Original Article

Integrating morphometric and spectral evidence for species discrimination in Amazonian trees: a case study of *Carapa* (Meliaceae)

Running Head: Morphometrics and NIR spectroscopy discriminate *Carapa*

Semírian Campos Amoêdo[1,3]*, Caroline da Cruz Vasconcelos[2], Isolde Dorothea Kossmann Ferraz[1],

[1]Instituto Nacional de Pesquisas da Amazônia (INPA), Programa de Pós-Graduação em Ciências Biológicas (Botânica), Manaus, Amazonas, Brazil

[2]Universidade Federal do Amazonas, Departamento de Ciências Florestais, Manaus, AM, Brazil

[3]Ação Ecológica Guaporé-Ecoporé, Porto Velho, RO, Brazil

*Corresponding author: semirian.bio@gmail.com

ORCIDs:

Semírian Campos Amoêdo: https://orcid.org/0000-0002-4365-0498

Caroline da Cruz Vasconcelos: https://orcid.org/0000-0002-7850-6672

Isolde Dorothea Kossmann Ferraz: https://orcid.org/0000-0003-3299-7856

Abstract

Reliable species-level identification of Amazonian trees from herbarium and field material is essential for conservation, management, and economic use of non-timber forest products, yet overlapping morphology and inconsistent identifications often hinder it. Using three Amazonian species of *Carapa* (*C. guianensis*, *C. surinamensis,* and *C. vasquezii*) as a case study, we tested whether leaf morphometry, near-infrared (NIR) reflectance spectroscopy, and their combination can discriminate species from vegetative material and whether sample imbalance explains the poor recognition of the rarest species. We measured 12 vegetative morphological characters and obtained NIR spectra (350–2500 nm) from herbarium specimens and field-collected leaves. Species discrimination was evaluated using Linear Discriminant Analysis (LDA), Partial Least Squares Discriminant Analysis (PLS-DA), and Random Forest (RF), considering both the original imbalanced dataset and a class-balanced dataset. Morphometric data provided the highest classification performance (balanced accuracy ≈ 0.68), outperforming models based on NIR spectra alone or on the combined morphometric and spectral datasets. Overall, LDA and PLS-DA performed better than Random Forest. Class imbalance suppressed recall of *C. surinamensis* (as low as ~0.01–0.22), which improved markedly (0.41–0.55) after correction, at the cost of *C. guianensis* recall, showing that imbalance explains much, but not all, of the difficulties in recognizing this species. Residual overlap, mainly

between *C. surinamensis* and *C. vasquezii*, is discussed as morphological and spectral similarity rather than evidence for or against taxonomic boundaries.

Keywords: Amazonian flora, *Carapa*, integrative taxonomy, leaf morphometry, near-infrared spectroscopy, species discrimination

Introduction

The species of the genus *Carapa* Aubl. (Meliaceae Juss.) have been the subject of several taxonomic revisions (De Candolle 1878; Harms 1940; Staner 1941; Styles 1981), the most recent of which (Kenfack 2011a) recognized 27 species, including new taxa and previously synonymized names. For the Brazilian Amazon, three species are currently accepted: *C. guianensis* Aubl., *C. surinamensis* Miq. and *C. vasquezii* Kenfack. In practice, however, these species remain difficult to recognize in the field and in herbarium collections, where a large proportion of specimens is still identified only as *C. guianensis*.

Additional uncertainty stems from evidence of hybridization between *C. guianensis* and *C. surinamensis* in the Guianas (Duminil *et al.,* 2012; Scotti-Saintagne *et al.,* 2013) and from phylogenetic analyses (Koenen *et al.,* 2015) showing that *C. surinamensis* is closely related to *C. vasquezii*, while *C. guianensis* groups with African congeners, albeit with limited support. Taken together, these taxonomic revisions, putative hybridization, and phylogenetic relationships highlight persistent uncertainty around species limits in the genus.

Traditional use also shapes how specimens are identified. *Carapa* species, collectively known as andiroba, are ecologically, economically, and medicinally important across tropical ecosystems: their timber is valued for its quality (Volpato *et al.,* 1972; Yared and Carpanezzi 1981), and the oil extracted from the seeds has anti-inflammatory, antiseptic, and repellent properties (Enriquez *et al.,* 2003; Mendonça and Ferraz 2007) that are important both for traditional medicine and for local economies. Because field researchers commonly record such material under a single vernacular name, most vouchers are attributed to *C. guianensis* in collections, often without revision in light of more recent taxonomic work.

This identification uncertainty compromises applications that depend on well-curated reference data, such as image-based identification (e.g., PlantNet) or leaf spectroscopy models trained on herbarium collections (Durgante *et al.,* 2013; Lang *et al.,* 2015; Prata *et al.,* 2018; Damasco *et al.,* 2019; Vasconcelos *et al.,* 2025; Cavender-Bares e*t al.,* 2026). Accurate classification is, in turn, important for conservation strategies and sustainable use, since it directly affects the quality and economic value of products such as oils and timber.

Part of the difficulty in species determination relates to similarities produced by convergent evolution, introgression and recent speciation, which generate cryptic species and complexes in which interspecific and intraspecific variation are hard to distinguish (Barraclough and Savolainen 2001; Fišer *et al.,* 2018). This is compounded because identification of any organism relies mainly on qualitative external macromorphology, which is subject to the limitations of human perception and varies among determiners (Rohwer *et al.,* 1991; Rohwer 2000; Gomes *et al.,* 2013). In plants, the problem is amplified because the main taxonomic characters lie in flowers and fruits, structures that are rarely present together on a single collected specimen.

Given the large number of herbarium specimens that include only vegetative material, a quantitative approach becomes essential for extracting reliable taxonomic information from otherwise difficult-to-identify specimens (Carvalho *et al.,* 2024). Amazonian long-term forest monitoring plots have accumulated extensive botanical collections through repeated sampling, yet their scientific potential remains underexploited because many specimens still lack reliable species-level identification (Hopkins 2007; Gomes *et al.,* 2013).

In this context, a quantitative approach to vegetative character variation not only supports species delimitation but also provides an objective framework for characterizing morphological diversity and evaluating the degree of differentiation and discrimination among taxa (Kenfack 2011a; Gaem *et al.*, 2022; Cardoso *et al.*, 2022). Unlike traditional approaches based on qualitative, subjective descriptions, morphometrics provides a consistent quantitative foundation for these analyses, increasing the objectivity and efficiency of morphological studies (Bookstein 1982; Cope *et al.,* 2012).

In parallel, other techniques have emerged to assist species determination, notably leaf reflectance spectroscopy, which has shown high predictive power and proved useful for resolving Amazonian plant complexes (Durgante *et al.,* 2013; Lang *et al.,* 2015; Prata *et al.,* 2018; Damasco *et al.,* 2019; Gaen et al., 2022), complementing traditional and morphometric approaches. This technique, based on the absorption and reflection of infrared light by biological samples, offers a non-destructive way to rapidly characterize plant tissue, processing many samples in a short time with minimal preparation (Pasquini 2018; Durgante *et al.,* 2013). It is inexpensive and particularly useful for biodiversity and conservation studies that require rapid, accurate species identification (Durgante *et al.,* 2026), provided species boundaries are well defined, and sampling is adequate to build robust predictive models.

Here we use *Carapa* as a case study to evaluate an integrative morphometric–spectral approach to species discrimination that is, in principle, applicable to other morphologically similar Amazonian tree taxa. Rather than attempting to resolve the geographic distribution or formal taxonomic limits of the three species, our focus is deliberately narrower: we ask how well vegetative morphometry and NIR

spectroscopy, alone and combined, discriminate among specimens already assigned to *C. guianensis*, *C. surinamensis* and *C. vasquezii*, and to what extent the well-known difficulty in recognizing the rarer *C. surinamensis* reflects sample imbalance rather than a lack of distinguishing signal.

Material and Methods

Sample coverage

Samples of *Carapa* spp. were obtained from the herbarium of the National Institute for Amazonian Research (INPA), from reference collections associated with permanent forest plots maintained by the Biological Dynamics of Forest Fragments Project (PDBFF) and the Biodiversity Research Program (PPBIO/INPA), and from field collections conducted specifically for this study. Because the reference collections from permanent plots consist predominantly of vouchers lacking reproductive structures, a common inclusion criterion was applied to all sources: specimens were required to have complete leaves suitable for vegetative morphometric analyses, with at least three leaflets per individual, representing the basal, apical, and central regions of the leaf.

For material collected specifically for this study, sampled trees were separated by a minimum distance of 50 m to reduce the chance of sampling closely related individuals and to obtain geographic representation of the target regions. Individuals previously identified as *C. surinamensis* were found only in planted stands along the Amazon basin. Voucher species assignment for all collections followed prior botanical determination using the taxonomic keys of Styles (1981) and Kenfack (2011b).

Table 1. Number of *Carapa* spp. samples by species, as determined on herbarium vouchers or by prior taxonomic determination for material collected in this study, used in morphometric (MP) and NIR spectroscopy (NIR) analyses.

Collection	*C. guianensis* (MP)	*C. guianensis* (NIR)	*C. surinamensis* (MP)	*C. surinamensis* (NIR)	*C. vasquezii* (MP)	*C. vasquezii* (NIR)
Herbarium/INPA	20	21	7	11	15	32
PDBFF	—	2	—	—	29	44
PPBIO	—	—	—	—	25	25
This study	75	79	20	18	26	29
Total	95	103	27	29	96	130

Note: Species not represented at localities covered by the permanent-plot reference collections are not known to occur at those sites.

Selection of morphological characters

Vegetative morphological characters were selected following the original species descriptions (Aublet 1775; De Candolle 1878; Kenfack 2011a). Twelve quantitative characters were measured once per sample, except petiole length, which was measured three times and averaged: petiole width (cm), petiole length (cm), rachis length (cm), number of leaflet pairs, mean petiolule length (cm), basal leaflet length (cm), basal leaflet width (cm), terminal leaflet length (cm), terminal leaflet width (cm),

basal leaflet length/width ratio, terminal leaflet length/width ratio, and number of secondary veins. Measurements were taken with a digital caliper and a millimeter tape and converted to centimeters for analysis.

Three qualitative characters were also recorded: leaflet shape (elliptic or oblong), apex shape (acuminate or obtuse), and base shape (cuneate, acute or rounded); these were coded as presence–absence for analysis. Ontogenetic effects could not be controlled owing to insufficient metadata in the sampled collections.

Morphometric analyses

An exploratory analysis of the morphological data was performed by Principal Component Analysis (PCA) to reduce dimensionality and identify patterns of variation among species, using the *prcomp* function in R (Rv.4.3.2, Core Team 2023).

To examine differences among species for each quantitative character, boxplots of absolute values were produced, and variation was tested by analysis of variance (ANOVA) followed by Tukey post-hoc comparisons, implemented with the *aov* and *glht* functions.

Pairwise dissimilarity among individuals, combining quantitative and qualitative vegetative characters, was computed with Gower's distance (*daisy* function) and used as input for hierarchical agglomerative clustering with Ward's minimum-variance method (*hclust*, method = "ward.D2").

A Discriminant Analysis of Principal Components (DAPC) was implemented in the *adegenet* package (Jombart *et al.,* 2023) to describe relationships among species and visualize clustering in reduced dimensional space; data were first ordinated by PCA and then subjected to discriminant analysis (Pritchard *et al.,* 2000). The relative contribution of each variable to the discriminant axes was calculated. DAPC scores were compared among species by ANOVA with Tukey post-hoc tests. Assignment probabilities of individuals to each species were visualized with the *compoplot* function.

