Chemical differentiation of Bolivian Cedrela species as a tool to trace illegal timber trade

Paredes-Villanueva, K., Espinoza, E., Ottenburghs, J., Sterken, M. G., Bongers, F., & Zuidema, P. A.

Chemical differentiation of Bolivian *Cedrela* species as a tool to trace illegal timber trade

Kathelyn Paredes-Villanueva[a,b,*], Edgard Espinoza[c], Jente Ottenburghs[d], Mark G. Sterken[e], Frans Bongers[a] and Pieter A. Zuidema[a]

[a] *Forest Ecology and Management Group, Wageningen University & Research, Droevendaalsesteeg 3, PO Box 47, 6700AA Wageningen, The Netherlands*

[b] *Carrera de Ingeniería Forestal, Universidad Autónoma Gabriel René Moreno, Facultad de Ciencias Agrícolas, Km 9 al Norte, El Vallecito, Santa Cruz, Bolivia*

[c] *U.S. Fish & Wildlife Service, National Forensics Laboratory, 1490 East Main Street, Ashland, OR 97520-1310 U.S.A.*

[d] *Resource Ecology Group, Wageningen University & Research, Droevendaalsesteeg 3a, 6708 PB, Wageningen, the Netherlands*

[e] *Laboratory of Nematology, Wageningen University & Research, Droevendaalsesteeg 2, 6708 PB Wageningen, The Netherlands*

*Corresponding author:

kathypavi@gmail.com

Abstract

Combating illegal timber trade requires the ability to identify species and verify geographic origin of timber. Forensic techniques that independently verify the declared species and geographic origin are needed, as current legality procedures are based on certificates and documents that can be falsified. Timber from the genus *Cedrela* is among the most economically valued tropical timbers worldwide. Three *Cedrela* species are included in the Appendix III of CITES: *C. fissilis*, *C. odorata*, and *C. angustifolia* (listed as *C. lilloi*). *Cedrela* timber is currently traded with false origin declarations and under a different species name, but tools to verify this are lacking. We used Direct Analysis in Real Time Time-of-Flight Mass Spectrometry (DART-TOFMS) to chemically identify *Cedrela* species and sites of origin. Heartwood samples from six *Cedrela* species (the three CITES-listed species plus *C. balansae*, *C. montana*, and *C. saltensis*) were collected at 11 sites throughout Bolivia. Mass spectra detected by DART-TOFMS comprised 1062 compounds; their relative intensities were analysed using Principal Component Analyses (PCA), Kernel Discriminant Analysis (KDA), and Random Forest analyses to check discrimination potential among species and sites. Species were identified with a mean discrimination error of 15-19%, with substantial variation in discrimination accuracy among species. The lowest error was observed in *C. fissilis* (Mean=4.4%). Site discrimination error was considerably higher: 43-54% for *C. fissilis* and 42-48% for *C. odorata*. These results provide good prospects to differentiate *C. fissilis* from other species, but at present there is no scope to do so for other tested species. Thus, discrimination is highly species specific. Our findings for tests of geographic origin suggest no potential to discriminate at the studied scale and for the studied species. Cross-checking results from different methods (KDA and Random Forest) reduced discrimination errors. In all, the DART-TOFMS technique allows independent verification of claimed identity of certain *Cedrela* species in timber trade.

Keywords: Illegal logging, *Cedrela*, mass spectrometry, discriminant analysis, Random Forest

Introduction

Illegal trade in timber is a worldwide environmental problem, resulting in damage of natural resources and economic loss. It has been estimated that 10% to 80% of the total timber trade is illegal (Seneca Creek Associates, 2004) and in some countries, such as Papua New Guinea, Liberia, and the Amazon countries (Stark and Pang Cheung, 2006; Lawson and MacFaul, 2010; Wit, *et al*., 2010), this percentage has been as high as 80-90% of all logging operations. The most common type of fraud concerns false declarations of species and geographic origin, as current legal procedures are generally based on certificates and documents which can be falsified. Most legislative measures focus at combating international illegal trade but a high proportion (70-90%) of illegal tropical timber is traded in domestic markets (Cerutti and Lescuyer, 2011; Kishor and Lescuyer, 2012; Lescuyer *et al*., 2014). Clearly, there is a need for forensic techniques to independently verify the origin of traded timber in both domestic and international markets.

The genus *Cedrela* (Meliaceae) delivers one of the most important tropical timbers (tropical cedar), but illegal logging of *Cedrela* has resulted in CITES-listing of several species in this genus (Compt and Christy, 2008). As a result, timber from these species can be traded internationally only if the appropriate permits have been obtained and presented for clearance at the port of entry or exit (CITES, 2017). The problem is that CITES-listed and non-listed *Cedrela* species are harvested and traded under the same name (Moya *et al*., 2013) and are often confused due to wood-anatomical similarities (Gasson, 2011; Gasson *et al*., 2011; Moya *et al*., 2013). For authorities enforcing CITES, methods to differentiate *Cedrela* species are needed.

Bolivia harbours as many as six *Cedrela* species, in different climatic zones, from moist to dry tropical forests, and from low to high altitudes (Mostacedo *et al*., 2003; Navarro, 2011; Navarro-Cerrillo *et al*., 2013): *Cedrela angustifolia* Sessé & Moc. Ex DC., *Cedrela balansae* C. DC., *Cedrela fissilis* Vell., *Cedrela montana* Moritz ex Turcz., *Cedrela odorata* L. and *Cedrela saltensis* M.A. Zapater & del Castillo. *Cedrela* species are highly valued locally (Mostacedo and Fredericksen, 1999) and used in carpentry, fine furniture, doors, windows, joinery, musical instruments, carvings, coatings and plywood (Toledo *et al*., 2008). However, *Cedrela* populations have declined considerably in recent years due to overexploitation (Mostacedo and Fredericksen, 1999, 2001). As a result, out of the six species, three are currently listed in Appendix III of CITES: *C. odorata*, *C. fissilis* and *C. angustifolia* (listed as *C. lilloi* C. DC.) (CITES, 2017). Despite legal harvesting limitations, these species remain at high risk because of continued illegal logging and timber trade (ABT, 2017). The high incidence of illegal trade indicates that control systems have limited effectiveness and methods for independent verification of species and legal origin are needed.

Chemical analysis tools, such as mass spectrometry (Fidelis *et al*., 2012), near-infrared spectroscopy (Braga *et al*., 2011; Bergo *et al*., 2016), and stable isotopes (Kagawa and Leavitt, 2010; Förstel *et al*., 2011; Vlam *et al*., 2018), can be used to discriminate species and verify the geographical origin of traded timber. For example, previous studies used a specific mass spectrometer to discriminate species that cannot be identified based on wood anatomy in the Americas (Espinoza *et al*., 2015), Africa (Deklerck *et al*., 2017), and Asia (McClure *et al*., 2015). In this study, we focus on chemical characterization by Direct Analysis in Real Time (DART) coupled with Time-of-Flight Mass Spectrometry (TOFMS). This technique has the potential to assist in enforcing protection of *Cedrela* species as it cannot be falsified, in contrast to current certificates used for declaration of species origin. In DART analysis, the mass spectrometer quickly identifies the chemical components by the differing mass to charge (*m/z*) of ions/compounds from specimens, without the need for sample preparation. The resulting chemical spectra can be used as a reference database for species identifications. Because this methodology has a high potential to identify species and locations, our aim is to test its applicability to differentiate *Cedrela* timber obtained from different species and geographic provenances.

We answer the following research questions: (1) To what extent can Bolivian *Cedrela* species be differentiated based on wood chemical composition? (2) To what extent can chemical composition help to differentiate timber sourced from different sites in Bolivia? (3) What is the accuracy for identification of each *Cedrela* species and site of origin within Bolivia based on their chemical profiles? As the geographical sites of the collected samples may have different environmental conditions, we expect to find distribution patterns of the wood composition that mirror these conditions (Zobel and van Buijtenen, 1989; Wilkins and Stamp, 1990; Mosedale and Ford, 1996; Moya and Calvo-Alvarado, 2012). We also expect that each *Cedrela* species will present specific chemicals that distinguish it from others (Chatterjee *et al*., 1971; Cordeiro *et al*., 2012; Eason and Setzer, 2007; Lago *et al*., 2004; Maia *et al*., 2000).

Methods

Study site and species

We studied heartwood samples from 6 *Cedrela* species in Bolivia, from 11 sites. In total we sampled 127 trees. Altitude of the sites ranged from 145 m.a.s.l. (meters above sea level) in Riberalta to 2022 m.a.s.l. in Postrervalle (Table 1). We selected sites taking into account the distribution of the study species and we maintained a minimum of 70 km distance between all site pairs to maximize the sampling coverage across the country (Table 1 and Figure 1). The maximum distance between pairs of sampled sites was 1300 km (Cobija-Roboré). We used these samples to perform two types of tests: differentiation of species and differentiation of geographic origin. In the species identification analyses, we included all *Cedrela* species in the sample collection to analyse cross-species discrimination. For

the geographic origin analysis, we only included the two species with the largest sample sizes that we had sampled at multiple sites: *C. odorata* from 3 sites and *C. fissilis* from 6 sites. The maximum distance between pairs of sites was 80 km for *C. fissilis* (Espejos-Yapacaní) and 425 km for *C. odorata* (Ribertalta-Rurrenabaque). Minimum distances between pairs of sites were 70 km (Concepción-Guarayos) for *C. fissilis* and 285 km (Ribertalta-Cobija) for *C. odorata*. We performed a stratified random sampling: in each of the *Cedrela* populations found, trees of diameter ≥10 cm were randomly selected with a minimum distance among trees of at least 50 m in order to obtain a homogeneous sampling in each site and to reduce genetic noise and confounding impact of sampling relatives on site (Gillies *et al*., 1999). This random selection of samples covered different types of forest strata.

