Received: 12 May 2017 | Revised: 28 June 2017 | Accepted: 29 June 2017

DOI: 10.1002/rcm.7939

WILEY Rapid Communications in Mass Spectrometry

RESEARCH ARTICLE

Comparison of species classification models of mass spectrometry data: Kernel Discriminant Analysis vs Random Forest; A case study of Afromosia (*Pericopsis elata* (Harms) Meeuwen)

V. Deklerck[1,4] | K. Finch[2] | P. Gasson[3] | J. Van den Bulcke[1] | J. Van Acker[1] | H. Beeckman[4] | E. Espinoza[5]

[1] Woodlab-UGent, Ghent University, Laboratory of Wood Technology, Department of Forest and Water Management, Coupure Links 653, B-9000 Ghent, Belgium

[2] Department of Botany and Plant Pathology, Oregon State University, Cordley Hall, 2701 SW Campus Way, Corvalis, OR, USA

[3] Royal Botanic Gardens, Kew, Richmond, TW9 3DS, UK

[4] Wood Biology Service, Royal Museum for Central Africa (RMCA), Leuvensesteenweg 13, 3080 Tervuren, Belgium

[5] U.S. National Fish and Wildlife Forensic Laboratory, 1490 East Main Street, Ashland, OR, USA

Correspondence
V. Deklerck, Woodlab-UGent, Ghent University, Laboratory of Wood Technology, Department of Forest and Water Management, Coupure Links 653, B-9000 Ghent, Belgium.
Email: victor.deklerck@ugent.be

Funding information
HerbaXylaRedd belspo-project, Grant/Award Number: BR/143/A3/HERBAXYLAREDD

Rationale: The genus *Pericopsis* includes four tree species of which only *Pericopsis elata* (Harms) Meeuwen is of commercial interest. Enforcement officers might have difficulties discerning this CITES-listed species from some other tropical African timber species. Therefore, we tested several methods to separate and identify these species rapidly in order to enable customs officials to uncover illegal trade. In this study, two classification methods using Direct Analysis in Real Time (DART™) ionization coupled with Time-of-Flight Mass Spectrometry (DART-TOFMS) data to discern between several species are presented.

Methods: Metabolome profiles were collected using DART™ ionization coupled with TOFMS analysis of heartwood specimens of all four *Pericopsis* species and *Haplormosia monophylla* (Harms) Harms, *Dalbergia melanoxylon* Guill. & Perr. Harms, and *Milicia excelsa* (Welw.) C.C. Berg. In total, 95 specimens were analysed and the spectra evaluated. Kernel Discriminant Analysis (KDA) and Random Forest classification were used to discern the species.

Results: DART-TOFMS spectra obtained from wood slivers and post-processing analysis using KDA and Random Forest classification separated *Pericopsis elata* from the other *Pericopsis* taxa and its lookalike timbers *Haplormosia monophylla*, *Milicia excelsa*, and *Dalbergia melanoxylon*. Only 50 ions were needed to achieve the highest accuracy.

Conclusions: DART-TOFMS spectra of the taxa were reproducible and the results of the chemometric analysis provided comparable accuracy. *Haplormosia monophylla* was visually distinguished based on the heatmap and was excluded from further analysis. Both classification methods, KDA and Random Forest, were capable of distinguishing *Pericopsis elata* from the other *Pericopsis* taxa, *Milicia excelsa*, and *Dalbergia melanoxylon*, timbers that are commonly traded.