NIR reflectance spectra collection

To evaluate species discrimination using leaf reflectance spectroscopy, dried leaves from 262 samples obtained from the collections listed in Table 1 and from material collected specifically for this study were analyzed (*C. guianensis*, n = 103; *C. surinamensis*, n = 29; *C. vasquezii*, n = 130; detailed sample numbers are provided in Table 1). For each sample, spectra were acquired from three leaflets, with one measurement taken from the adaxial and one from the abaxial surface of each leaflet, resulting in six spectral readings per sample.

Spectral reflectance was measured using an ASD FieldSpec® 3 spectrometer (Analytical Spectral Devices, ASD Inc., USA), provided by the Graduate Program in Forest Sciences at INPA. The

instrument covers the 350–2500 nm wavelength range, encompassing the visible (VIS), near-infrared (NIR), and short-wave infrared (SWIR) regions. Each spectrum comprises 2,151 reflectance values, with an optical resolution of 3 nm in the VIS (350–700 nm) and NIR (700–1000 nm) regions and 8 nm in the SWIR (1000–2500 nm) region. Spectral data were recorded at 10 nm resolution. Wavelengths in the visible region (350–700 nm), regions associated with water absorption (960–980, 1170–1190, 1235–1255, 1300–1460, 1750–2030, 2040–2060, 2135–2155 and 2153–2173 nm; Thenkabail and Lyon 2016; Wang and Gamon 2019), and the noisiest edges of the sensor range (350–399 and 2401–2500 nm; Cavender-Bares *et al.,* 2016) were excluded from analysis. Each raw spectrum comprised 2,151 reflectance values, of which 1,104 were retained after these exclusions. The spectrometer was calibrated against a white reference disc (defining a 1.0 baseline, i.e., 100% reflectance) before each sample reading. Individuals with clearly noisy or erroneous readings were removed during quality screening; a small number of ambiguous readings were retained because it was not possible to determine with confidence whether the deviation reflected noise or a genuine spectral signature.

NIR reflectance spectra analysis

The six scans per sample were averaged, and the mean spectrum was first analyzed by DAPC (adegenet package; Jombart et al., 2023), following the same approach used for the morphometric data.

The same data set was submitted to Linear Discriminant Analysis (LDA), analyzed using the Leave-One-Out (LOO) cross-validation. This procedure was repeated for all samples, ensuring that each one was used for both training and testing, at different times. Once the discriminant function was generated, it was possible to predict the group to which the eliminated sample belonged, and prediction results for all samples were obtained. Finally, confusion matrices were generated from the model's successful and error predictions. The analyses were performed using the R packages "plyr" (Wickham 2023), "MASS" (Ripley et al., 2024), "vegan" (Oksanen et al., 2024), "caret" (Kuhn et al., 2023) and "NIRtools" (Perdiz 2023).

Comparison of classification algorithms across data sources

To directly compare the discriminative performance of morphometry, NIR spectroscopy, and their combination, we additionally evaluated three families of supervised classifiers on each of three data sources: morphometric variables alone (MORPHO), spectral variables alone (NIR, represented by their leading principal components), and a combined dataset concatenating both blocks for the subset of individuals with paired morphometric and spectral measurements (COMBINED).

For each data source, we compared Linear Discriminant Analysis (LDA), Partial Least Squares Discriminant Analysis (PLS-DA), and Random Forest (RF), the latter implemented with class weighting (*ranger* package; Wright and Ziegler 2017) and, separately, with random oversampling of the minority class in the training partition ("upsampled"). Combined with an "original" and a class-balanced training condition, this produced up to seven model variants per data source (21 data-source × model combinations in total). Models were evaluated with repeated 5-fold cross-validation (20 repetitions, 100 folds in total), using the same fold partitions across all model variants within a data source to allow paired comparisons.

Performance was summarized as balanced accuracy, with 95% confidence intervals obtained by bootstrap resampling of the cross-validated predictions, and as per-species recall, which we treat as the primary metrics because they are not inflated by the majority classes. Macro-F1 was retained as a secondary, contrastive metric, since gains in recall for the rare class necessarily increase false positives. It can therefore lower macro-averaged precision-sensitive metrics even when discrimination of the rare class genuinely improves.

Class imbalance was corrected in two complementary ways: for LDA, by using equal prior probabilities across species instead of sample-frequency priors; for PLS-DA and RF, by random oversampling of *C. surinamensis* in the training folds; and, for RF, additionally by class weighting in *ranger*. The effect of balancing was quantified as the paired difference (Δ) in balanced accuracy and in per-species recall between the original and balanced conditions, computed on the same cross-validation folds, with 95% confidence intervals for Δ obtained by bootstrap.

As an independent confirmation of the imbalance effect, we additionally repeated a fully separate design based on repeated undersampling: each species was resampled, without replacement, to the size of the smallest class (*C. surinamensis*), and balanced accuracy of LDA was evaluated by internal cross-validation across 100 independent undersampled replicates for each data source. Convergence between this undersampling-based estimate and the equal-prior LDA result from the main design was used as a robustness check that the imbalance effect is not an artifact of a single correction technique.

All analyses were conducted in R v.4.3.2 (R Core Team 2023), using the packages *MASS* (Ripley *et al.,* 2024), *plyr* (Wickham 2023), *vegan* (Oksanen *et al.,* 2024), *caret* (Kuhn *et al.,* 2023), *adegenet* (Jombart *et al.,* 2023), *NIRtools* (Perdiz 2023), and *ranger* (Wright and Ziegler 2017); the exact function calls used for PLS-DA and for the upsampling and bootstrap procedures should be confirmed against the analysis scripts before submission.

Results

Morphological data

Exploratory PCA of the morphometric data (Fig. 1) showed that the first two principal components explained 59.6% of total variance (PC1 = 41.2%, PC2 = 18.4%). Samples, colored by assigned species, showed some overlap, indicating similarities in a subset of characters, but also separation, particularly along PC1, suggesting that some morphological variables contribute strongly to species differentiation.

ANOVA showed statistically significant differences among species for most morphometric variables (Table 2). Petiole length and width, rachis length, mean petiolule length, and basal and terminal leaflet dimensions all had $P < 0.001$. Leaflet length/width ratios were also highly significant, while the number of leaflet pairs was moderately significant ($0.01 \leq P < 0.05$). Only the number of secondary veins showed no significant difference among species ($P \geq 0.05$).

Table 2. Analysis of variance (ANOVA) results for morphometric variables, showing degrees of freedom, sums of squares, mean squares, F values, P values, and significance level for each variable. Significance codes: ** *P < 0.001;* *P < 0.01;* P < 0.05; ns P ≥ 0.05.

Variable	df	F	P	Sig.
Petiole length (cm)	2	26.1	7.11e-11	*
Petiole width (cm)	2	14.3	1.42e-06	*
Rachis length (cm)	2	29.5	4.49e-12	*
Leaflet pairs (n)	2	3.2	0.0403	
Mean petiolule length (cm)	2	23.2	7.14e-10	*
Basal leaflet length (cm)	2	6.8	0.00132	
Basal leaflet width (cm)	2	19.9	1.18e-08	*
Terminal leaflet length (cm)	2	9.9	7.36e-05	*
Terminal leaflet width (cm)	2	13.1	4.22e-06	*
Secondary veins (n)	2	1.9	0.145	ns
Basal leaflet length/width	2	7.4	0.000774	*
Terminal leaflet length/width	2	8.0	0.000423	*

C. surinamensis and *C. vasquezii* had, on average, longer and wider petioles than *C. guianensis*. *C. vasquezii* stood out for its longer rachis (mean 61.7 cm) and higher mean number of leaflet pairs (6.2), consistent with its longer rachis. *C. surinamensis* and *C. vasquezii* also had longer basal and terminal leaflets than *C. guianensis*: for basal leaflets, *C. vasquezii* had the highest mean (16.4 cm) and *C. guianensis* the lowest (13.8 cm); for terminal leaflets, *C. vasquezii* again had the highest mean (26.3 cm) against 21.8 cm in *C. guianensis*. Leaflet width followed a similar pattern, while the mean number of secondary veins was similar across species (close to 10 in all three), consistent with the non-significant ANOVA result. Leaflet length/width ratios showed a comparatively narrow range across species (basal leaflets: 0.50–0.58; terminal leaflets: 0.36–0.42).

Table 3. Percentage distribution of leaflet shape, apex, and base categories among *C. guianensis*, *C. surinamensis,* and *C. vasquezii* specimens.

Leaflet character	*C. guianensis*	*C. surinamensis*	*C. vasquezii*
Oblong	6.3%	100%	97.9%
Elliptic	93.7%	—	2.1%
Acuminate apex	98.9%	—	—
Obtuse apex	1.1%	100%	100%
Acute base	17.9%	—	2.1%
Rounded base	6.3%	100%	96.9%
Cuneate base	75.8%	—	1.0%

Leaflet shape clearly distinguished *C. guianensis* from the other two species: *C. guianensis* was predominantly elliptic (93.7%), acuminate (98.9%), and cuneate-based (75.8%), whereas *C. surinamensis* was uniformly oblong, obtuse, and rounded-based (100% in all three characters), and *C. vasquezii* showed a similar, though not identical, pattern (97.9% oblong, 100% obtuse, and 96.9% rounded-based, with small proportions of acute or cuneate bases and elliptic leaflets).

Hierarchical clustering based on Gower dissimilarity confirmed these exploratory patterns (Fig. 2): *C. guianensis* formed a clearly separated cluster from a second group combining *C. surinamensis* and *C. vasquezii* samples, and no consistent sub-clustering distinguished *C. surinamensis* from *C. vasquezii within that second group*.

DAPC provided better visual resolution than the dissimilarity-based clustering, particularly between *C. surinamensis* and *C. vasquezii* (Fig. 3). Applied to all three species, DAPC produced three distinguishable clusters, with the first discriminant axis (DAPC1) accounting for 73% of the discriminant variance. Nonetheless, considerable overlap of individual points suggests that part of the observed variation reflects differences among individuals rather than clear-cut differences among species. Restricting the analysis to *C. surinamensis* and *C. vasquezii* (Fig. 4) showed overlapping but distinguishable distributions along the discriminant axis, indicating that although morphometric variability overlaps between the two species, mean differences remain considerable. The variables contributing most to the DAPC axes were mean petiolule length, petiole length and width, and basal leaflet length (Fig. 5).

Spectral data

Mean reflectance spectra of the three species showed broadly similar shapes, with the largest differences occurring in the near-infrared (700–1300 nm) and short-wave infrared (1500–2500 nm) regions, likely reflecting differences in cell structure and chemical composition among species (Fig. 6).

DAPC applied to the spectral data indicated separation among groups, with DAPC1 explaining 86% and DAPC2 14% of the variance (Fig. 7). *C. guianensis* was more dispersed along DAPC1, while *C. surinamensis* and *C. vasquezii* overlapped more extensively, reflecting spectral similarity between these two species that parallels the pattern observed in the morphometric data.

The result of the linear discriminant analysis (LDA-LOO), visualized in the form of a confusion matrix, showed that the correct predictions in relation to the model were 90% in *Carapa guianensis,* 93% in *C. vasquezii,* and 68% in *C. surinamensis* (Fig. 8). The matrix also showed that the model tends to confuse *C. surinamensis* with the other species, suggesting overlap or similarity between the spectral characteristics from the model's perspective; because of the low number of samples, it was not possible to predict this class correctly. In addition, *C. vasquezii* had the highest rate of correct predictions, probably because it is the most prevalent class in the dataset. Classes with the highest number of samples tend to influence the model's predictions more, which may make it more likely to predict this class when it is uncertain. Therefore, the results showed the potential of spectral data but were not definitive in delimiting the species.