Table 1. *Cedrela* species and sites included in the study. Sample size refers to the number of trees sampled; botanical samples to the number of trees from which botanical samples were obtained for verification of identification by taxonomists.

Species	Sites	Sample size	Botanical samples	Altitude (m.a.s.l.)
C. angustifolia	Monteagudo	2	2	1705
	Postrervalle	13	12	2022
C. balansae	Concepción	10	10	432
C. fissilis	Bajo Paraguá	10	*	287
	Concepción	13	13	432
	Espejos	6	6	553
	Guarayos	13	9	260
	Roboré	10	*	632
	Yapacaní	10	5	318
C. montana	Postrervalle	2	2	2022
C. odorata	Cobija	10	*	274
	Riberalta	10	*	145
	Rurrenabaque	10	4	309
C. saltensis	Monteagudo	8	2	1705
Total		**127**	**65**	

*No botanical samples were collected, but identification was based on previous collections.

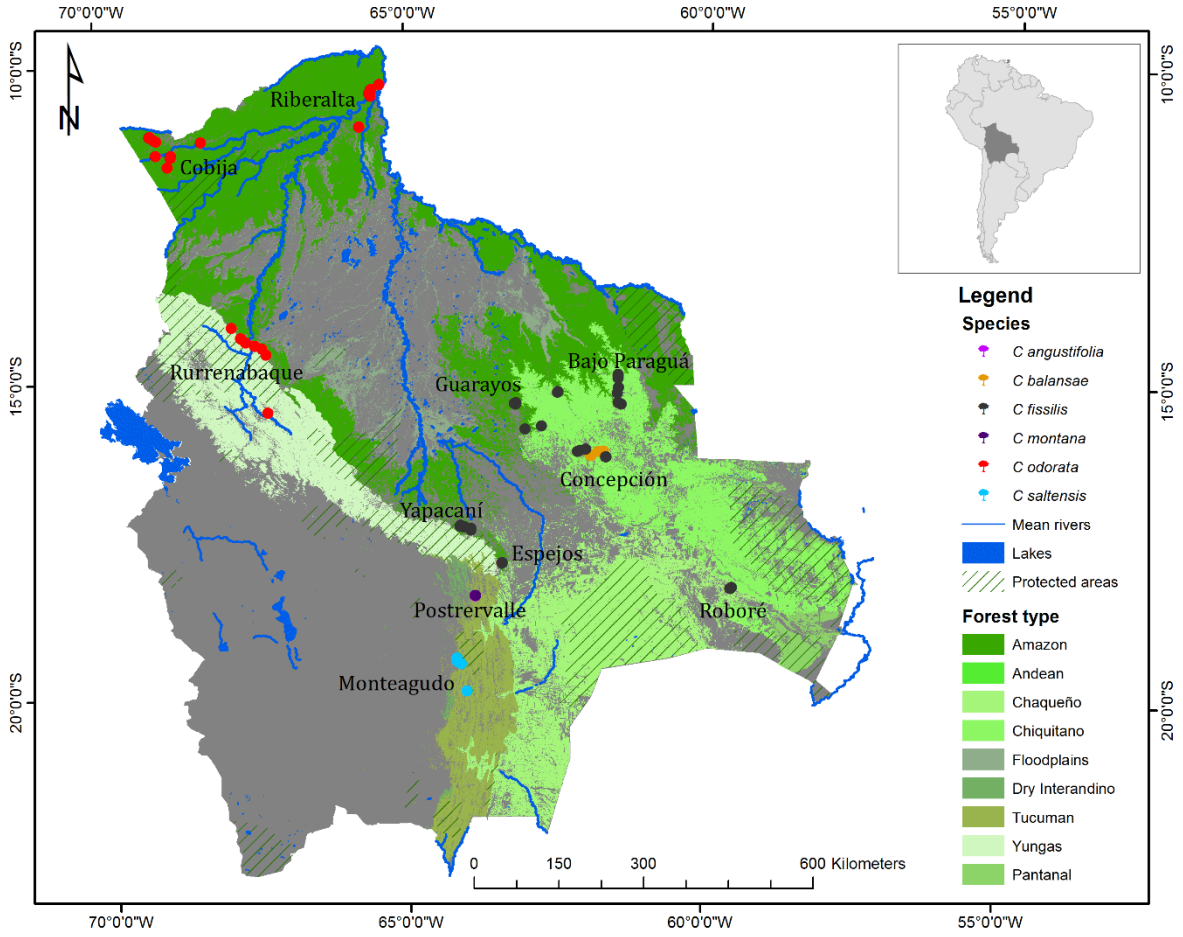


Figure 1 Locations of sampled trees belonging to six *Cedrela* species in Bolivia. Forest cover: Autoridad de Bosques y Tierra, 2015.

Preliminary analyses of sapwood and heartwood showed a wider variation of compounds in heartwood (70.0629-1086.567 *m/z*) compared with sapwood with a dominance of sugars and starch (69.0285-958.4909 *m/z*) that were not species-specific. Based on these results, we decided to only include heartwood samples in our analyses. A single heartwood sample was collected from each tree using a 5 mm diameter increment borer (Haglöf) at 50-100 cm stem height. Species were morphologically identified *in situ* with the help of local guides. In addition, botanical samples were collected for species confirmation when identification in the field was not possible. This was done for 53% of the sampled trees. The voucher preparation and confirmation of the species based on herbarium collections were carried out by an experienced botanist, A. Araujo Murakami at the Museo de Historia Natural Noel Kempff Mercado (Bolivia).

Chemical analysis

We used Direct Analysis in Real Time Time-of-Flight- Mass Spectrometry (DART-TOFMS) to differentiate *Cedrela* species and to explore if geographical origin could be determined based on

chemical composition of heartwood. The DART source consists of an ionization technique that occurs at atmospheric pressure and is discussed by Cody *et al*. (2005). Once the molecules from the sample are ionized, they are directed towards the time-of-flight mass spectrometer (TOFMS) (Cody *et al*., 2005). The mass spectrometer will then characterize the molecules from the sample by determining the mass to charge (*m/z*) of the ions in their protonated forms.

The principal ionization mechanisms for DART-TOFMS have been thoroughly discussed and it has been used to identify timber species with an accuracy of 70% to 95% (Lancaster and Espinoza, 2012; Evans *et al*., 2017). To describe the chemotaxonomic relationship of our *Cedrela* samples, mass spectra were acquired using a DART ion source (IonSense, Saugus, MA, USA) coupled to a JEOL AccuTOF time-of-flight mass spectrometer (JEOL USA, Peabody, MA, USA) in positive ion mode. To check if preparation of wood was needed, we tested the maximum number of compounds by soaking wood in methanol versus using wood with no previous treatment. We did not observe any enhancement with previous preparation of wood samples (data not shown). Hence, we decided to use untreated heartwood samples. We cut slivers of heartwood no wider than 4 mm from each sample with a scalpel. These slivers were held in the DART helium gas stream for 8 seconds. A mass calibration standard of polyethylene glycol 600 (Ultra, Kingstown, RI, USA) was run between each 5 samples. The DART source parameters were: needle voltage, 3.5 kV; electrode 1 voltage, 150 V; electrode 2 voltage, 250 V; and gas heater temperature, 350°C. The mass spectrometer settings included: rings lens voltage, 5 V; orifice 1 voltage, 20 V; orifice 2, 5 V; cone temperature, 120°C; peaks voltage, 600 V; ion guide bias, 28 V; focus lens voltage, -120 V; reflectron voltage, 870 V; pusher voltage, 778 V; pulling voltage, -778 V; suppression voltage, 0.00 V; flight tube voltage, -7000 V; and detector voltage, 2000 V. Spectra covered the mass range of 70 to 1100 mass-to-charge ratios (*m/z*) and were obtained at 1 scan per second. The helium flow rate for the DART source was 2.0 mL s^{-1}. The resolving power of the mass spectrometer, as stated by the manufacturer, was ±2.0 millimass units (mmu). The diagnostic compounds for spectrum classification were selected with 250 mmu and 1% threshold (Deklerck *et al*., 2017). TSS Unity, a mass-spec data-processing software (Shrader Software Solutions, Inc., Grosse Pointe Park, MI, USA), was used to export the data as text files for further analysis.

Statistical analysis

Our analysis of the masses (*m/z*) detected and relative intensities obtained from the DART TOFMS consisted of several steps. To evaluate if chemotypes can enable differentiation of *Cedrela* species (research question 1: species identification) we first evaluated the existence of species specific compounds, reduced the sample-compound data matrix using Principal Component Analysis (PCA) and finally performed a discriminant analysis to classify the species, determine the importance of each compound and predict sample assignment. For these analyses we used Kernel Discriminant Analyses with package ks 1.10.5 (Duong, 2007, 2017), and Random Forest model with package randomForest

4.6.12 (Liaw and Wiener, 2002) and dplyr 0.7.4 (Wickham *et al*., 2017) in R version 3.3.3 (R Development Core Team, 2017). We used PCA in order to reduce the number of variables (compounds) into principal components that can then be used as input for a first type of discriminant analysis (KDA). A second discriminant analysis (Random Forest) was used to identify the most important compounds that can differentiate between species. Both discriminant analyses were based on randomized samples and variables in every run. The classification results allowed us to assess the classification success by evaluating frequencies of correct and erroneous identifications. For the analysis of geographic origin (research question 2: geographic origin identification) for *C. fissilis* and *C. odorata* (Table 1), we followed the same steps. Based on the classifications, cross validation errors were estimated for species and site assignments (research question 3: identification accuracy).

