fourier transform infrared (FTIR) spectroscopy, Photosynth. Res. 101 (2009) 157–170, https://doi.org/10.1007/s11120-009-9439-x.
[32] B.F.J. Manly, J.A.N. Alberto, Multivariate Statistical Methods: a Primer, Chapman and Hall/CRC, 2016, p. 270p.
[33] D.I. Warton, Eco-Stats: Data Analysis in Ecology: from T-Tests to Multivariate Abundances, Springer Nature, 2022, p. 434p.
[34] D. Ami, P. Mereghetti, S.M. Doglia, Multivariate analysis for Fourier transform infrared spectra of complex biological systems and processes, Multivariate analysis in management, engineering and the sciences (2013) 189–220, https://doi.org/10.5772/53850.
[35] F. Pedregosa, et al., Scikit-learn: machine learning in Python, the Journal of machine Learning research 12 (2011) 2825–2830, https://doi.org/10.5555/1953048.2078195.
[36] G.C. Cawley, N.L.C. Talbot, On over-fitting in model selection and subsequent selection bias in performance evaluation, J. Mach. Learn. Res. 11 (2010) 2079–2107, https://doi.org/10.5555/1756006.1859921.
[37] T. Gunasegaran, Y.-N. Cheah, Evolutionary cross validation, in: 2017 8th International Conference on Information Technology (ICIT), IEEE, 2017, pp. 89–95, https://doi.org/10.1109/ICITECH.2017.8079960.
[38] R. Falcioni, et al., Classification and prediction by pigment content in lettuce (*Lactuca sativa* L.) varieties using machine learning and ATR-FTIR spectroscopy, 11, Plants 24 (2022) 3413, https://doi.org/10.3390/plants11243413.
[39] M.S. Selvam, et al., Assessment of phytochemical, FT-IR and GC-MS fingerprint profiling of marine angiosperms *Enhalus acoroides* (Lf) Royle and *Syringodium isoetifolium* (Asch) Dandy, Gulf of mannar Biosphere Reserve, Tamil nadu, Asian J. Biol. Life Sci. 11 (2) (2022) 469, https://doi.org/10.5530/ajbls.2022.11.64.
[40] B. Ribeiro Da Luz, Attenuated total reflectance spectroscopy of plant leaves: a tool for ecological and botanical studies, New Phytol. 172 (2) (2006) 305–318, https://doi.org/10.1111/j.1469-8137.2006.01823.x.
[41] K. Kucharska-Ambrożej, et al., Quality control of mint species based on UV-VIS and FTIR spectral data supported by chemometric tools, Food Control 129 (2021) 108228, https://doi.org/10.1016/j.foodcont.2021.108228.
[42] H. Lorenzi, Arvores brasileiras: manual de identificação e cultivo de plantas arbóreas nativas do Brasil, 7a Ed Vol. 1, – 2 – 3, Instituto Plantarum de Estudos da Flora, São Paulo, 2016, p. 1152p.
[43] Q. He, et al., Genus Sapium (Euphorbiaceae): a review on traditional uses, phytochemistry, and pharmacology, J. Ethnopharmacol. 277 (2021) 114206, https://doi.org/10.1016/j.jep.2021.114206.
[44] L.V. Laskoski, et al., Phytochemical prospection and evaluation of antimicrobial, antioxidant and antibiofilm activities of extracts and essential oil from leaves of *Myrsine umbellata* Mart. (Primulaceae), Braz. J. Biol. 82 (2022), https://doi.org/10.1590/1519-6984.263865.
[45] T.G. Rios, et al., FTIR spectroscopy with machine learning: a new approach to animal DNA polymorphism screening, Spectrochim. Acta Mol. Biomol. Spectrosc. 261 (2021) 120036, https://doi.org/10.1016/j.saa.2021.120036.
[46] V.C. Souza, H. Lorenzi, Botânica sistemática: guia ilustrado para identificação das famílias de fanerógamas nativas e exóticas no Brasil, baseado em APG IV, 4a Ed., Nova Odessa, Plantarum, 2019, p. 768.
[47] ANGIOSPERM PHYLOGENY GROUP, An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV, et al. Bot. J. Linn. Soc. 181 (1) (2016) 1–20, https://doi.org/10.1111/boj.12385.
[48] P.C.S. Saibert, M.B. Romagnolo, A.L.M. Albiero, Comparação Morfoanatômica de Folhas de *Campomanesia xanthorcapa* O. Berg e *Campomanesia guaviroba* (DC.) Kiaersk. (Myrtaceae) Como Contribuição a Farmacognosia, Visão Acadêmica 19 (2018) 3, https://doi.org/10.5380/acd.v19i3.60594.
[49]] M. Ferreira Junior, A.O.S. Vieira, Espécies arbóreo-arbustivas da família Rubiaceae Juss. na bacia do rio Tibagi, PR, Brasil, HOEHNEA 42 (2015) 289–336, https://doi.org/10.1590/2236-8906-10/2015.
[50] C. Dos Santos Passos, et al., Monoamine oxidase inhibition by monoterpene indole alkaloids and fractions obtained from *Psychotria suterella* and *Psychotria laciniata*, J. Enzym. Inhib. Med. Chem. 28 (3) (2013) 611–618, https://doi.org/10.3109/14756366.2012.666536.
[51] A. Dos Santos Cavalcanti, et al., Volatiles composition and extraction kinetics from *Schinus terebinthifolius* and *Schinus molle* leaves and fruit, Revista Brasileira de Farmacognosia 25 (2015) 356–362, https://doi.org/10.1016/j.bjp.2015.07.003.
[52] T. Liu, et al., De novo assembly and characterization of transcriptome using Illumina paired-end sequencing and identification of CesA gene in ramie (*Boehmeria nivea* L. Gaud), BMC Genom. 14 (1) (2013) 1–11, https://doi.org/10.1186/1471-2164-14-125.
[53] Q. Bi, et al., Complete mitochondrial genome of a Chinese oil tree yellowhorn, *Xanthoceras sorbifolium* (Sapindales, Sapindaceae), Mitochondrial DNA Part B 4 (1) (2019) 1492–1493, https://doi.org/10.1080/23802359.2019.1601038.
[54] C. Zhang, et al., Asterid phylogenomics/phylotranscriptomics uncover morphological evolutionary histories and support phylogenetic placement for numerous whole-genome duplications, Mol. Biol. Evol. 37 (11) (2020) 3188–3210, https://doi.org/10.1093/molbev/msaa160.