Comparison of morphometry, NIR and combined data across classification algorithms

Across the 100 cross-validation folds run for the 21 data-source × model combinations, morphometry alone (MORPHO) was consistently the best-performing data source for LDA and PLS-DA, both under balanced-prior/upsampled training conditions (Table 4). NIR alone and the combined dataset (COMBINED) performed essentially equivalently to one another and about three percentage points of balanced accuracy below MORPHO for LDA and PLS-DA, with confidence intervals that did not fully overlap those of MORPHO, indicating a consistent rather than incidental difference. RF was the weakest algorithm on every data source, roughly four to five percentage points below LDA/PLS-DA, and was also the least sensitive to class balancing.

Table 4. Mean Balanced Accuracy (95% bootstrap CI) under the class-balanced training condition, by data source and classification algorithm.

Data source	LDA	PLS-DA	RF (weighted)	RF (upsampled)
MORPHO	0.678 (0.664–0.693)	0.671 (0.657–0.685)	0.629 (0.619–0.640)	0.651 (0.640–0.664)
NIR	0.651 (0.640–0.661)	0.648 (0.638–0.657)	0.614 (0.603–0.624)	0.609 (0.598–0.620)
COMBINED	0.652 (0.641–0.662)	0.648 (0.638–0.658)	0.614 (0.605–0.624)	0.611 (0.600–0.623)

Notably, combining NIR with morphometry did not improve discrimination relative to morphometry alone: COMBINED generally tracked NIR rather than MORPHO. Thus, under the data-fusion and classification strategies used here, adding the NIR block provided no measurable gain in discriminative performance over morphometry alone. This may reflect limited additional taxonomic signal in the NIR data, redundancy between the two data sources, or interference from the high-dimensional spectral block, even after dimensionality reduction, when both sources are combined within the same LDA/PLS-DA framework. These alternatives cannot be distinguished from classification performance alone.

Effect of class balancing on balanced accuracy

Correcting for class imbalance produced a consistent, statistically clear improvement in balanced accuracy for LDA and PLS-DA across all three data sources (Table 5; all 95% confidence intervals for Δ excluded zero). For RF, class weighting had no detectable effect on balanced accuracy in any data source (all Δ confidence intervals spanned zero).

Table 5. Mean paired change in Balanced Accuracy (ΔBA) after correcting for class imbalance, by data source and algorithm, computed on the same cross-validation folds. The RF row shows the class-weighted variant.

Data source	Algorithm	ΔBA	95% CI	Significant?
MORPHO	LDA	+0.032	0.023 to 0.042	Yes
MORPHO	PLS-DA	+0.048	0.036 to 0.060	Yes
MORPHO	RF (weighted)	+0.004	−0.001 to 0.009	No
NIR	LDA	+0.024	0.013 to 0.034	Yes
NIR	PLS-DA	+0.026	0.018 to 0.035	Yes
NIR	RF (weighted)	−0.001	−0.006 to 0.004	No
COMBINED	LDA	+0.024	0.013 to 0.034	Yes
COMBINED	PLS-DA	+0.026	0.017 to 0.034	Yes
COMBINED	RF (weighted)	−0.002	−0.006 to 0.001	No

Recall of C. surinamensis: the central result

The most consistent result of this study concerns the recall of *C. surinamensis*, the least-sampled species. Under original training conditions, *C. surinamensis* was almost invisible to the classifiers: recall ranged from near 0 (PLS-DA on NIR and on COMBINED) to 0.22 (LDA on MORPHO) (Table 6). Applying equal priors (LDA) or upsampling in training (PLS-DA) raised recall to 0.41–0.55, with 95% confidence intervals for the change that excluded zero in every case, indicating a real and substantial, not merely stochastic, effect.

Table 6. Recall of *C. surinamensis* under original and class-balanced training conditions, by data source and algorithm.

Data source	Algorithm	Recall (original)	Recall (balanced)	ΔRecall
MORPHO	LDA	0.223	0.527	+0.304
MORPHO	PLS-DA	0.070	0.550	+0.480
MORPHO	RF (weighted)	0.103	0.135	+0.032
MORPHO	RF (upsampled)	0.103	0.331	+0.228
NIR	LDA	0.085	0.490	+0.405
NIR	PLS-DA	0.006	0.409	+0.403
NIR	RF (weighted)	0.051	0.070	+0.019
COMBINED	LDA	0.087	0.490	+0.403
COMBINED	PLS-DA	0.008	0.420	+0.412
COMBINED	RF (weighted)	0.053	0.070	+0.018

This pattern supports the hypothesis that poor original performance was driven largely by class imbalance rather than a complete absence of discriminative signal: the classifiers could capture information distinguishing *C. surinamensis* but had little "incentive" to predict this class when its low training frequency biased decisions toward the majority species.

The trade-off: recall loss in the majority species

The gain in *C. surinamensis* recall came predominantly at the expense of *C. guianensis* recall, which dropped by 13 to 26 percentage points across data sources and algorithms, while *C. vasquezii* recall remained comparatively stable. This trade-off also explains why Macro-F1 sometimes decreased after balancing (e.g., PLS-DA on COMBINED: 0.642 to 0.516): the sensitivity gain for the rare class increased false positives, lowering average precision. For this reason, balanced accuracy and per-species recall are reported here as the primary performance metrics, with Macro-F1 retained only as a secondary, contrastive indicator, since it can mask the very effect this analysis was designed to detect.

Random Forest: class weighting versus upsampling

Random Forest behaved differently from LDA and PLS-DA: class weighting in *ranger* had almost no effect on *C. surinamensis* recall (+0.02 to +0.03), whereas upsampling in training was considerably more effective (Table 6), raising recall from 0.05–0.10 to 0.21–0.33 depending on data source. Even so, upsampled RF remained well below balanced LDA/PLS-DA (recall ≈ 0.20–0.33 versus ≈ 0.41–0.55). Consequently, when reporting a "corrected" RF result, the upsampled variant, not the weighted variant, should be used, with the caveat that RF, even corrected, underperforms LDA/PLS-DA for the rare class in this dataset.

Independent confirmation: repeated undersampling

Using a fully independent design, repeatedly resampling each species down to the size of the smallest class and validating by internal cross-validation across 100 replicates, LDA Balanced Accuracy was 0.655 (95% CI 0.647–0.663) for MORPHO, 0.643 (0.636–0.650) for COMBINED, and 0.642 (0.635–0.649) for NIR (Table 7). These values closely match the equal-prior LDA results from the main design (0.678, 0.652, and 0.651, respectively), showing that two entirely different balancing strategies (undersampling versus equal priors) converge on a similar performance range. This convergence supports the conclusion that the imbalance effect is a genuine property of the data rather than an artifact of a single correction technique.

Table 7. Balanced Accuracy of LDA under an independent repeated-undersampling design (100 replicates), by data source.

Data source	Mean BA	95% CI
MORPHO	0.655	0.647–0.663
COMBINED	0.643	0.636–0.650
NIR	0.642	0.635–0.649

Discussion

The concept of a “species complex” refers to groups of morphologically similar, frequently confused species that generate taxonomic uncertainty (Fišer *et al.,* 2018; Damasco *et al.,* 2019). In *Carapa*, the

morphological similarity among *C. surinamensis*, *C. vasquezii* and *C. guianensis*, together with the phenotypic variability observed here, illustrates the difficulty of delimiting these species from vegetative characters alone.

Kenfack (2011a) previously showed that vegetative characters alone were insufficient to separate *C. surinamensis* from *C. vasquezii*. Reliably delimiting the three species required a combination of vegetative and reproductive characters. The present results are consistent with, rather than contradictory to, that conclusion: DAPC recovered statistically distinguishable group means for these two species along the discriminant axis, and the cross-validated classifiers achieved recall well above chance (1/3) once class imbalance was corrected, but neither analysis produced clean, non-overlapping separation at the level of individual specimens. Separation of group centroids and classification of individual specimens are different standards of evidence, and only the latter is directly relevant to routine specimen identification.

Petiolule length, petiole length and width, and leaflet base shape were, in that order, the most important vegetative characters distinguishing species in this study, a result driven mainly by DAPC. This agrees with earlier work on the genus as a whole, which identified petiole length relative to total leaf length and the number of leaflets per leaf as the most distinctive morphometric characters (Kenfack 2011a, 2011b) using multivariate analysis. The clear morphological separation of *C. guianensis* from the other two species, evident here in both the dendrogram and the DAPC ordination, underscores the diagnostic value of leaflet shape for distinguishing *C. guianensis* from *C. surinamensis*/*C. vasquezii*, even though it does not resolve the latter pair as clearly.

A separate study focused on seed and seedling characters found that the shape of the first cataphylls and the length and width of the seed hilum effectively distinguished the three species (Gama 2019), pointing to alternative characters usable for field identification in the absence of flowers. These characters were not available for the specimens used here. Likewise, geometric morphometrics based on leaflet outline landmarks, a method that has proved useful for distinguishing complex Amazonian plant taxa elsewhere (Prata *et al.,* 2018; Holanda *et al.,* 2021; Vasconcelos 2024), was not analyzed in the present study, although the underlying image data were collected and remain available for future analysis.

The overlap observed among individual DAPC scores may reflect within-species variability, possibly driven by phenotypic plasticity across the range of environments these species occupy. This variability underscores the general challenge of relying solely on morphological characters for species identification (Padial *et al.,* 2010; Pante *et al.,* 2015; Draper *et al.,* 2020).

Phenotypic and, by extension, spectral variability is a plausible explanation for the comparatively weak, imbalance-sensitive performance obtained here for *C. surinamensis*, given that leaf reflectance

spectroscopy has elsewhere achieved correct identification rates above 96% for other Amazonian plant complexes (Durgante *et al.,* 2013; Lang *et al.,* 2017; Prata *et al.,* 2018; Damasco *et al.,* 2019). Other factors that plausibly reduce NIR performance in this dataset include imprecise taxonomic identification of the reference material, since model calibration requires a solid taxonomic basis to maximize accuracy, and, as our results show, the marked disparity in sample size among species.

That combining NIR with morphometry (COMBINED) did not outperform morphometry alone (MORPHO) was, on reflection, not an intuitive outcome: it suggests either that leaf spectra carry comparatively little information that is complementary to vegetative morphometry for separating these three species, or that the higher dimensionality of the NIR block, even after data reduction, adds noise that competes with the morphometric signal when the two blocks are combined linearly, as in LDA/PLS-DA. Formal multi-block fusion methods that explicitly weight complementary blocks (rather than simple concatenation) might recover a combined-data advantage that linear concatenation obscures and are identified below as a priority for future analysis.

The systematically weaker performance of Random Forest relative to LDA/PLS-DA, and its comparative insensitivity to class weighting, is consistent with the relatively small, moderately high-dimensional dataset used here: linear discriminant methods can be more sample-efficient than tree ensembles under these conditions, and resampling-based correction (oversampling) evidently interacted more effectively with RF's decision-tree structure than did weighting the loss function directly. This is a useful applied finding for future work in this system: if Random Forest is used, address class imbalance through resampling rather than class weighting alone.

Taken together, the balancing experiments support class imbalance as a major, but not sole, cause of the historically poor recognition of *C. surinamensis*. Recall recovered substantially (from near zero to 0.41–0.55) once imbalance was corrected, and this recovery was replicated by an entirely independent undersampling design, arguing against a statistical artifact. At the same time, corrected recall plateaued well below 1.0, and its gain came at a real cost to *C. guianensis* recall. This residual difficulty is compatible with genuine morphological and spectral similarity between *C. surinamensis* and *C. vasquezii*, plausibly related to their close phylogenetic relationship (Koenen *et al.,* 2015) and to reports of hybridization and introgression at the *C. guianensis*/*C. surinamensis* boundary elsewhere in the range of the genus (Duminil *et al.,* 2012; Scotti-Saintagne *et al.,* 2013), rather than with a purely technical shortcoming of the classifiers or the sampling design.