In detail the method involved four main steps. First, we produced a heat-map graph to visualize the chemical profiles (or chemotypes) of the specimens and to verify whether heartwood samples of a particular species contain diagnostic molecules (expressed as mass-to-charge ratio: *m/z*) that allow it to be distinguished from other species. The heatmap is a graphical representation of the raw mass spectra measured by DART-TOFMS and is created using the Mass Mountaineer software (RBC Software, Peabody, MA, USA). It illustrates the mass-to-charge ratio (*m/z*) of the detected compounds and their intensities in a spectrum.

Second, to reduce the large data matrix into a set of variables so that the variation within each set is maximized (Gotelli and Ellison, 2004), a PCA was necessary for the set which consisted of 125 samples and 1062 compounds. PCA aims to find the linear combinations of variables by using the covariance matrix of data. The first axis reflects the linear fit capturing most of the variation and the successive orthogonal axes reflect the linear capturing of remaining variation not captured in each of the previous components. We extracted six principal components from the sample-molecule matrix, reflecting the greatest variation in the data matrix. The loadings of all 125 samples on the first six axes were retained and this new matrix was used as input in the discriminant analysis. We excluded *C. montana* due to its small sample size (2 individuals).

Third, we performed Kernel Discriminant Analysis (KDA) to test species identification and geographic origin. As KDA cannot cope with more than 6 variables, we performed the PCA analysis described above, and used the first 6 PCA axes. KDA separates the samples based on an *a priori* classification assignment (to species and sites classes) and looks for the optimal non-linear combination of variables (here the 6 component loadings) for maximal separation of the samples in the six dimensional space (Baudat and Anouar, 2000). KDA's learning algorithm uses Bayes discriminant rule which allocates a point x in the sample space to one (and only one) of the sampled populations. Each population is associated to a kernel density which was estimated implementing a diagonal data-driven (constrained, symmetric and positive-definite) bandwidth matrix (Duong, 2007). This learning algorithm needs to be

trained in order to assess the discrimination power of KDA. Therefore, our data were split in two sets: 80% for training and 20% for testing the model. The pre-smoothed data were then applied to estimate a Smoothed Cross Validation (SCV) error (Duong, 2007) as a different procedure to test correctness of the assignment tests. This delivers the classification error (%) which is the probability that samples are incorrectly assigned to a provenance. A cross validation error of 0% indicates that all the samples were correctly assigned.

Finally, we used Random Forest analysis to generate a sample classification model in which splits are based on just one chemical compound. One Random Forest run created 500 'Random Forests' which are used to obtain a final model (Breiman, 2001; Liaw and Wiener, 2002). As with KDA, the algorithm uses 80% of the dataset for training and 20% for model validation. Every run of Random Forest uses a different training set and may lead to different results. Therefore, we ran Random Forest 100 times and averaged the results. In this way, a total of 50.000 Random Forests (100 runs x 500 Random Forests) were built. For each run, the model provided a list of compounds, with their value of importance. We selected the most important compounds that occurred in >40% of the runs and calculated their frequency. These tentative assignments were based on 351 molecules described either for the *Cedrela* genera or the Meliaceae family (Afendi *et al*., 2012). Chemical composition in wood can vary not only among species but also for a given tree species or even a given tree (Pettersen, 1984), but heartwood extractive and exudates can also be species specific (Hillis, 1987). Therefore we used the list of the most important compounds to check if any species indicative compound was present.

Random Forest analysis allowed us to identify specific chemical compounds that separate one species or site from the other. The Out-of-Bag (OOB, take one out) error rate and species class error were estimated for each of the 100 runs and used to calculate the standard deviation (SD) of these estimates. The OOB estimate is equivalent to the SCV error of the KDA analysis.

Both KDA and Random Forest analyses generated confusion matrices showing the frequency at which each species/site was wrongly classified. In addition, the total of samples tested for each species after 100 randomization runs allowed us to check with what species a single sample could be confused. Finally, the mean errors per species for site identification across the 100 runs were obtained together with their corresponding standard deviation.

Results

A total of 1062 ions were characterized and their respective intensities were described, in 6 *Cedrela* species from 11 sites across Bolivia, from the DART-TOMFS spectra. The results were analysed for species and sites identification separately. A first inspection of chemical data in the heatmap (Figure 2) suggests species-specific patterns in the chemical profiles. Further cross-checking with the actual mass

spectra confirmed that the *C. odorata* samples had a higher intensity of compounds with molecular masses around *m/z* 212 and 480, *C. fissilis* had higher intensities for compounds in the *m/z* 484-502 range, *C. balansae* at *m/z* 478 and 680, and *C. angustifolia* showed high intensities for compounds at *m/z* 212 and 400. Although the samples of *C. montana* showed distinctive ions at *m/z* 275 and 398, this species was excluded from further analyses due to small sample size.

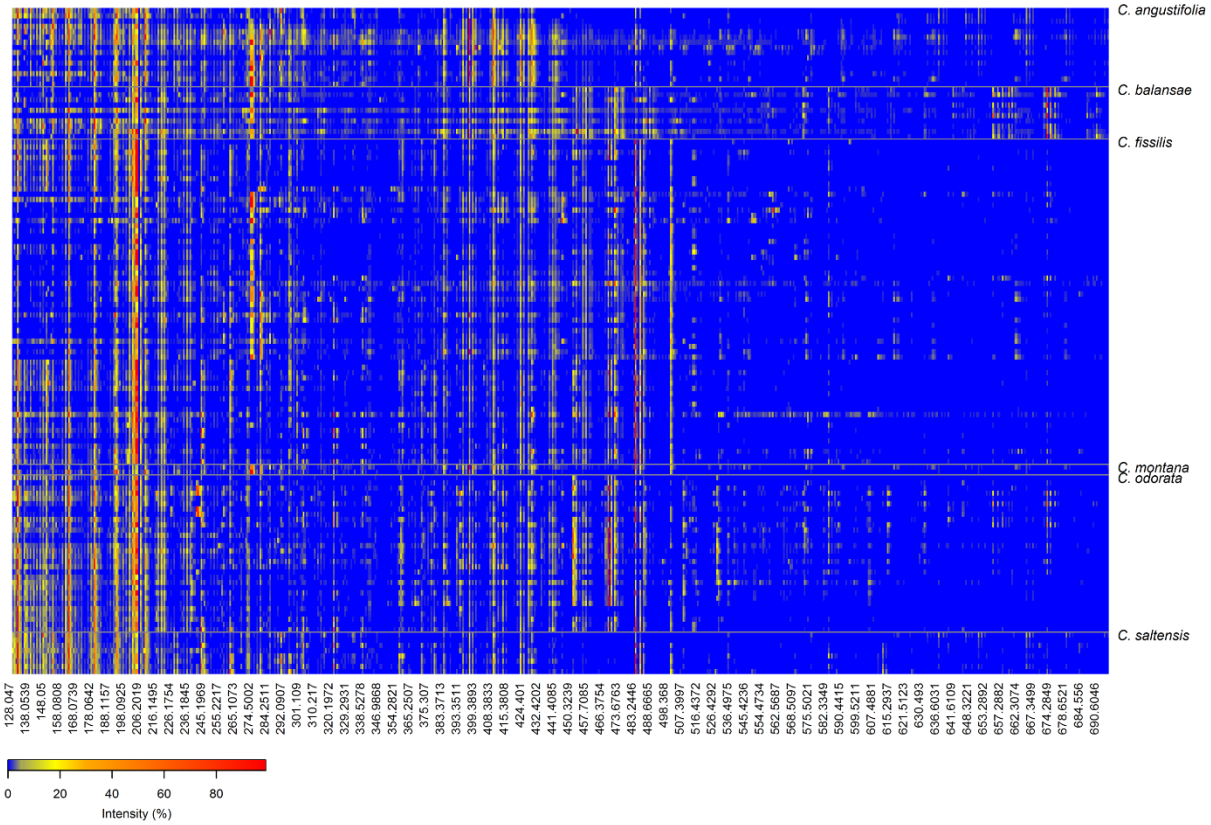


Figure 2 Heatmap of the output of the DART-TOMFS for 6 *Cedrela* species in Bolivia. Each row represents one sample (one tree). Each column represents a specific mass-to-charge ratio (*m/z*) of an ion. Colour gradient represents relative compound intensity (relative to the most abundant compound).

Identification of species

The analysis for species differentiation included five species*: C. angustifolia*, *C. balansae*, *C. fissilis*, *C. odorata*, and *C. saltensis* (Table 1). The PCA analysis showed that the six most important components together explained 72.1% of the variation across the samples and that the samples were reasonably well separated in the PCA space (Figure 3). The variances explained by the 6 principal components (PCs) were: 24%, 20%, 10%, 8%, 6%, and 4% for PC 1-6 correspondingly. These six components were used as input for the KDA. The KDA (of the 80% sample) resulted in a clear separation of the species (Table 2).