Species discrimination versus taxonomic delimitation

It is important to be explicit about what these results do and do not demonstrate. The analyses reported here address species *discrimination*: the statistical ability to assign specimens to groups that were already defined by prior taxonomic determination, using vegetative morphometry and/or NIR spectra.

They do not constitute, and should not be read as, a test of species *delimitation* in the taxonomic sense, that is, of whether *C. guianensis*, *C. surinamensis* and *C. vasquezii* are correctly circumscribed as three distinct evolutionary lineages. Delimitation requires integrating multiple, largely independent lines of evidence (reproductive morphology, population genetics or genomics, geography and, where available, experimental crossing or reproductive-isolation data) within a formal taxonomic framework (Padial *et al.,* 2010; Pante *et al.,* 2015). High or low classification accuracy under the taxonomy as currently applied cannot, by itself, confirm or refute that taxonomy, because misclassification could equally arise from mislabelled reference vouchers, character overlap within otherwise valid species, or genuinely blurred species boundaries, and the present design cannot distinguish among these possibilities.

What the present results do support is the practical value of an integrative morphometric and NIR pipeline as an identification and collection-curation tool: a way to flag likely misidentified vouchers, prioritize specimens for expert review, and guide targeted collecting of the under-sampled *C. surinamensis*, without asserting a taxonomic conclusion about the validity of the three species as currently circumscribed.

Limitations

Several limitations should be considered. *C. surinamensis* was represented by a relatively small and geographically restricted sample, mostly from planted stands, limiting statistical power and the generalisability of the balancing results. Ontogenetic effects could not be controlled, and reproductive characters important for species delimitation (Kenfack 2011a) were unavailable.

The combined analyses were restricted to paired samples and used simple concatenation rather than a dedicated data-fusion method. NIR reproducibility across sessions or operators was not formally assessed. Finally, because model evaluation relied on the prior taxonomic identifications used as training labels, classification errors cannot be unequivocally distinguished from potential errors in the reference identifications.

Conclusion

Using *Carapa* as a case study, this study compared leaf morphometry, NIR reflectance spectroscopy, and their combination for discriminating three morphologically similar Amazonian tree species and tested whether sample imbalance contributes to the poor recognition of the least represented species, *C. surinamensis*.

Morphometry was the strongest single source of discriminative information in this dataset, driven mainly by petiole, petiolule, and leaflet dimensions and leaflet shape. NIR spectroscopy also contained

a taxonomically informative signal, but neither NIR alone nor its combination with morphometry improved discrimination over morphometry alone. Under the modelling framework used here, therefore, adding spectral information provided no measurable predictive advantage over morphological measurements. LDA and PLS-DA generally performed better than Random Forest across data sources and balancing strategies.

Class imbalance explained a substantial, but not complete, component of the poor recognition of *C. surinamensis*. Correcting for imbalance markedly increased its recall in LDA and PLS-DA, with statistically supported gains, while independent undersampling confirmed that the relative advantage of morphometry persisted when class sizes were equalized. Importantly, confusion matrices showed that *C. surinamensis* was predominantly misclassified as *C. guianensis* in the original dels and models and that this asymmetry was strongly reduced after balancing. Thus, the original classification failure cannot be attributed primarily to overlap with *C. vasquezii*. Nevertheless, recall remained moderate after balancing, indicating residual limitations in species separability that cannot be explained by class prevalence alone.

These findings concern species discrimination under the taxonomy currently applied rather than taxonomic *delimitation*, which would require independent reproductive, molecular, and geographic evidence within an integrative taxonomic framework. Within this scope, morphometric and NIR approaches provide complementary opportunities for specimen screening, although morphometry currently offers the stronger discriminatory performance. Rather than replacing expert identification, these tools may help curate herbarium collections, identify uncertain or atypical specimens for reassessment, and prioritize material for integrative taxonomic study. This framework could be extended to other morphologically similar Amazonian tree groups facing comparable identification challenges.

Acknowledgments

The authors thank the institutions and individuals who supported this study, including the herbarium of the National Institute for Amazonian Research (INPA), the Biological Dynamics of Forest Fragments Project (PDBFF) and the Biodiversity Research Program (PPBIO), and the Graduate Program in Forest Sciences of INPA for the loan of the ASD FieldSpec© 3 spectrometer. Funding is acknowledged separately below.

Authors' Contributions

Semírian Campos Amoêdo was responsible for study conception and design, data collection, statistical analyses, interpretation of the results, and initial drafting of the manuscript. Caroline da Cruz Vasconcelos contributed to the statistical analyses, interpretation of the results, and critical revision of

the manuscript. Isolde Dorothea Kossmann Ferraz contributed to the study conception and critical revision of the intellectual content. All authors read and approved the final version of the manuscript.

Conflict of Interest

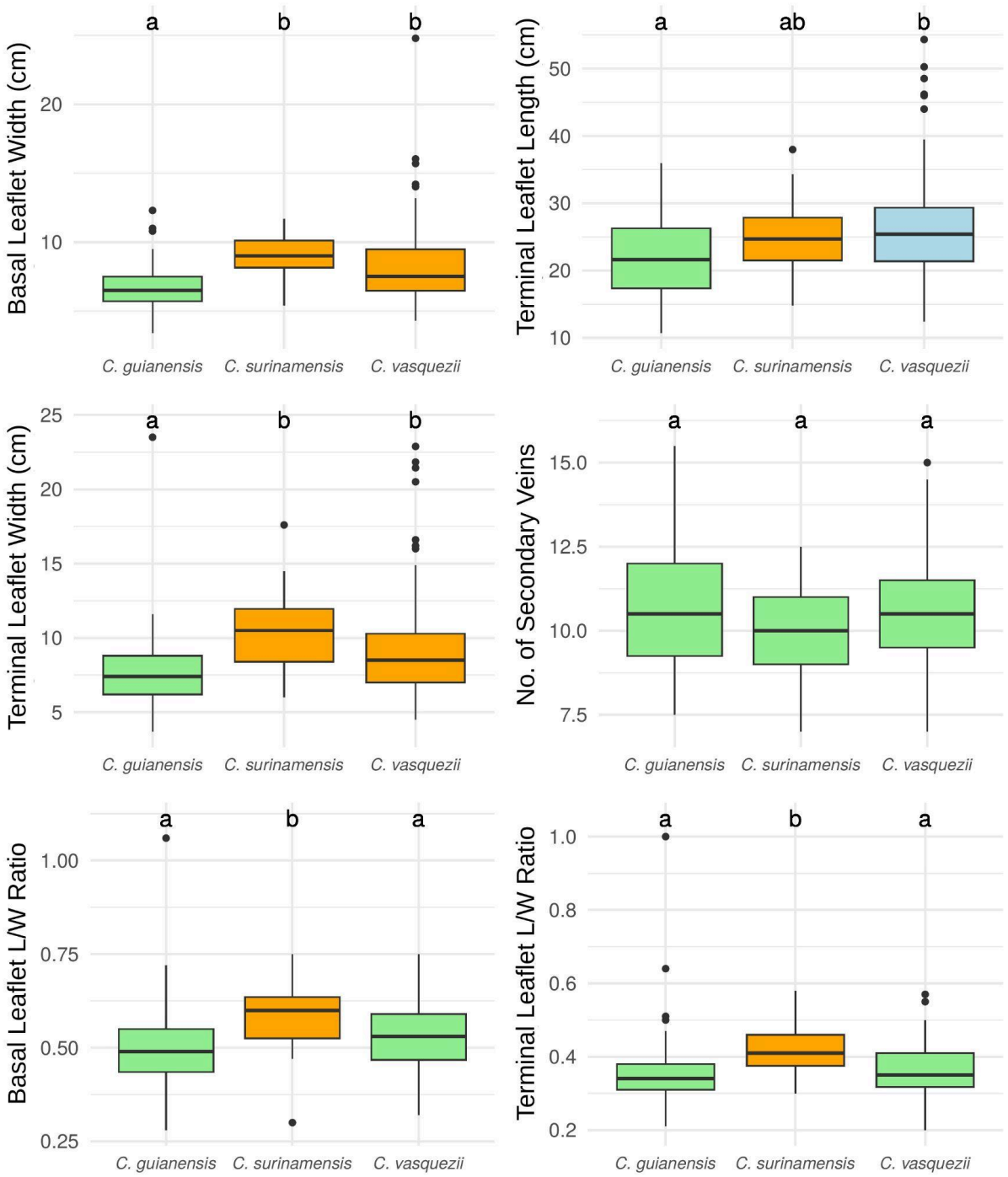
The authors declare that there are no conflicts of interest related to the publication of this manuscript.

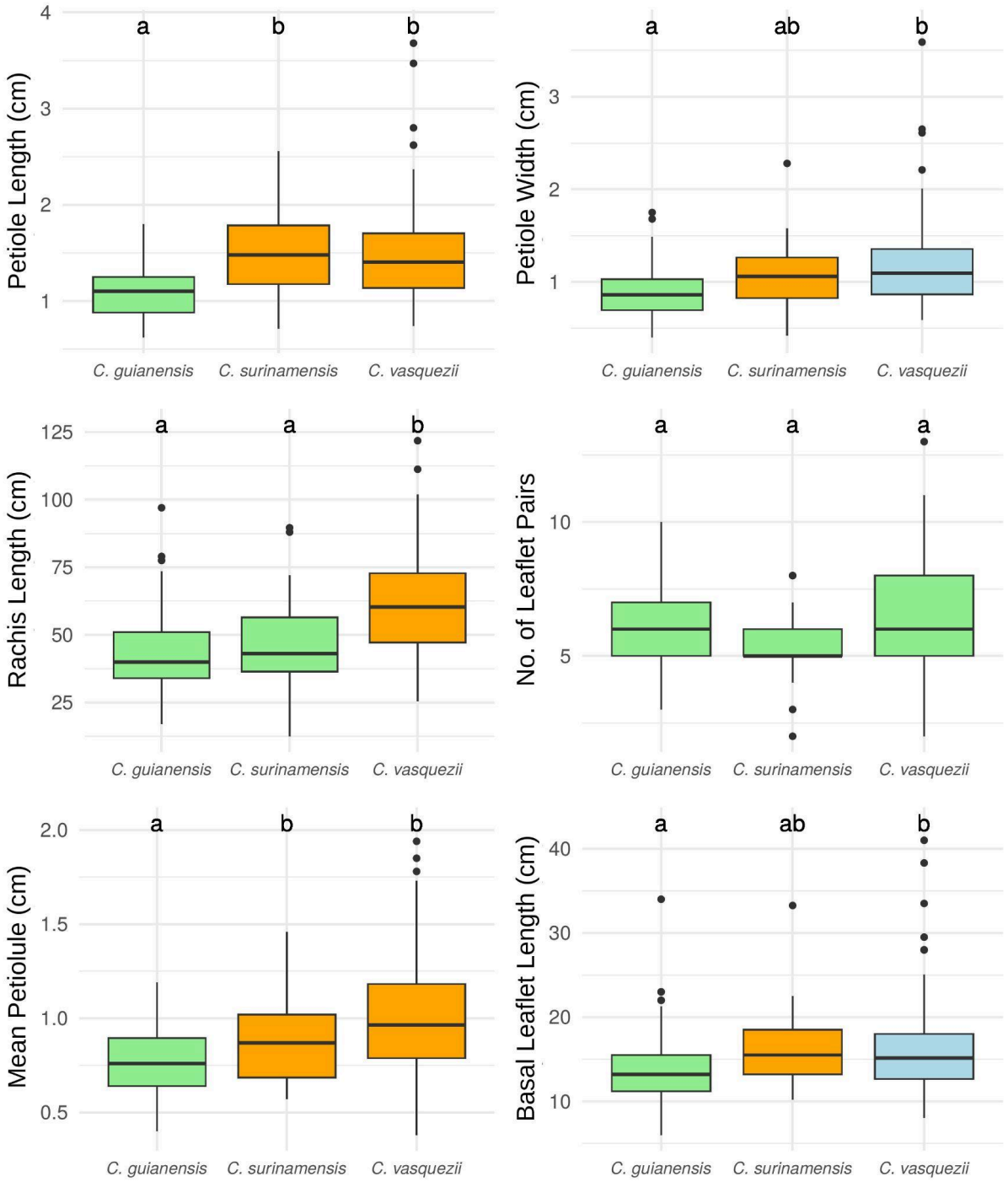
Data Availability

The data that support the findings of this study are available, upon reasonable request from the corresponding author.

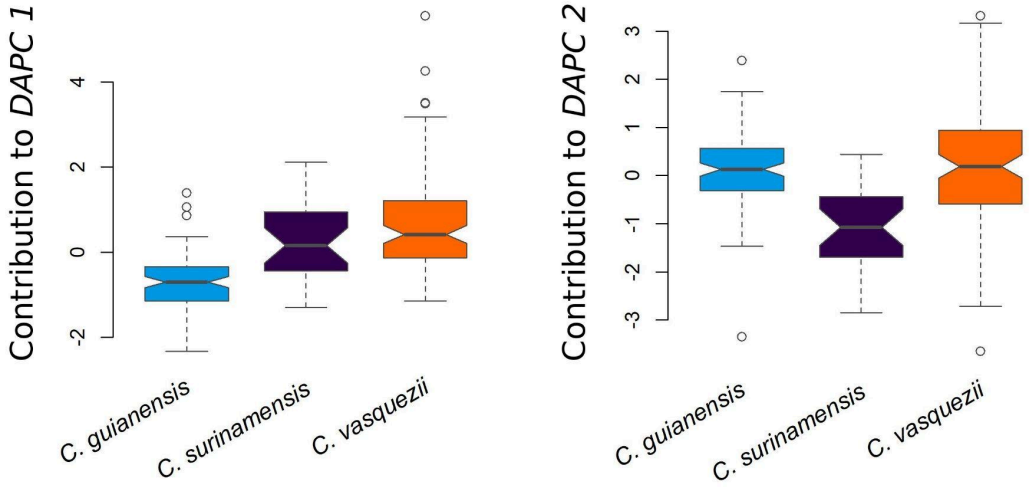
Supplementary Material

Supplementary Figure S1. Full set of per-variable ANOVA/Tukey boxplots for the 12 morphometric variables (panels A and B).


Basal Leaflet Width (cm)
a
b
b
20
10
C. guianensis
C. surinamensis
C. vasquezii
Terminal Leaflet Length (cm)
a
ab
b
50
40
30
20
10
C. guianensis
C. surinamensis
C. vasquezii
Terminal Leaflet Width (cm)
a
b
b
25
20
15
10
5
C. guianensis
C. surinamensis
C. vasquezii
No. of Secondary Veins
a
a
a
15.0
12.5
10.0
7.5
C. guianensis
C. surinamensis
C. vasquezii
Basal Leaflet L/W Ratio
a
b
a
1.00
0.75
0.50
0.25
C. guianensis
C. surinamensis
C. vasquezii
Terminal Leaflet L/W Ratio
a
b
a
1.0
0.8
0.6
0.4
0.2
C. guianensis
C. surinamensis
C. vasquezii


Petiole Length (cm)
4
3
2
1
a
b
b
C. guianensis
C. surinamensis
C. vasquezii
Petiole Width (cm)
3
2
1
a
ab
b
C. guianensis
C. surinamensis
C. vasquezii
Rachis Length (cm)
125
100
75
50
25
a
a
b
C. guianensis
C. surinamensis
C. vasquezii
No. of Leaflet Pairs
10
5
a
a
a
C. guianensis
C. surinamensis
C. vasquezii
Mean Petiolule (cm)
2.0
1.5
1.0
0.5
a
b
b
C. guianensis
C. surinamensis
C. vasquezii
Basal Leaflet Length (cm)
40
30
20
10
a
ab
b
C. guianensis
C. surinamensis
C. vasquezii

Supplementary Figure S2. DAPC variable-contribution plots for discriminant axes 1 and 2 (morphometric data).



Supplementary Table S1. Complete descriptive statistics (minimum, maximum, mode, mean, standard deviation, Tukey groups) for the 12 morphometric variables by species.

	C. guianensis						*C. surinamensis*						*C. vasquezii*					
Morphometric variables	min	max	mode	mean	sd		min	max	mode	mean	sd		min	max	mode	mean	sd	
1. Petiole length (cm)	0.62	1.8	1.1	1.09	0.3	a	0.71	2.56	1.22	1.50	0.5	b	0.74	3.68	1.63	1.53	0.5	b
2. Petiole width (cm)	0.4	1.7	1.23	0.90	0.3	a	0.42	2.28	1.1	1.07	0.4	ab	0.59	3.59	1.09	1.20	0.5	b
3. Rachis length (cm)	17	97	34.5	42.76	14.0	a	12.5	89.61	40.5	46.91	18.0	a	25.5	121.7	48	61.67	19.9	b
4. Number of leaflet pairs	3	10	6	5.80	1.3	a	2	8	5	5.41	1.4	a	2	13	5	6.23	2.0	a
5. Mean petiolule length (cm)	0.4	1.1	0.74	0.77	0.2	a	0.57	1.46	0.67	0.89	0.2	b	0.38	1.94	0.94	1.01	0.3	b
6. Basal leaflet length (cm)	6	34	15	13.79	4.2	a	10.2	33.26	19.5	16.24	4.8	ab	8	41	14.5	16.35	5.8	b
7. Basal leaflet width (cm)	3.4	12.3	6	6.64	1.6	a	5.4	11.7	8.3	9.07	1.6	b	4.3	24.79	6	8.43	2.9	b
8. Terminal leaflet length (cm)	10.7	36	21	21.81	5.9	a	14.8	37.98	28	24.84	5.3	ab	12.4	54.34	18.5	26.26	8.2	b
9. Terminal leaflet width (cm)	3.7	23.5	10	7.60	2.4	a	6	17.6	7.7	10.43	2.6	b	4.5	22.88	7	9.32	3.6	b
10. Number of secondary veins	7.5	15.5	10	10.72	1.9	a	7	12.5	9	9.98	1.3	a	7	15	11	10.44	1.7	a
11. Basal leaflet length/width ratio	0.28	1.1	0.49	0.50	0.1	a	0.3	0.75	0.6	0.58	0.1	b	0.32	0.75	0.5	0.53	0.1	a
12. Terminal leaflet length/width ratio	0.21	1	0.37	0.36	0.1	a	0.3	0.58	0.38	0.42	0.1	b	0.2	0.57	0.32	0.36	0.1	a

Funding Information

SCA acknowledges support from the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) through a doctoral scholarship, and the Fundação de Amparo à Pesquisa do Estado do Amazonas (FAPEAM) through the POSGRAD Program (2017 and 2019), which provided financial support for field assistance. CCV acknowledges the research grant awarded by the International Association for Plant Taxonomy (IAPT, 2025 grant round).

Table and Figure Captions

Figures:

Fig. 1. Principal Component Analysis of morphometric variables for *Carapa* spp. samples from the Amazonian lowlands; colors indicate species assignment.

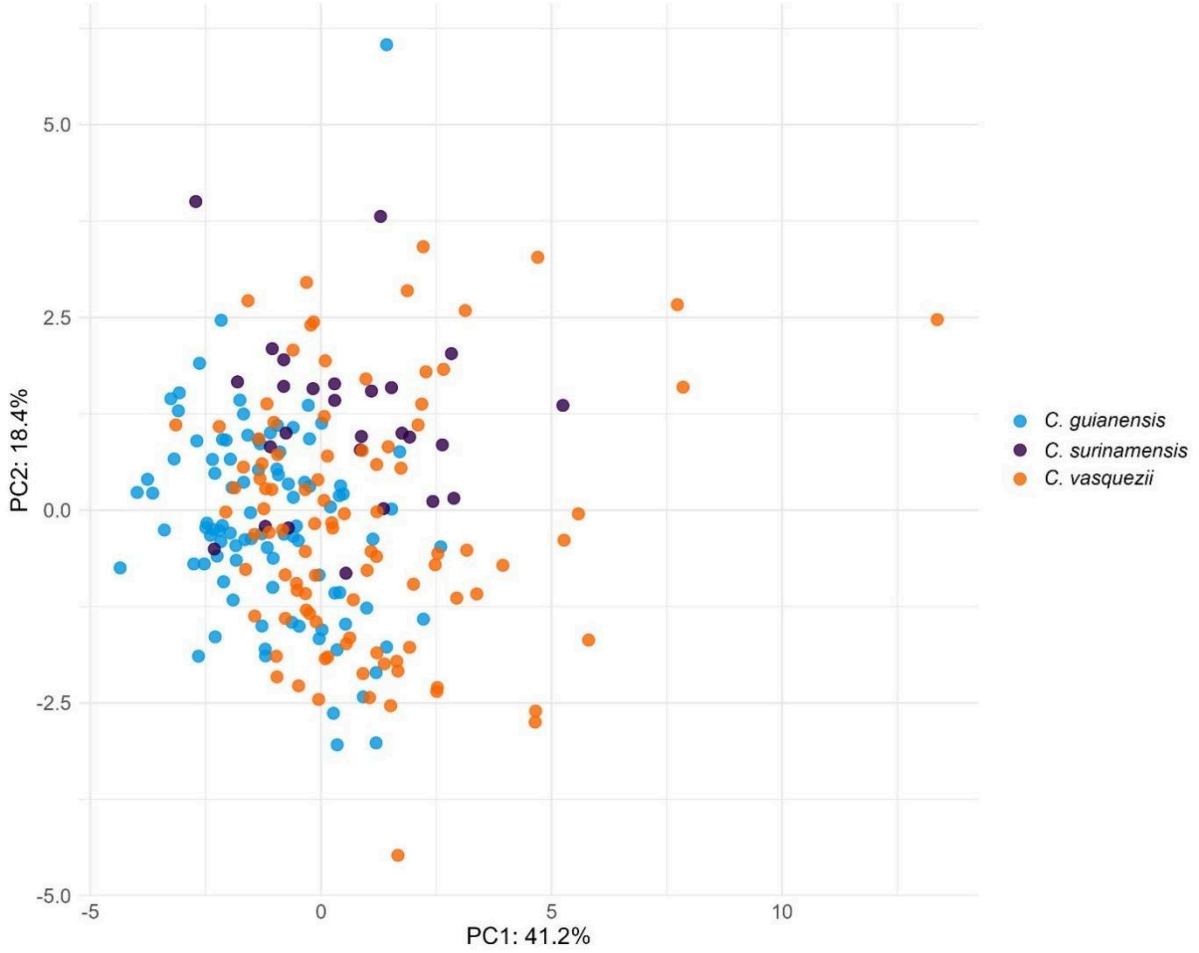


Fig. 2. Hierarchical clustering dendrogram from quantitative and qualitative morphometric data, using Gower dissimilarity and Ward's ("ward.D2") clustering method. *Carapa guianensis* (blue), *C. surinamensis* (purple), and *C. vasquezii* (orange).