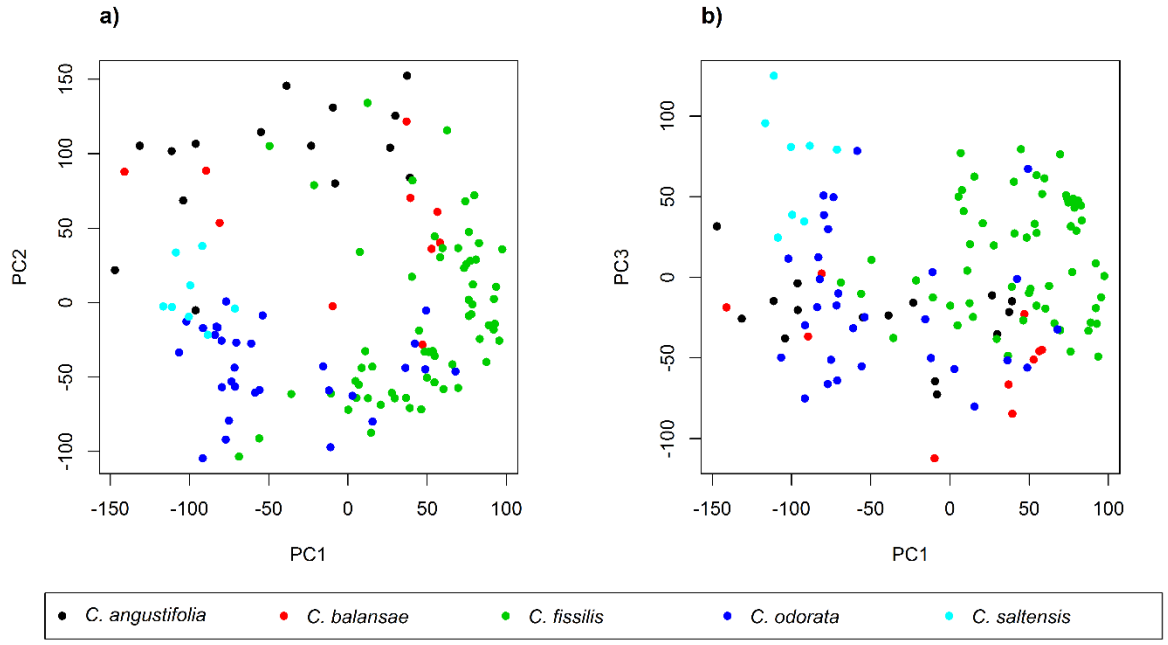


Figure 3 Results of Principal Component Analysis used for KDA analyses for species. Scatterplots combining (a) PC1 and PC2, and (b) PC1 and PC3.

Table 2. Error classification for species. Mean Smoothed Cross Validation (SCV) error, mean Out-of-bag (OOB) error for classification and their corresponding standard deviations (SD) were estimated after 100 runs for KDA and Random Forest, respectively.

Species	KDA		Random Forest	
	Mean error (%)	SD (%)	Mean error (%)	SD (%)
C. angustifolia	26.5	28.0	33.9	7.9
C. balansae	46.1	36.8	42.4	17.7
C. fissilis	8.7	7.1	4.4	1.8
C. odorata	22.3	17.5	15.8	5.3
C. saltensis	20.2	32.4	29.6	17.2
Mean	**18.9**	**7.0**	**14.9**	**2.3**

The KDA had a total mean error of 19% for the SCV test (Table 2). Species-specific errors differed strongly, from 8.7% for *C. fissilis* to 46.1% for *C. balansae*. The mean error per species (OOB) from the 100 Random Forest analyses was 15%, representing a mean identification accuracy of 85% (Table 2; Supplementary Data Figure A.1). Again, these errors differed substantially between species with the lowest value of 4.4% for *C. fissilis* and highest error of 42.4% for *C. balansae* (Supplementary Data Figure A.1).

In the KDA analysis, identification errors for *C. angustifolia* and *C. balansae* included wrong assignments to all the other species. *C. fissilis* was wrongly identified as all the species except as *C.*

279 *saltensis*. *C. odorata* was mostly identified as *C. fissilis* (118 samples out of 562) and in some cases
280 wrongly identified as *C. saltensis* (11 samples out of 562). It was rarely classified as *C. angustifolia* (1
281 sample out of 562) and never as *C. balansae*. *C. saltensis* was mostly confused with *C. odorata* (33
282 samples out of 188), in some cases with *C. angustifolia* (10 samples out of 188), rarely as *C. balansae*
283 (2 samples out of 188) but never as *C. fissilis* (Supplementary Data Table A.1).

284 From the Random Forest analyses, the most important compounds for species discrimination were
285 selected (Supplementary Data Table A.3). In total, 15 compounds were most frequent in over 58% of
286 the runs (100 runs). For some compounds we were able to infer the molecular formula and make
287 tentative assignments.

288 In most cases of the Random Forest analyses, each species was confused with three other species
289 (Supplementary Data Table A.2): *C. angustifolia* was mostly classified as *C. fissilis* or *C. saltensis* and
290 on one occasion as *C. balansae*. *C. balansae* was confused with all species except for *C. saltensis*. A
291 similar pattern holds for *C. fissilis*, although this species was mostly confused with *C. odorata*. Vice
292 versa, *C. odorata* was mostly confused with *C fissilis*, in addition to two samples that were mistakenly
293 identified as *C. saltensis*. Finally, *C. saltensis* was confused with all species, except for *C. balansae*.

294 *Identification of geographic origin*

295 The analysis for geographic origin was done for *C. fissilis* and *C. odorata* separately. Classification
296 performance was higher for Random Forest compared to Kernel Discriminant analysis. Furthermore,
297 Random Forest showed similar error rates for both species while Kernel Discriminant showed a
298 difference of 6.3% between *C. fissilis* and *C. odorata* (Table 3a and b). The error rate for site
299 identification was highly variable for both methods. The first six PCs were selected from the PCA
300 analysis (Figure 4) as they explained the highest variance: 78.9% in the case of *C. fissilis* and 86.2% for
301 *C. odorata*.

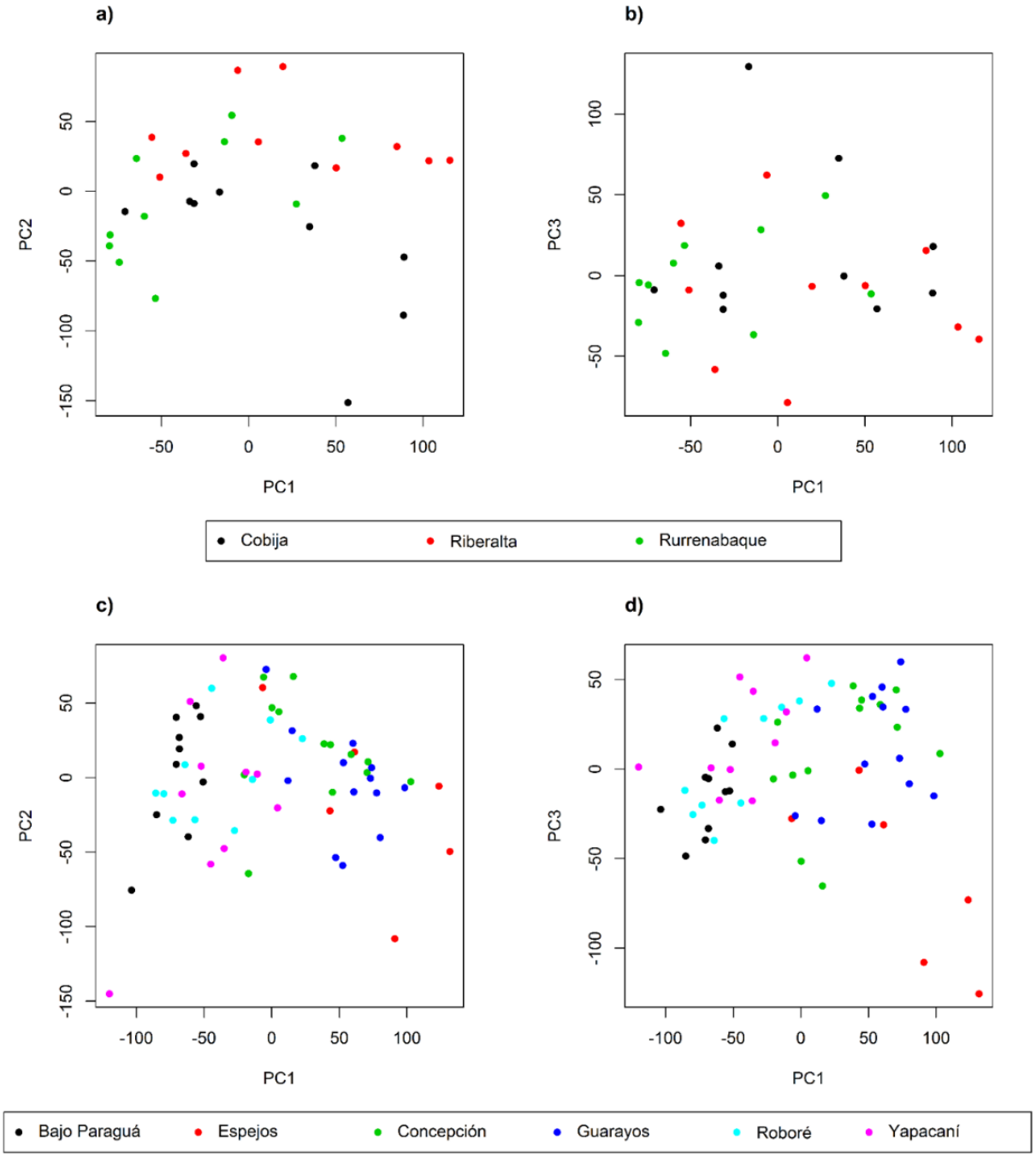


Figure 4 Results of Principal Component Analysis used for KDA analyses for geographic origin. Scatterplots combining (a) PC1 and PC2, (b) PC1 and PC3 for *C. odorata*, and (c) PC1 and PC2 and (d) PC1 and PC3 for *C. fissilis*.

306 **Table 3. Error classification for sites of *C. fissilis* (a) and *C. odorata* (b) based on KDA with 6 PCs,**
307 **and Random Forest analyses.** Mean classification and standard deviation (SD) were estimated using
308 the classification error per site after 100 runs with different training and testing sets.

309

	KDA		Random Forest	
	Mean error (%)	**SD (%)**	**Mean error (%)**	**SD (%)**
a) *C. fissilis* sites				
Bajo Paraguá	45.8	36.5	38.5	16.9
Espejos	43.3	44.4	86.7	11.7
Concepción	37.5	36.8	23.5	13.3
Guarayos	39.7	32.2	36.9	17.5
Roboré	80.9	33.3	57.3	17.9
Yapacaní	60.9	36.8	48.2	20.7
Mean	**53.9**	**12.5**	**42.7**	**4.8**
b) *C. odorata* sites				
Cobija	47.8	38.5	60.7	15.8
Riberalta	38.4	38.6	40.9	19
Rurrenabaque	48.5	38.5	30.4	21.6
Mean	**47.7**	**19.7**	**42.4**	**8.6**

310

311 KDA classification errors for *C. fissilis* samples were on average 53.9% (range 37.5% to 80.9%), while
312 those for *C. odorata* averaged 47.7% (range 38.4% to 48.5%, Table 3). Roboré and Yapacaní showed
313 the highest total mean error for sites discrimination of *C. fissilis* samples (Table 3a, Supplementary Data
314 Figure A.1c), and Concepción and Guarayos the lowest. Rurrenabaque showed the highest mean error
315 and Riberalta the lowest error for *C. odorata* sample classification (Table 3b, Supplementary Data
316 Figure A.1d).

317 There was misclassification between 3-4 other sites of origin (Supplementary Data Table A.4) with the
318 trained algorithm in KDA. However, some sites showed chemical characteristics clearly distinct from
319 other sites. For example, samples from Roboré and Bajo Paraguá were distinct from Espejos but this
320 site was often confused with Concepción and Guarayos. Samples from Bajo Paraguá and Espejos were
321 distinct from each other but wrongly assigned to Concepción and Yapacaní. Guarayos and Espejos were
322 distinct from Bajo Paraguá but were wrongly assigned to Concepción and Yapacaní.

323 Similarly to KDA, there was misclassification between 2-3 other sites of origin in the Random Forest
324 analyses (Supplementary Data Table A.5). For example, Bajo Paraguá was distinct from 3 sites:
325 Espejos, Concepción and Guarayos but some samples were wrongly assigned to Roboré and Yapacaní.
326 Roboré samples had a higher chance of being wrongly assigned to Bajo Paraguá compared with

Yapacaní. Samples from Espejos were wrongly assigned to all sites except Bajo Paraguá. The highest error for the identification of *C. fissilis* sites using Random Forest was observed in Espejos followed by Roboré and Yapacaní while Concepción showed the best performance with the lowest error rate (Supplementary Data Figure A.1e and f) (Table 3).

On the other hand, *C. odorata* showed the highest classification error for Rurrenabaque and lowest error for Riberalta. Samples from Cobija were confused with samples from Rurrenabaque and Riberalta (Supplementary Data Table A.6). However, Rurrenabaque samples were mostly assigned to Riberalta followed by Cobija. Riberalta was confused by the other two sites but it had the highest number of correct assignments.

Although Random Forest included a higher number of samples from different sites compared with KDA (24 samples), it performed similarly in error rates and assignments. Samples from Cobija were mostly wrongly assigned to Rurrenabaque and to a lesser extent to Riberalta (Supplementary Data Table A.7). Samples from Rurrenabaque were wrongly assigned to Riberalta and Cobija, at roughly equal frequencies. With this method, Rurrenabaque showed the highest number of correct assignments. In each of the 100 Random Forest analyses, the most important compounds for site discrimination were selected (Supplementary Data Table A.8).

Discussion

To combat the illegal trade in timber, independent methods to identify species and verify geographical origin need to be developed. In this study, we assessed the effectiveness of DART-TOFMS spectra followed by multivariate statistical analysis to determine the potential for differentiating *Cedrela* species and geographic origin of *Cedrela* timber. Overall species differentiation error was 15-19% (range for two statistical methods), while that for geographic origin was significantly higher (42-54%). These discrimination errors are higher compared with previous studies that applied DART-TOFMS, which reached discrimination errors of less than 10% for species discrimination (Lancaster and Espinoza, 2012; Musah *et al*., 2015; Evans *et al*., 2017) and of ~30% in distinguishing between sites of origin (Finch *et al*., 2017). We also found strong differences in discrimination error between species. Possible explanations for these differences include (1) low sample sizes for some species, (2) variation within species, (3) misidentification by the curator, or (4) variation across the sites where the species are found (e.g. some species are found together as *C. fissilis* and *C. balansae*). We will discuss these possible causes below.

Low sample size can lead to higher error rates. This is exemplified by *C. montana*, of which only two samples were collected. Including this species in the analyses increased the error of species identification from 15 to 30%. Yet other studies that applied DART-TOFMS in species with small

sample sizes have successfully discriminated between species (Lancaster and Espinoza, 2012; McClure *et al*., 2015; Wiemann and Espinoza, 2017). This discrepancy depends on the degree of chemical variation which is much smaller in some species than in others. This variability was evidenced by *C. fissilis* and *C. odorata* which showed the lowest error rates in the species discrimination analysis compared with *C. angustifolia* and *C. balansae* which showed the highest error rates. This indicates that the accuracy of discrimination is highly species specific which thwarts extrapolating these results to other species and sites. Nevertheless, a more accurate conclusion can be reached by identifying representative chemical compounds in a heatmap. This graphical overview facilitates the discovery of particular trends, such as species-specific chemicals. Another possible source of error is misidentification by the curator. This possible observer bias could be solved by having multiple curators identify and compare herbarium samples before further analysis. In this study, the samples identified were based on a large herbarium collection and previous identifications of *Cedrela* samples throughout Bolivia.

The low accuracy of site discrimination may also be caused by local conditions such as climate, soil characteristics and nutrient availability which seem to affect tree performance and composition (Gentry *et al*., 1995; Medina, 1995; Oliveira-Filho *et al*., 1998; Toledo *et al*., 2011). In the Meliaceae family, Noldt *et al*. (2001) found that some species were more sensitive to environmental conditions due to root systems in the upper soil layers. The *Cedrela* samples in our study also showed superficial tree roots and site-specific growth variation (Paredes-Villanueva *et al*., 2016) which indicates that these trees display site-specific characteristics that may have played an important role in wood formation. Such site characteristics vary from large scale, e.g. ecosystem under different climatic regimes to small scale *e.g.* the micro site factors that contribute to tree development (Reifsnyder *et al*., 1971). The scale variation of site identification may have played a role in our discrimination among sites: *C. odorata* sites were more distant than *C. fissilis* sites. This was confirmed when only Bajo Paraguá, Roboré and Yapacaní (the most distant sites of *C. fissilis*) were analysed: the accuracy remained similar with Random Forest (57%) and increased to 53% with KDA (data not shown). These results suggest that discriminating between more distant regions or locations may result in higher accuracies than discriminating among neighboring sites.

Apart from these external factors that influence discrimination error, the two statistical analyses we used (Random Forest and KDA) also resulted in different error rates. These errors can be reduced by comparing the probabilities of being assigned to another group. Therefore, KDA and Random Forest would best be used alongside each other as triangulation methods. Comparing and cross-checking results between groups and statistical methods will increase the certainty in identifying species and site of origin. In addition, results should also be complemented by other independent statistical tools. Consistent results of these statistical methods could increase the confidence of correct identification

when analysing the spectra generated by DART-TOFMS. Decision Trees (Kamiński *et al*., 2018; Therneau *et al*., 2018) or other machine learning algorithms that would also provide information of the less abundant chemicals could also be used as multiple approximation methods.

Finally, DART-TOFMS is a qualitative analysis; in order to investigate the role of distance, rainfall, altitude and the chemical composition of *Cedrela* trees in predicting the likelihood of belonging to the conditions of a specific site, it is necessary to apply a quantitative chemical approach. Such an analysis, in which the effects of sample size and time on the detection accuracy of the chemical signals are measured, will allow us to interpret the resulting molecular mass spectra across different spatial and temporal scales. The within-the-tree variation and among-site differentiation of the chemical compounds of the same species represents a great potential for more specific characterization.

All samples in this study were collected in Bolivia, a country that is severely impacted by illegal trade in timber, including *Cedrela* species. The methods used in this study showed the high potential of mass spectrometry for use in *Cedrela* species identification in Bolivia, with the highest confidence in identifying *C. fissilis*. DART TOFMS analysis can easily separate *Cedrela* genus trees from the other look-alike species, like *Swietenia macrophylla* King and *Carapa guianensis* Aubl. (Braga *et al*., 2011; Bergo *et al*., 2016), and this would help when false declarations and documents are being used. Previous studies also found that most of the difficulties of *Cedrela* identifications were at the species level rather than at the genus level (Gasson, 2011). Also, the accuracy of identification between samples from the genera *Dalbergia* and *Machaerium* was >95% (Espinoza *et al*., 2015; Lancaster and Espinoza, 2012). This suggests that DART TOFMS analysis may perform better in distinguishing between *Cedrela* and other look-alike genera, but suffers in species specific assignment within the taxa.