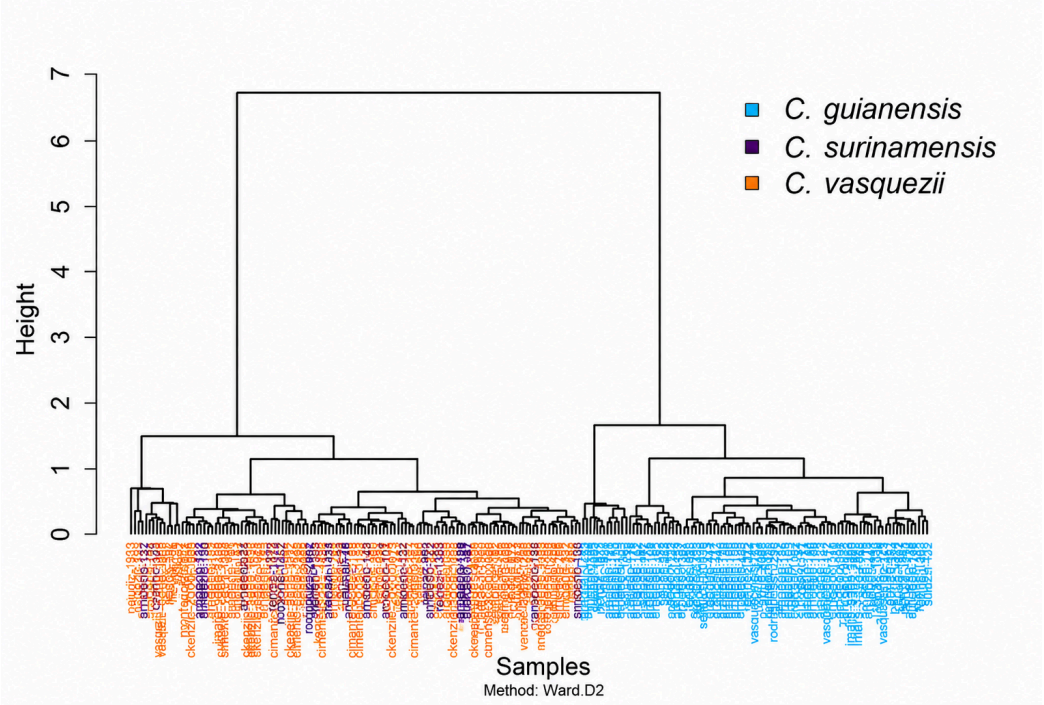


Fig. 3. Discriminant Analysis of Principal Components (DAPC) applied to morphometric data for the three *Carapa* species.

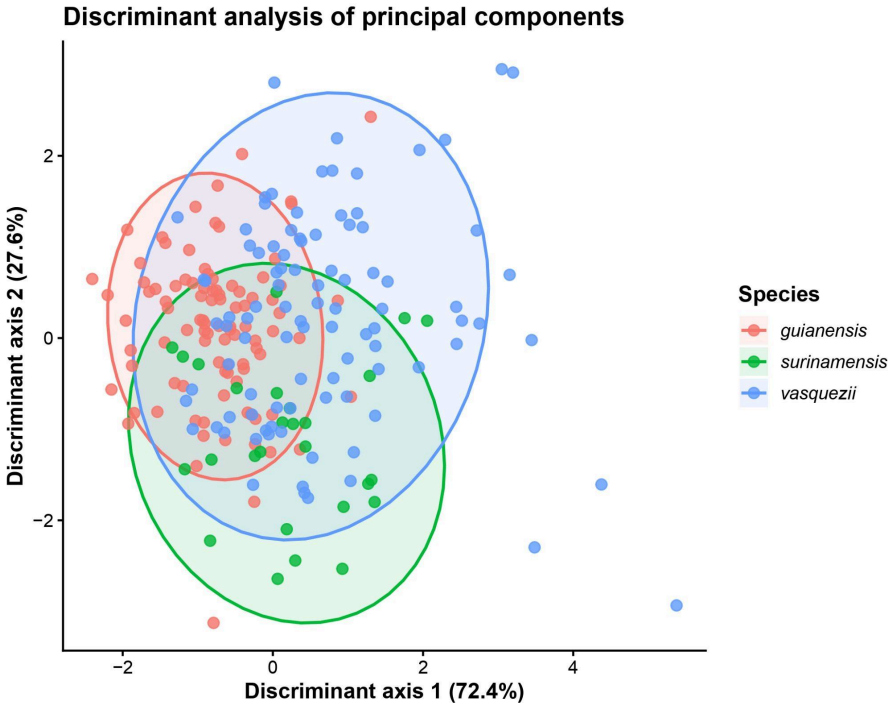


Fig. 4. DAPC showing the distinction between *C. surinamensis* and *C. vasquezii* based on morphometric characters; curves represent species distributions along the discriminant function.

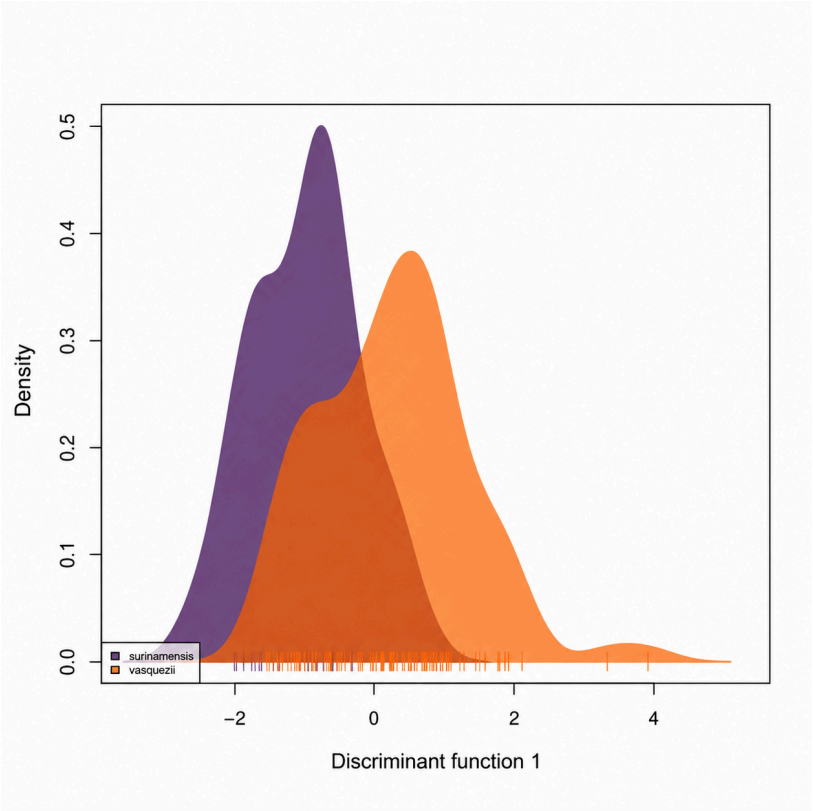


Fig. 5. Contribution of morphometric variables to DAPC axes 1 and 2.

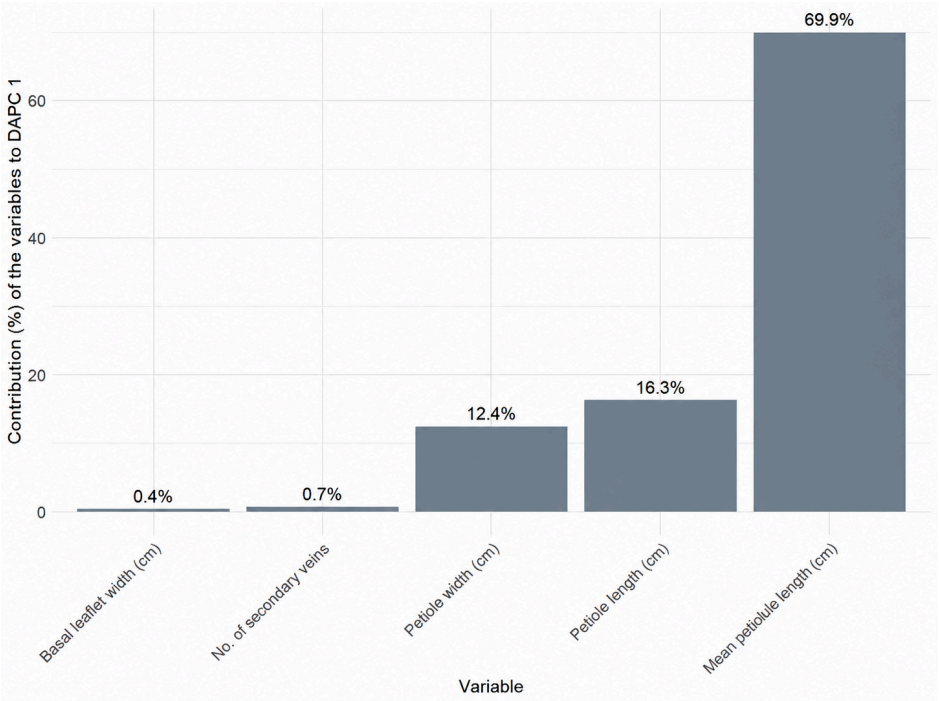


Fig. 6. Mean leaf reflectance spectral signatures of *Carapa* spp. samples.

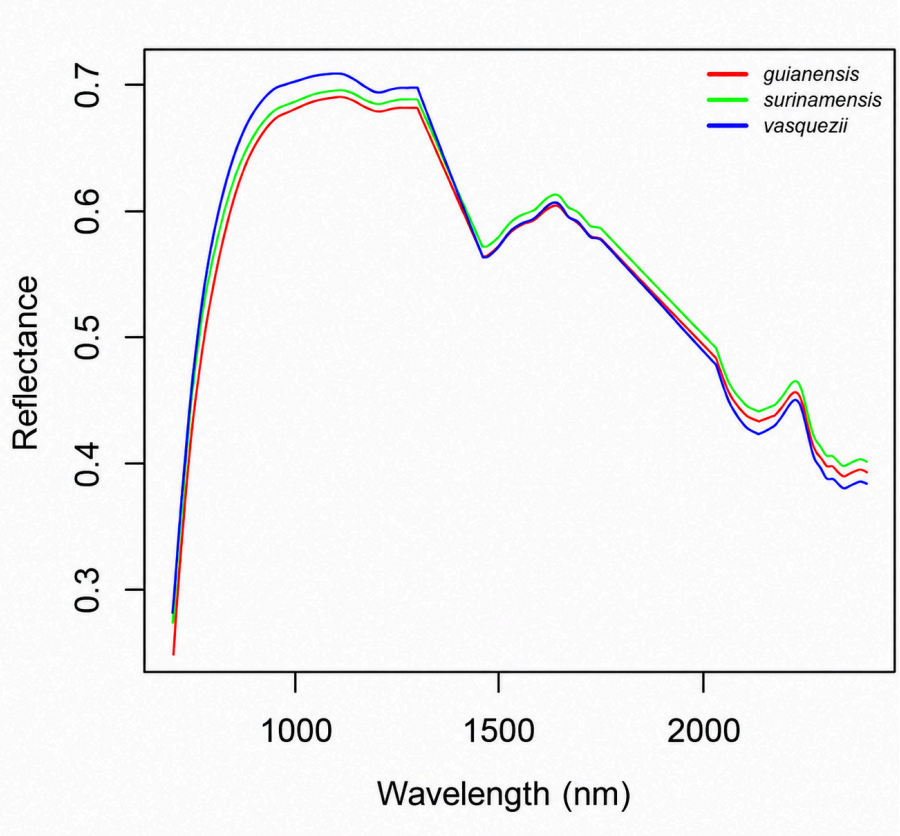


Fig. 7. DAPC applied to spectral data for the three *Carapa* species.

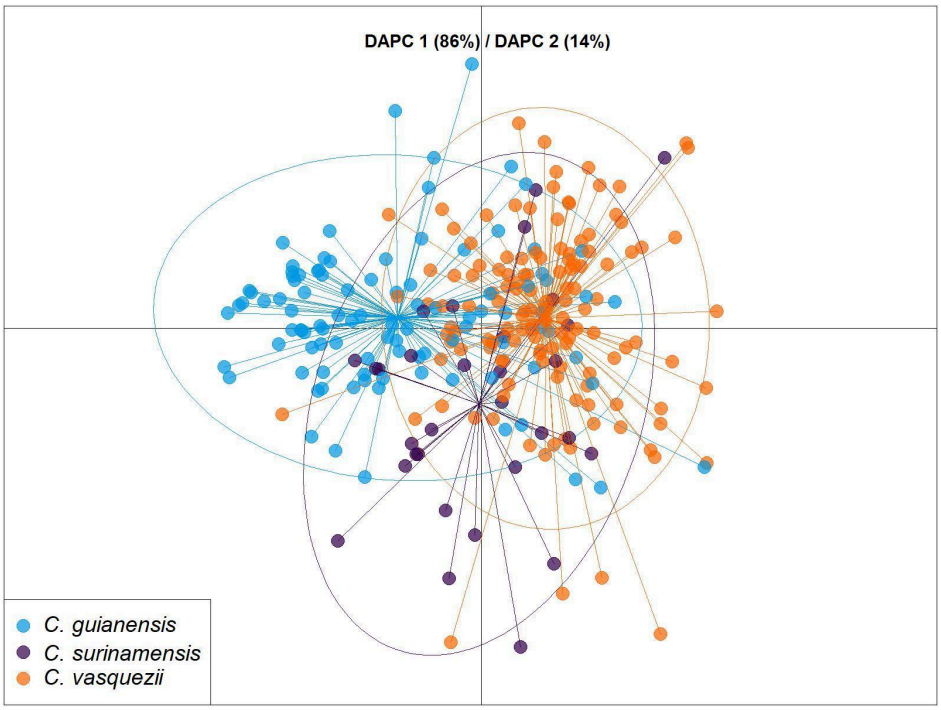
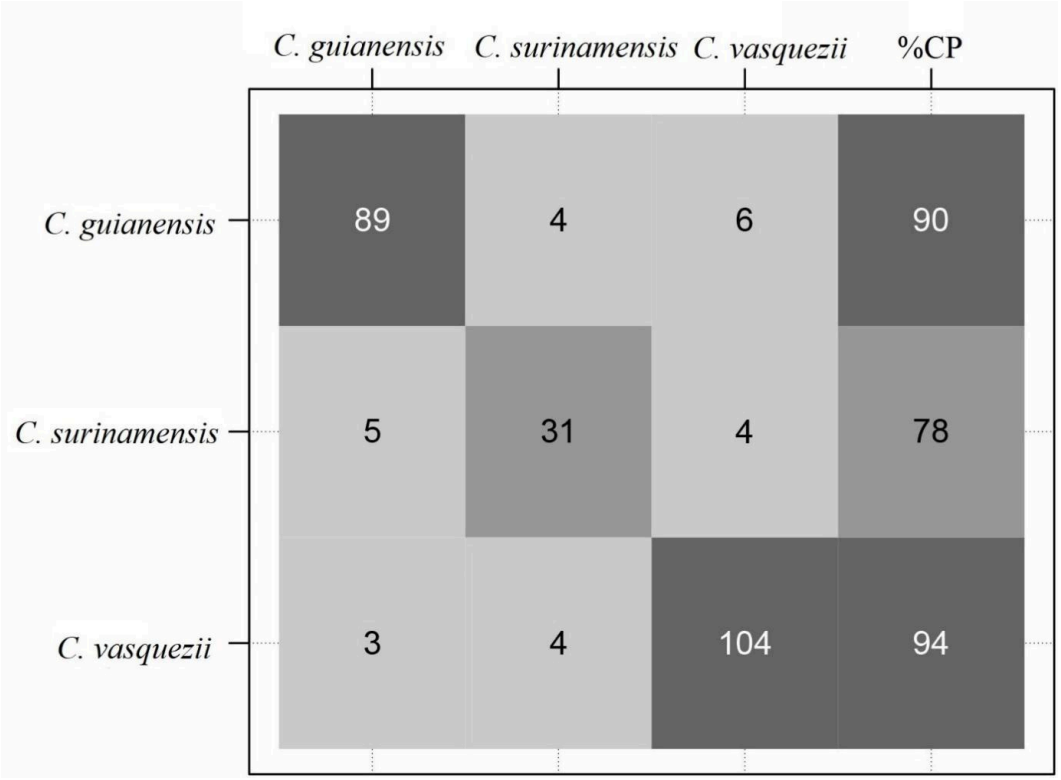


Fig. 8. Confusion matrix for Linear Discriminant Analysis with "Leave-One-Out" cross-validation (LDA–LOO). Sample prediction based on prior identification. The matrix cells show the number of predictions in each category, comparing the values predicted by the model to the actual values. The rows represent the true data categories. The columns represent the categories predicted by the model. %CP: percentage of correct predictions.



Tables:

Table 1. Number of *Carapa* spp. samples by species and collection, used in morphometric (MP) and NIR spectroscopy analyses.

Collection	***C. guianensis***		***C. surinamensis***		***C. vasquezii***	
	MP	NIR	MP	NIR	MP	NIR
Herbarium/INPA	20	21	7	11	15	32
PDBFF		2			29	44
PPBIO					25	25
This study	75	79	20	18	26	29
Total	**95**	**103**	**27**	**29**	**96**	**130**

Table 2. Analysis of variance (ANOVA) results for the 12 morphometric variables.

Morphometric variable	Degrees of freedom	Sum of squares	Mean square	F-value	Pr(>F)	
1. Petiole length (cm)	2	10	5.0	26.1	7.11e-11	***
2. Petiole width (cm)	2	4	2.2	14.3	1.42e-06	***
3. Rachis length (cm)	2	17749	8874	29.5	4.49e-12	***
4. Number of leaflet pairs	2	17	8.7	3.2	0.0403	*
5. Mean petiolule length (cm)	2	2	1.3	23.2	7.14e-10	***
6. Basal leaflet length (cm)	2	346	173.0	6.8	0.00132	**
7. Basal leaflet width (cm)	2	206	103.4	19.9	1.18e-08	***
8. Terminal leaflet length (cm)	2	957	478.5	9.9	7.36e-05	***
9. Terminal leaflet width (cm)	2	232	116.3	13.1	4.22e-06	***
10. Number of secondary veins	2	12	6.1	1.9	0.145	ns
11. Basal leaflet length/width ratio	2	0.1	0.07	7.4	0.000774	***
12. Terminal leaflet length/width ratio	2	0,1	0.05	8.0	0.000423	***

** para $P<0.001$; ** para $P<0.01$; * para $P<0.05$; ns para valores de $P\geq0.05$

Table 3. Percentage distribution of qualitative leaflet characters among the three species.

Leaflet character	*C. guianensis*	*C. surinamensis*	*C. vasquezii*
Oblong	6,3%	100%	97,9%
Elliptic	93,7%	NA	2,1%
Acuminate apex	98,9%	NA	NA
Obtuse apex	1,1%	100%	100%
Acute base	17,9%	NA	2,1%
Rounded base	6,3%	100%	96,9%
Cuneate base	75,8%	NA	1,0%

Table 4. Mean Balanced Accuracy by data source (MORPHO, NIR, COMBINED) and classification algorithm (LDA, PLS-DA, RF), balanced condition.

Data source	LDA	PLS-DA	RF (weighted)	RF (upsampled)
MORPHO	0.678 (0.664–0.693)	0.671 (0.657–0.685)	0.629 (0.619–0.640)	0.651 (0.640–0.664)
NIR	0.651 (0.640–0.661)	0.648 (0.638–0.657)	0.614 (0.603–0.624)	0.609 (0.598–0.620)
COMBINED	0.652 (0.641–0.662)	0.648 (0.638–0.658)	0.614 (0.605–0.624)	0.611 (0.600–0.623)

Table 5. Mean paired change in Balanced Accuracy (ΔBA) after correcting for class imbalance, by data source and algorithm, computed on the same cross-validation folds. The RF row shows the class-weighted variant.

Data source	Algorithm	ΔBA	95% CI	Significant?
MORPHO	LDA	+0.032	0.023 to 0.042	Yes
MORPHO	PLS-DA	+0.048	0.036 to 0.060	Yes
MORPHO	RF (weighted)	+0.004	−0.001 to 0.009	No
NIR	LDA	+0.024	0.013 to 0.034	Yes

Data source	Algorithm	ΔBA	95% CI	Significant?
NIR	PLS-DA	+0.026	0.018 to 0.035	Yes
NIR	RF (weighted)	−0.001	−0.006 to 0.004	No
COMBINED	LDA	+0.024	0.013 to 0.034	Yes
COMBINED	PLS-DA	+0.026	0.017 to 0.034	Yes
COMBINED	RF (weighted)	−0.002	−0.006 to 0.001	No

Table 6. Recall of *C. surinamensis* under original and class-balanced training conditions, by data source and algorithm.

Data source	Algorithm	Recall (original)	Recall (balanced)	ΔRecall
MORPHO	LDA	0.223	0.527	+0.304
MORPHO	PLS-DA	0.070	0.550	+0.480
MORPHO	RF (weighted)	0.103	0.135	+0.032
MORPHO	RF (upsampled)	0.103	0.331	+0.228
NIR	LDA	0.085	0.490	+0.405
NIR	PLS-DA	0.006	0.409	+0.403
NIR	RF (weighted)	0.051	0.070	+0.019
COMBINED	LDA	0.087	0.490	+0.403
COMBINED	PLS-DA	0.008	0.420	+0.412
COMBINED	RF (weighted)	0.053	0.070	+0.018

Table 7. Balanced Accuracy of LDA under an independent repeated-undersampling design (100 replicates), by data source.

Data source	Mean BA	95% CI
MORPHO	0.655	0.647–0.663
COMBINED	0.643	0.636–0.650
NIR	0.642	0.635–0.649

References

Aublet, J.B.C.F. 1775. *Histoire des Plantes de la Guiane Françoise*. Pierre-François Didot, London and Paris. 85–86p.

Barraclough, T.G.; Savolainen, V. 2001. Evolutionary rates and species diversity in flowering plants. *Evolution* 55: 677–683.

Bookstein, F.L. 1982. Foundations of Morphometrics. *Annual Review of Ecology, Evolution, and Systematics* 13: 451–470.

Cardoso PH, Menini Neto L, Somavilla NS, Trovó M. 2022. A morphometric approach and recircumscription of the *Stachytarpheta longispicata* complex (Verbenaceae). European Journal of Taxonomy 833: 12–45. doi: 10.5852/ejt.2022.833.1881.

Carvalho TCR, Vasconcelos CC, Terra-Araujo MH. 2024. Testing the morphotype-based classification of two leaf shapes in *Ecclinusa guianensis* (Sapotaceae) from Central Amazonia. Brittonia 76: 398–410. doi: 10.1007/s12228-024-09809-8.

Cavender-Bares, J.; Meireles, J.E.; Couture, J.J.; Kaproth, M.A.; Kingdon, C.C.; Singh, A.; *et al.,* 2016. Associations of leaf spectra with genetic and phylogenetic variation in oaks: Prospects for remote detection of biodiversity. *Remote Sensing* 8: 221.