Conclusion

Cedrela species belong to a timber genus that has been overexploited in the last couple of years. The regulation of their trading has presented many challenges, given that the identification of those species that belong to the CITES list is difficult because of similar wood anatomical characteristics. Our approach offers a strategy for improving identification certainty of *Cedrela* species by using a complementary approach contributing to their proper forest management and conservation. DART-TOFMS offers an alternative for identification and chemical discrimination among such species. There are several statistical methods to analyse the data generated by DART-TOFMS. Consistent results of two statistical methods (discriminant analyses: KDA and Random Forest) were found in this study, and applying both methods on the same dataset is recommended. Our results reveal potential for *Cedrela* species assignment (81-85% accuracy), particularly for *C. fissilis* (95.6%). Our results also show that discrimination of geographical origin is not possible due to low assignment (with accuracies of 46-57% for *C. fissilis* and 52-58% for *C. odorata*). Thus, the mass spectrometric approach used here can help to

identify species provenance of certain Bolivian *Cedrela* timbers, but not geographic provenance within the country.

Acknowledgements

We would like to thank the Museo de Historia Natural Noel Kempff Mercado for their support on the species identification. This research was financed by the NFP/Nuffic fellowship (the Netherlands). Fieldwork was logistically supported by Universidad Autónoma Gabriel René Moreno and financed by Alberta Mennega Stichting and The Rufford Foundation. We are also grateful to Christoph Ruttkies and Tarn Duong for their recommendations and assistance with the statistical analysis. We are also thankful to Pieter Baas and Nicolien Sol for their help during the collection of samples and analysis of data.

Conflict of interest statement

None declared.

Funding

This work was supported by the Dutch organization for internationalization in education (NUFFIC) [NFP-PhD.14/61]. Fieldwork was financed by Alberta Mennega Stichting; and The Rufford Foundation [18 670–1].

References

ABT. 2017 Audiencia pública inicial y parcial de rendición de cuentas (de enero hasta agosto, gestión 2017). Autoridad de Fiscalización y Control Social de Bosques y Tierra (ABT), Santa Cruz, Bolivia.

Afendi, F.M., Okada, T., Yamazaki, M., Hirai-Morita, A., Nakamura, Y., Nakamura, K. *et al.* 2012 KNApSAcK family databases: integrated metabolite–plant species databases for multifaceted plant research. *Plant and Cell Physiology*, **53**, e1(1-12).

Baudat, G. and Anouar, F. 2000 Generalized Discriminant Analysis using a Kernel approach. *Neural Comput.*, **12** (10), 2385-2404.

Bergo, M.C.J., Pastore, T.C.M., Coradin, V.T.R., Wiedenhoeft, A.C. and Braga, J.W.B. 2016 NIRS identification of *Swietenia macrophylla* is robust across specimens from 27 countries. *IAWA Journal*, **37** (3), 420-430.

Braga, J.W.B., Pastore, T.C.M., Coradin, V.T.R., Camargos, J.A.A. and da Silva, A.R. 2011 The use of Near Infrared Spectroscopy to identify solid wood specimens of *Swietenia macrophylla* (Cites Appendix II). *IAWA Journal*, **32** (2), 285-296.

Breiman, L. 2001 Random forests. *Machine learning*, **45** (1), 5-32.

Cerutti, P.O. and Lescuyer, G. 2011 *The domestic market for small-scale chainsaw milling in Cameroon: Present situation, opportunities and challenges*. Center for International Forestry Research (CIFOR): Bogor, Indonesia.

Chatterjee, A., Chakrabortty, T. and Chandrasekharan, S. 1971 Chemical investigation of *Cedrela toona*. *Phytochemistry*, **10** (10), 2533-2535.

CITES. 2017 Convention on International Trade in Endangered Species of Wild Fauna and Flora. Appendices I, II and III. https://cites.org/eng/app/appendices.php (accessed 09 December 2017).

Cody, R.B., Laramée, J.A. and Durst, H.D. 2005 Versatile new ion source for the analysis of materials in open air under ambient conditions. *Analytical chemistry*, **77** (8), 2297-2302.

Compt, J. and Christy, T. 2008 The 14th meeting of the Conference of the Parties to CITES. *Traffic Bulletin*, **21** (3), 101.

Cordeiro, J.R., Martinez, M.I.V., Li, R.W.C., Cardoso, A.P., Nunes, L.C., Krug, F.J. *et al.* 2012 Identification of four wood species by an electronic nose and by LIBS. *International Journal of Electrochemistry*, **2012**, 5.

Deklerck, V., Finch, K., Gasson, P., Van den Bulcke, J., Van Acker, J., Beeckman, H. *et al.* 2017 Comparison of species classification models of mass spectrometry data: Kernel Discriminant Analysis vs Random Forest; A case study of Afrormosia (*Pericopsis elata* (Harms) Meeuwen). *Rapid communications in mass spectrometry : RCM*, **31** (19), 1582-1588.

Duong, T. 2007 ks: Kernel density estimation and kernel discriminant analysis for multivariate data in R. *Journal of Statistical Software*, **21** (7), 1-16.

Duong, T. 2017 ks: Kernel Smoothing. https://cran.r-project.org/web/packages/ks/.

Eason, H.M. and Setzer, W.N. 2007 Bark essential oil composition of *Cedrela tonduzii* C. DC. (Meliaceae) from Monteverde, Costa Rica *Records of Natural Products*, **1** (2-3), 24-27.

Espinoza, E.O., Wiemann, M.C., Barajas-Morales, J., Chavarria, G.D. and McClure, P.J. 2015 Forensic analysis of CITES-protected *Dalbergia* timber from the Americas. *IAWA Journal*, **36** (3), 311-325.

Evans, P.D., Mundo, I.A., Wiemann, M.C., Chavarria, G.D., McClure, P.J., Voin, D. *et al.* 2017 Identification of selected CITES-protected Araucariaceae using DART TOFMS. *IAWA Journal*, **38** (2), 266-S263.

Fidelis, C.H.V., Augusto, F., Sampaio, P.T.B., Krainovic, P.M. and Barata, L.E.S. 2012 Chemical characterization of rosewood (*Aniba rosaeodora* Ducke) leaf essential oil by comprehensive two-dimensional gas chromatography coupled with quadrupole mass spectrometry. *Journal of Essential Oil Research*, **24** (3), 245-251.

Finch, K., Espinoza, E., Jones, F.A. and Cronn, R. 2017 Source identification of western Oregon Douglas-fir wood cores using mass spectrometry and random forest classification. *Applications in plant sciences*, **5** (5), 1600158.

Förstel, H., Boner, M., Höltken, A., Fladung, M., Degen, B. and Zahnen, J. 2011 *Fighting illegal logging through the introduction of a combination of the isotope method for identifying the origins of timber and DNA analysis for differentiation of tree species*. WWF Germany: Berlin, Germany.

Gasson, P. 2011 How precise can wood identification be? Wood anatomy's role in support of the legal timber trade, especially cites. *IAWA Journal*, **32** (2), 137-154.

Gasson, P., Baas, P. and Wheeler, E. 2011 Wood anatomy of Cites-listed tree species. *IAWA Journal*, **32** (2), 155-198.

Gentry, A.H., Bullock, S.H., Mooney, H.A. and Medina, E. 1995 Seasonally dry tropical forests. *Seasonally dry tropical forests*.

Gillies, A., Navarro, C., Lowe, A., Newton, A.C., Hernandez, M., Wilson, J. *et al.* 1999 Genetic diversity in Mesoamerican populations of mahogany (*Swietenia macrophylla*), assessed using RAPDs. *Heredity (Edinb)*, **83** (6), 722-732.

Gotelli, N. and Ellison, A. 2004 *A primer of ecological statistics. Sinauer Associates Inc*. Sunderland, 614 pp.

Hillis, W.E. 1987 *Heartwood and tree exudates*. Springer-Verlag.

Kagawa, A. and Leavitt, S.W. 2010 Stable carbon isotopes of tree rings as a tool to pinpoint the geographic origin of timber. *Journal of Wood Science*, **56** (3), 175-183.

Kamiński, B., Jakubczyk, M. and Szufel, P. 2018 A framework for sensitivity analysis of decision trees. *Central European Journal of Operations Research*, **26** (1), 135-159.

Kishor, N. and Lescuyer, G. 2012 Controlling illegal logging in domestic and international markets by harnessing multi-level governance opportunities. *International Journal of the Commons*, **6**, 255-270.

Lago, J.H.G., de Ávila, P., de Aquino, E.M., Moreno, P.R.H., Ohara, M.T., Limberger, R.P. *et al.* 2004 Volatile oils from leaves and stem barks of *Cedrela fissilis* (Meliaceae): chemical composition and antibacterial activities. *Flavour Fragr. J.*, **19** (5), 448-451.

Lancaster, C. and Espinoza, E. 2012 Analysis of select *Dalbergia* and trade timber using direct analysis in real time and time-of-flight mass spectrometry for CITES enforcement. *Rapid Commun. Mass Spectrom.*, **26** (9), 1147-1156.

Lawson, S. and MacFaul, L. 2010 *Illegal Logging and Related Trade: Indicators of the Global Response*. Chatham House: London, UK.

Lescuyer, G., Ndotit, S., Ndong, L.B.B., Tsanga, R. and Cerutti, P.O. 2014 *Policy options for improved integration of domestic timber markets under the voluntary partnership agreement (VPA) regime in Gabon*. Center for International Forestry Research (CIFOR): Bogor, Indonesia, 4 pp.

Liaw, A. and Wiener, M. 2002 Classification and regression by randomForest. *R news*, **2** (3), 18-22.

Maia, B.H.L.N.S., Paula, J.R.d., Sant'Ana, J., Silva, M.F.d.G.F.d., Fernandes, J.B., Vieira, P.C. *et al.* 2000 Essential oils of *Toona* and *Cedrela* species (Meliaceae): taxonomic and ecological implications. *Journal of the Brazilian Chemical Society*, **11**, 629-639.

McClure, P.J., Chavarria, G.D. and Espinoza, E. 2015 Metabolic chemotypes of CITES protected *Dalbergia* timbers from Africa, Madagascar, and Asia. *Rapid Commun. Mass Spectrom.*, **29** (9), 783-788.

Medina, E. 1995 Diversity of life forms of higher plants in neotropical dry forests. In *Seasonally Dry Tropical Forests.* E. Medina, H.A. Mooney and S.H. Bullock (eds.), Cambridge University Press, Cambridge, pp. 221-242.

Mosedale, J.R. and Ford, A. 1996 Variation of the flavour and extractives of European oak wood from two French forests. *J. Sci. Food Agric.*, **70** (3), 273-287.

Mostacedo, B. and Fredericksen, T. 2001 *Regeneración y silvicultura de bosques tropicales en Bolivia*. Editora El País: Santa Cruz, Bolivia.

Mostacedo, B. and Fredericksen, T.S. 1999 Regeneration status of important tropical forest tree species in Bolivia: assessment and recommendations. *Forest Ecology and Management*, **124** (2), 263-273.

Mostacedo, B., Justiniano, J., Toledo, M. and Fredericksen, T. 2003 *Guía dendrológica de especies forestales de Bolivia*. Segunda Edición. Proyecto de Manejo Forestal Sostenible.

Moya, R. and Calvo-Alvarado, J. 2012 Variation of wood color parameters of *Tectona grandis* and its relationship with physical environmental factors. *Annals of Forest Science*, **69** (8), 947-959.

Moya, R., Wiemann, M.C. and Olivares, C. 2013 Identification of endangered or threatened Costa Rican tree species by wood anatomy and fluorescence activity. *Revista de Biología Tropical*, **61**, 1113-1156.

Musah, R.A., Espinoza, E.O., Cody, R.B., Lesiak, A.D., Christensen, E.D., Moore, H.E. *et al.* 2015 A high throughput ambient mass spectrometric approach to species identification and classification from chemical fingerprint signatures. *Scientific Reports*, **5**, 11520.

Navarro-Cerrillo, R.M., Agote, N., Pizarro, F., Ceacero, C.J. and Palacios, G. 2013 Elements for a non-detriment finding of *Cedrela* spp. in Bolivia—A CITES implementation case study. *Journal for Nature Conservation*, **21** (4), 241-252.

Navarro, G. 2011 *Clasificación de la vegetación de Bolivia*. Centro de Ecología Difusión Simón I. Patiño: Santa Cruz, 713 pp.

Noldt, G., Bauch, J., Koch, G. and Schmitt, U. 2001 Fine roots of *Carapa guianensis* Aubl. and *Swietenia macrophylla* King: cell structure and adaptation to the dry season in Central Amazonia. *Journal of Applied Botany*, **75** (3-4), 152-158.

Oliveira-Filho, A.T., Curi, N., Vilela, E.A. and Carvalho, D.A. 1998 Effects of canopy gaps, topography, and soils on the distribution of woody species in a central Brazilian deciduous dry forest. *Biotropica*, **30** (3), 362-375.

Paredes-Villanueva, K., López, L. and Navarro-Cerrillo, R.M. 2016 Regional chronologies of *Cedrela fissilis* and *Cedrela angustifolia* in three forest types and their relation to climate. *Trees*, **30** (5), 1581-1593.

Pettersen, R.C. 1984 The chemical composition of wood. In *The Chemistry of Solid Wood.* R. Rowell (ed.), American Chemical Society, Washington, D.C., pp. 57-126.

R Development Core Team. 2017 *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/

Reifsnyder, W.E., Furnival, G. and Horowitz, J. 1971 Spatial and temporal distribution of solar radiation beneath forest canopies. *Agricultural Meteorology*, **9**, 21-37.

Seneca Creek Associates, L. 2004 *"Illegal" logging and global wood markets: The competitive impacts on the U.S. wood products industry*. American Forest & Paper Association.

Stark, T. and Pang Cheung, S. 2006 *Sharing the blame: Global consumption and China 's role in ancient forest destruction*: Beijing, China.

Therneau, T., Atkinson, B. and Ripley, B. 2018 rpart: Recursive Partitioning and Regression Trees. https://CRAN.R-project.org/package=rpart.

Toledo, M., Chevallier, B., Villaroel, D. and Mostacedo, B. 2008 *Ecología y silvicultura. de especies menos conocidas: Cedro, Cedrela spp.* Proyecto BOLFOR II/ Instituto Boliviano de Investigación Forestal: Santa Cruz, Bolivia.

Toledo, M., Poorter, L., Peña-Claros, M., Alarcón, A., Balcázar, J., Leaño, C. *et al.* 2011 Climate is a stronger driver of tree and forest growth rates than soil and disturbance. *Journal of Ecology*, **99** (1), 254-264.

Vlam, M., de Groot, G.A., Boom, A., Copini, P., Laros, I., Veldhuijzen, K. *et al.* 2018 Developing forensic tools for an African timber: Regional origin is revealed by genetic characteristics, but not by isotopic signature. *Biological Conservation*, **220**, 262-271.

Wickham, H., Francois, R., Henry, L., Müller, K. and RStudio. 2017 dplyr: A grammar of data manipulation. A fast, consistent tool for working with data frame like objects, both in memory and out of memory. https://CRAN.R-project.org/package=dplyr.

Wiemann, M.C. and Espinoza, E.O. 2017 Species verification of *Dalbergia nigra* and *Dalbergia spruceana* samples in the wood collection at the Forest Products Laboratory. *Review Process: Informally Refereed (Peer-Reviewed)*.

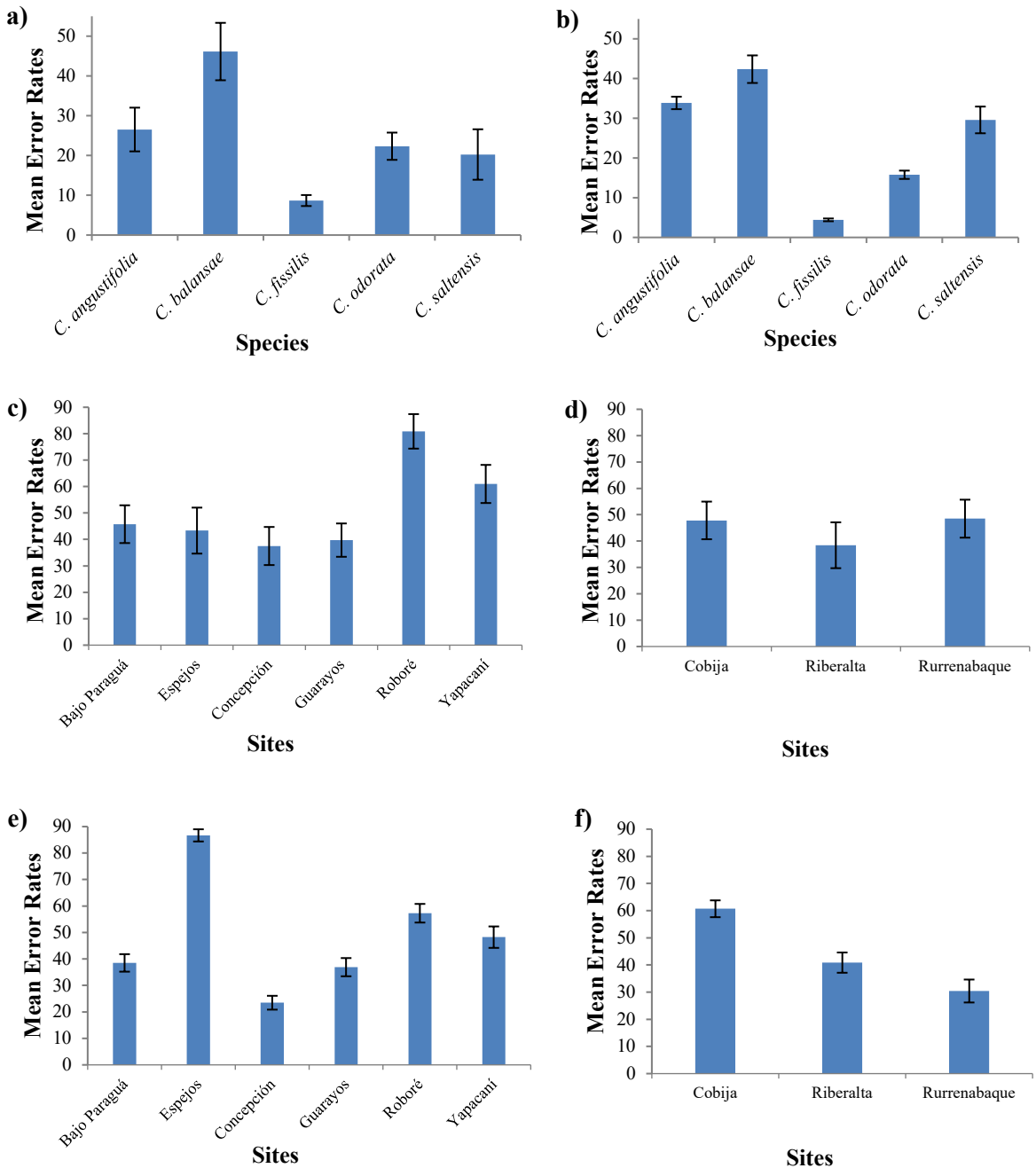
Wilkins, A.P. and Stamp, C.M. 1990 Relationship between wood colour, silvicultural treatment and rate of growth in *Eucalyptus grandis* Hill (Maiden). *Wood Science and Technology*, **24** (4), 297-304.