Cavender-Bares J, White DM, Iwanycki Ahlstrand N, Austin MW, Bastianelli D, Bazan S, Boughalmi K, Cardinal-McTeague W, Chacon-Madrigal E, Couvreur TLP, Davis C, Durgante FM, Grace OM, Guzmán Q JA, Hansen K, Hernández-Leal MS, Hopkins MJG, Jackson R, Kothari S, Lee AK, Léveillé-Bourret E, Pinto-Ledezma J, Quinteros Casaverde NL, Meireles JE, Neto-Bradley B, Nichodemus CO, Ree RH, Schmull M, Soltis DE, Soltis PS, Tuomisto H, Ustin SL, Vasconcelos CC. 2026. Next-generation specimen digitization: capturing reflectance spectra from the world's herbaria for modeling plant biology across time, space, and taxa. New Phytologist 251: 645–665. doi: 10.1111/nph.70645.

Cope, J.S.; Corney, D.; Clark, J.Y.; Remagnino, P.; Wilkin, P. 2012. Plant species identification using digital morphometrics: A review. *Expert Systems with Applications* 39: 7562–7573.

Damasco, G.; Daly, D.C.; Vicentini, A.; Fine, P.V.A. 2019. Reestablishment of *Protium cordatum* (Burseraceae) based on integrative taxonomy. *Taxon* 68: 34–46.

De Candolle, C. 1878. Meliaceae. In: De Candolle, A.; De Candolle, C. (Eds.). *Monographiae Phanegamarum*. Masson, Paris. p.389–758.

Draper, F.C.; Baker, T.R.; Baraloto, C.; Chave, J.; Costa, F.; Martin, R.E.; *et al.,* 2020. Quantifying tropical plant diversity requires an integrated technological approach. *Trends in Ecology & Evolution* 35: 1100–1109.

Duminil, J.; Kenfack, D.; Viscosi, V.; Grumiau, L.; Hardy, O.J. 2012. Testing species delimitation in sympatric species complexes: The case of an African tropical tree, *Carapa* spp. (Meliaceae). *Molecular Phylogenetics and Evolution* 62: 275–285.

Durgante, F.M.; Higuchi, N.; Almeida, A.; Vicentini, A. 2013. Species spectral signature: Discriminating closely related plant species in the Amazon with Near-Infrared Leaf-Spectroscopy. *Forest Ecology and Management* 291: 240–248.

Durgante FM, Vasconcelos CC, Hadlich HL, Mallmann CL, Perdiz RO, Cunha KCM, Chaves SR, Pasquini C, Torralvo K, Hopkins MJG, Schöngart J, Piedade MTF, Sá PA, Conde MLG, Wittmann F. 2026. Species recognition in herbaria and standing trees using an accessible portable spectrometer. Preprint.

Enriquez, G.V.E.; Silva, M.A.; Cabral, E. 2003. *Biodiversidade da Amazônia: usos e potencialidades dos mais importantes produtos naturais do Pará*. Embrapa, Belém. 179p.

Fišer, C.; Robinson, C.T.; Malard, F. 2018. Cryptic species as a window into the paradigm shift of the species concept. *Molecular Ecology* 27: 613–635.

Gaem PH, Andrade A, Mazine FF, Vicentini A. 2022. Tree species delimitation in tropical forest inventories: Perspectives from a taxonomically challenging case study. Forest Ecology and Management 505: 119900. doi: 10.1016/j.foreco.2021.119900.

Gama, I.G. da 2019. Distinção de sementes, frutos e plântulas das três espécies do gênero *Carapa* Aubl. na Amazônia Central., Dissertação de mestrado, Instituto Nacional de Pesquisas da Amazônia (INPA), Brasil. 73p.

Gomes, A.C.S.; Andrade, A.C.S.; Barreto-Silva, J.S.; Brenes-Arguedas, T.; López, D.C.; Freitas, C.C.; *et al.,* 2013. Local plant species delimitation in a highly diverse Amazonian forest: Do we all see the same species? *Journal of Vegetation Science* 24: 70–79.

Harms, H. 1940. Meliaceae. In: Engler, A.; Prantl, K. (Eds.). *Die natürlichen Pflanzenfamilien*. W. Engelmann, Leipzig. p.1–172.

Holanda, A.S.S.; Vicentini, A.; Chave, J.; Carvalho, F.A.; Zartman, C.E. 2021. Quantitative morphometrics suggest that the widespread Neotropical *Humiria balsamifera* (Aubl.) St. Hil. is a species complex. *Acta Botanica Brasilica* 35: 339–351.

Hopkins MJG. 2007. Modelling the known and unknown plant biodiversity of the Amazon Basin. Journal of Biogeography 34: 1400–1411. doi: 10.1111/j.1365-2699.2007.01737.x.

Jombart, T.; Kamvar, Z.N.; Collins, C.; Lustrik, R.; Beugin, M.-P.; Knaus, B.J.; *et al.,* 2023. adegenet: Exploratory Analysis of Genetic and Genomic Data. R package version 2.1.10. (https://rdrr.io/cran/adegenet/). Accessed on 20 Jan 2024.

Kenfack, D. 2011a. Resurrection in *Carapa* (Meliaceae): a reassessment of morphological variation and species boundaries using multivariate methods in a phylogenetic context. *Botanical Journal of the Linnean Society* 165: 186–221.

Kenfack, D. 2011b. A synoptic revision of *Carapa* (Meliaceae). *Harvard Papers in Botany* 16: 171–231.

Kenfack, D. 2011c. *Carapa vasquezii* (Meliaceae), a new species from western Amazonia. *Brittonia* 63: 7–10.

Koenen, E.J.M.; Clarkson, J.J.; Pennington, T.D.; Chatrou, L.W. 2015. Recently evolved diversity and convergent radiations of rainforest mahoganies (Meliaceae) shed new light on the origins of rainforest hyperdiversity. *New Phytologist* 207: 327–339.

Kuhn, M.; Wing, J.; Weston, S.; Williams, A.; Keefer, C.; Engelhardt, A.; *et al.,* 2023. caret: Classification and Regression Training. R package version 6.0-94. (https://rdrr.io/cran/caret/). Accessed on 20 Jan 2024.

Lang, C.; Almeida, D.R.A.; Costa, F.R.C. 2017. Discrimination of taxonomic identity at species, genus and family levels using Fourier Transformed Near-Infrared Spectroscopy (FT-NIR). *Forest Ecology and Management* 406: 219–227.

Lang, C.; Costa, F.R.C.; Camargo, J.L.C.; Durgante, F.M.; Vicentini, A. 2015. Near infrared spectroscopy facilitates rapid identification of both young and mature Amazonian tree species. *PLoS One* 10: e0134521.

Mendonça, A.P.; Ferraz, I.D.K. 2007. Óleo de andiroba: processo tradicional da extração, uso e aspectos sociais no estado do Amazonas, Brasil. *Acta Amazonica* 37: 353–364.

Oksanen, J.; Simpson, G.L.; Blanchet, F.G.; Kindt, R.; Legendre, P.; Minchin, P.R.; *et al.,* 2024. vegan: Community Ecology Package. R package version 2.6-8. (https://rdrr.io/cran/vegan/). Accessed on 20 Jan 2024.

Padial, J.M.; Miralles, A.; De la Riva, I.; Vences, M. 2010. The integrative future of taxonomy. *Frontiers in Zoology* 7: 16.

Pante, E.; Schoelinck, C.; Puillandre, N. 2015. From integrative taxonomy to species description: one step beyond. *Systematic Biology* 64: 152–160.

Pasquini C. 2018. Near infrared spectroscopy: A mature analytical technique with new perspectives – A review. Analytica Chimica Acta 1026: 8–36. doi: 10.1016/j.aca.2018.04.004.

Perdiz, R.O. 2023. NIRtools: Tools to deal with near infrared (NIR) spectroscopy data. R package version 0.0.1. (https://rdrr.io/github/ricoperdiz/NIRtools/). Accessed on 20 Jan 2024.

Prata, E.M.B.; Sass, C.; Rodrigues, D.P.; Domingos, F.M.C.B.; Specht, C.D.; Damasco, G.; *et al.,* 2018. Towards integrative taxonomy in Neotropical botany: Disentangling the *Pagamea guianensis* species complex (Rubiaceae). *Botanical Journal of the Linnean Society* 188: 213–231.

Pritchard, J.K.; Stephens, M.; Donnelly, P. 2000. Inference of population structure using multilocus genotype data. *Genetics* 155: 945–959.

R Core Team 2023. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. (https://www.r-project.org/). Accessed on 15 Apr 2023.

Ripley, B.; Venables, B.; Bates, D.M.; Hornik, K.; Gebhardt, A.; Firth, D. 2024. MASS: Support Functions and Datasets for Venables and Ripley's MASS. R package version 7.3-61. (https://rdrr.io/cran/MASS/). Accessed on 20 Jan 2024.

Rohwer, J.G. 2000. Toward a phylogenetic classification of the Lauraceae: evidence from matK sequences. *Systematic Botany* 25: 60–71.

Rohwer, J.G.; Richter, H.G.; van der Werff, H. 1991. Two new genera of Neotropical Lauraceae and critical remarks on the generic delimitation. *Annals of the Missouri Botanical Garden* 78: 388–400.

Scotti-Saintagne, C.; Dick, C.W.; Caron, H.; Vendramin, G.G.; Guichoux, E.; Buonamici, A.; *et al.,* 2013. Phylogeography of a species complex of lowland Neotropical rain forest trees (*Carapa*, Meliaceae). *Journal of Biogeography* 40: 676–692.

Staner, P. 1941. Les Méliacées du Congo Belge. *Bulletin du Jardin botanique de l'État à Bruxelles* 16: 109–251.

Styles, B.T. 1981. Swietenioideae. In: Pennington, T.D.; Styles, B.T.; Taylor, D.A.H. (Eds.). *Flora Neotropica Monograph No 28: Meliaceae*. The New York Botanical Garden Press, New York. p.359–420.

Thenkabail, P.; Lyon, J. 2016. *Hyperspectral Remote Sensing of Vegetation*. CRC Press, Boca Raton (FL).

Volpato, E.; Schmidt, P.B.; Araújo, V.C. de 1972. *Carapa guianensis* Aubl. (Andiroba). Estudos comparativos de tratamentos silviculturais. *Acta Amazonica* 2: 75–81.

Wang, R.; Gamon, J.A. 2019. Remote sensing of terrestrial plant biodiversity. *Remote Sensing of Environment* 231: 111218.

Wickham, H. 2023. plyr: Tools for Splitting, Applying and Combining Data. R package version 1.8.9. (https://rdrr.io/cran/plyr/). Accessed on 20 Jan 2024.

Wright, M.N.; Ziegler, A. 2017. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software* 77: 1–17.

Yared, J.A.G.; Carpanezzi, A.A. 1981. *Conversão de capoeira alta da Amazônia em povoamento de produção madeireira: o método do 'recru' e espécies promissoras*. Embrapa, Belém. 27p.

Vasconcelos CC. 2024. Taxonomia integrativa para a delimitação e descrição de espécies em Chrysophylloideae neotropical (Sapotaceae). PhD Thesis, Instituto Nacional de Pesquisas da Amazônia, Manaus.

Vasconcelos CC, Ferraz IDK, Durgante FM, Wittmann F, Schöngart J, Piedade MTF, Camargo JLC, Terra-Araujo MH. 2025. *Ecclinusa nervosa* (Sapotaceae, Chrysophylloideae), a new species discovered in Central Amazonia. Systematic Botany 50: 55–66.