Wit, M., Van Dam, J., Cerutti, P.O., Lescuyer, G., Kerrett, R. and McKeown, J.P. 2010 *Chainsaw milling: supplier to local markets - a synthesis*. ETFRN: Wageningen, The Netherlands.

Zobel, B.J. and van Buijtenen, J.P. 1989 *Wood variation: Its causes and control*. Springer Berlin Heidelberg.

607 **Supplementary material**

608



609

610

611 **Figure A.1. Mean error rates of a) the Kernel Discriminant Analysis and b) Random Forest**
612 **analysis for species analyses. Mean error rates for c) Kernel Discriminant Analysis per site for *C.***
613 ***fissilis* and d) Kernel Discriminant Analysis per site for *C. odorata*, e) Random Forest analyses**
614 **per site for *C. fissilis* and f) Random Forest analyses per site for *C. odorata*.** The whiskers show the
615 standard error of the data.

Table A.1. Confusion matrix of species discrimination and frequency of species (%) in each randomized classification table using KDA

KDA		*C. angustifolia*	*C. balansae*	*C. fissilis*	*C. odorata*	*C. saltensis*	Total
C. angustifolia	**Total samples**	232	23	20	20	20	315
	%	73.7	7.3	6.3	6.3	6.3	100
C. balansae	**Total samples**	44	112	44	12	10	222
	%	19.8	50.5	19.8	5.4	4.5	100
C. fissilis	**Total samples**	32	18	1107	56	0	1213
	%	2.6	1.5	91.3	4.6	0	100
C. odorata	**Total samples**	1	0	118	432	11	562
	%	0.2	0	21.0	76.9	2.0	100
C. saltensis	**Total samples**	10	2	0	33	143	188
	%	5.3	1.1	0	17.6	76.1	100

Table A.2. Confusion matrix of species discrimination and frequency of species (%) in each randomized classification table using Random Forest

Random Forest		*C. angustifolia*	*C. balansae*	*C. fissilis*	*C. odorata*	*C. saltensis*	Total
C. angustifolia	**Total samples**	796	1	236	0	162	1195
	%	66.6	0.1	19.7	0	13.6	100
C. balansae	**Total samples**	92	470	177	54	0	793
	%	11.6	59.3	22.3	6.8	0	100
C. fissilis	**Total samples**	4	4	4752	211	0	4971
	%	0.1	0.1	95.6	4.2	0	100
C. odorata	**Total samples**	0	0	369	2010	2	2381
	%	0	0	15.5	84.4	0.1	100
C. saltensis	**Total samples**	91	0	13	81	475	660
	%	13.8	0	2.0	12.3	72.0	100

622 **Table A.3. List of the 15 most important chemical compounds obtained from 50,000 runs of**
623 **Random Forests.** The numbers are the mass-to-charge ratios (*m/z*).

m/z	% of runs including the compound	Molecular formula	Tentative assignments
501.278	100	$C_{28}H_{38}O_8$ -H	3,7-Dideacetylkhivorin
500.265	100	$C_{28}H_{34}O_7$ +NH_4	Gedunin
484.245	100	$C_{27}H_{34}O_9$ – H_2O	Cedrodorin
483.244	100	$C_{28}H_{34}O_7$ +H	Gedunin
469.344	99	$C_{27}H_{32}O_7$ +H	Mexicanolide
528.412	92	-	-
451.337	92	$C_{27}H_{32}O_7$ - OH	Mexicanolide
229.200	84	-	-
227.095	83	$C_{15}H_{24}$ +Na	delta-Cadinene
357.136	79	$C_{21}H_{24}O_5$ +H	-
470.335	74	$C_{27}H_{36}O_8$ –H_2O	Swiemahogin A
452.307	74	$C_{27}H_{34}O7$ –H_2O	Methyl angolensate
527.418	71	$C_{29}H_{36}O_{10}$ -OH	6-Acetoxycedrodorin
471.347	67	$C_{27}H_{34}O_7$ +H	Methyl angolensate
507.399	58	$C_{30}H_{36}O_7$ -H	Mahonin

624

625 **Table A.4. Confusion matrix of sites discrimination and frequency of sites (%) in each**
626 **randomized classification table using KDA in *C. fissilis***

KDA		Bajo Paraguá	Espejos	Concepción	Guarayos	Roboré	Yapacaní	Total
Bajo Paraguá	**Total samples**	101	0	1	0	82	25	209
	%	48.3	0	0.5	0	39.2	12.0	100
Espejos	**Total samples**	0	83	39	30	0	3	155
	%	0	53.5	25.2	19.4	0	1.9	100
Concepción	**Total samples**	0	0	144	49	17	43	253
	%	0	0	56.9	19.3	6.7	17.0	100
Guarayos	**Total samples**	0	2	95	157	4	6	264
	%	0	0.8	36.0	59.5	1.5	2.3	100
Roboré	**Total samples**	120	0	40	0	36	17	213
	%	56.3	0	18.8	0	16.9	8.0	100
Yapacaní	**Total samples**	48	22	8	0	50	78	206
	%	23.3	10.7	3.9	0	24.3	37.9	100

627

Table A.5. Confusion matrix of sites discrimination and frequency of sites (%) in each randomized classification table using Random Forest in *C. fissilis*

Random Forest		Bajo Paraguá	Espejos	Concepción	Guarayos	Roboré	Yapacaní	Total
Bajo Paraguá	**Total samples**	507	0	0	0	235	65	807
	%	62.8	0	0	0	29.1	8.1	100
Espejos	**Total samples**	0	68	216	181	1	3	469
	%	0	14.5	46.1	38.6	0.2	0.6	100
Concepción	**Total samples**	0	12	805	221	0	0	1038
	%	0	1.2	77.6	21.3	0	0	100
Guarayos	**Total samples**	0	11	342	649	0	0	1002
	%	0	1.1	34.1	64.8	0	0	100
Roboré	**Total samples**	369	0	13	0	361	66	809
	%	45.6	0	1.6	0	44.6	8.2	100
Yapacaní	**Total samples**	186	0	17	0	152	420	775
	%	24.0	0	2.2	0	19.6	54.2	100

Table A.6. Confusion matrix of sites discrimination and frequency of sites (%) in each randomized classification table using KDA in *C. odorata*

KDA		Cobija	Riberalta	Rurrenabaque	Total
Cobija	**Total samples**	107	23	80	210
	%	51.0	11.0	38.1	100
Riberalta	**Total samples**	6	114	74	194
	%	3.1	58.8	38.1	100
Rurrenabaque	**Total samples**	34	69	93	196
	%	17.3	35.2	47.4	100

Table A.7. Confusion matrix of sites discrimination and frequency of sites (%) in each randomized classification table using Random Forest in *C. odorata*

Random Forest		Cobija	Riberalta	Rurrenabaque	Total
Cobija	**Total samples**	319	182	288	789
	%	40.4	23.1	36.5	100
Riberalta	**Total samples**	181	485	137	803
	%	22.5	60.4	17.1	100
Rurrenabaque	**Total samples**	112	117	579	808
	%	13.9	14.5	71.7	100

Table A.8. List of the most important chemical compounds obtained from 50,000 of Random Forests to identify site of origin for *C. fissilis* and C. *odorata*. The numbers are the mass-to-charge ratios (*m/z*).

C. fissilis				*C. odorata*			
m/z	**% of runs including the compound**	**Molecular formula**	**Tentative assignment**	***m/z***	**% of runs including the compound**	**Molecular formula**	**Tentative assignment**
149.123	100	-	-	468.307	97	-	-
122.075	100	-	-	467.344	96	-	-
121.067	100	-	-	527.418	88	$C_{29}H_{36}O_9$ -H	Methyl 3beta-acetoxy-6-hydroxy-1-oxomeliac-14-enoate
109.098	100	-	-	673.281	77	$C_{35}H_{46}O_{14}$ - OH	Meliacarpinin D
279.165	100	-	-	583.22	77	-	-
123.044	99	-	-	81.035	63	-	-
280.164	97	-	-	486.349	58	-	-
274.112	96	-	-	99.044	58	-	-
274.5	95	-	-	192.142	58	-	-
150.072	94	-	-	470.335	50	$C_{27}H_{34}O_7$	Methyl angolensate
95.087	84	-	-	117.053	48	-	-
135.103	83	$C_{10}H_{14}$ +H	p-Cymene	303.449	44	-	-
206.201	79	-	-	528.412	43	-	-
104.069	53	-	-	469.344	41	$C_{27}H_{32}O_7$ +H	Mexicanolide
275.276	50	-	-				
81.035	47	-	-				
379.292	44	-	-				
204.186	40	$C_{15}H_{26}O$ – H_2O	T-Muurolol				