1 | INTRODUCTION

1.1 | Species characteristics and international trade

Pericopsis elata, commonly known as Afrormosia, is an emblematic species of the African rainforests that has been protected by the Convention on International Trade in Endangered Species (CITES)[1] since 1992. Its heartwood is characterized by high natural durability, mechanical strength, and dimensional stability. This combination makes it suitable for the most demanding applications of wood, especially for exterior joinery. The decorative value of the wood is also appreciated for the production of luxury furniture and parquetry. In some parts of the rainforest belt, the species is common and available in quantities large enough for industrial logging and the international timber trade. The market discovered the species as a precious wood, named Afrormosia, after the Second World War. The *P. elata* populations of Ghana were logged followed by those of Côte-d'Ivoire soon after.[2] These loggings were not based on management plans aiming at a sustainable yield and resulted in overexploitation after only a few decades. The West-African countries are therefore no longer considered a source of *P. elata* timber. The logging shifted gradually to the Central African rainforests of Cameroon, the Republic of Congo

(where the species is relatively rare), and the Kisangani region of the Democratic Republic of the Congo (DRC) in this order. There are two additional *Pericopsis* species in Africa: *P. angolensis* (Baker) Meeuwen and *P. laxiflora* (Baker) Meeuwen, whereas a single species is endemic to Asia: *P. mooniana*. *P. angolensis* also produces high-quality durable timber. However, trees of this species are less abundant and too small or poorly shaped for commercial exploitation, except in Mozambique, where the wood is sometimes traded as Muwanga or mixed with harvested *P. elata* timber.[2] *P. laxiflora* is similar, with the same uses but is not available in large sizes, and is by some researchers considered a subspecies of *P. angolensis*.[2] The Asian species, *P. mooniana*, which ranges from Sri Lanka east to New Guinea and Micronesia, is mainly exported from Indonesia to Japan.[3]

Because of law enforcement concerns, there is a need to distinguish *P. elata* from the other *Pericopsis* species and lookalike timbers. There have been documented fraudulent imports of *P. elata* declared as *Milicia excelsa*, a non-CITES listed species. The timber of *P. elata* can also be confused with *Dalbergia melanoxylon* (CITES App. II) from Africa. Traditional identification of wood has relied on anatomical features such as those in the extensive online database InsideWood.[4] When searching InsideWood using standardized wood anatomical features of *P. elata*, the results indicate that several other species, such as *D. melanoxylon* and *Haplormosia monophylla*, have similar wood structures. *D. melanoxylon* is a timber species that also occurs in Central Africa. *H. monophylla*, which is taxonomically closely related to *P. elata*, also occurs in Africa and it is traded by the common name of Idewa. To a lesser extent, the timber of the three other *Pericopsis* species might also be sold or confused with *P. elata*.[2]

1.2 | Species identification based on wood anatomy

It is therefore important for law enforcement officers to be able to discriminate between the abovementioned species. The anatomical features described in the IAWA Hardwood List[5] and used in InsideWood are adequate for narrowing down the number of possible identities of a hardwood sample, but their discriminatory ability is limited for closely related taxa with very similar features. This is the case with *Pericopsis* and *Haplormosia*, which have similar paratracheal axial parenchyma (ranging from scanty paratracheal through vasicentric to aliform to confluent and banded, especially in *Haplormosia*), storied axial parenchyma in mainly four-celled strands, and rays generally up to three or four cells wide and storied. *P. mooniana* appears to have sparser vessels and more distinct aliform and confluent parenchyma than the other *Pericopsis* species, but this observation is based on a single microscope slide in Kew's reference collection (Royal Botanic Gardens, Richmond, UK), and the literature in InsideWood and Plant resources of South-East Asia 5 (PROSEA).[3] It becomes more difficult to differentiate between *P. angolensis*, *P. elata*, and *P. laxiflora* based on wood anatomical features. Comparing these species using InsideWood leads to minor and variable differences. Only *P. elata* appears to have vasicentric axial parenchyma. However, vasicentric axial parenchyma was also present in two transverse sections of *P. angolensis*, provided by the Royal Museum for Central Africa (RMCA, Tervuren, Belgium). *P. laxiflora* appears to have more bands of parenchyma and few to no high rays. This was observed by comparison of two transverse and tangential sections with two and three slides of *P. elata* and *P. angolensis*, respectively. Another interesting feature is the presence of unilateral parenchyma in *P. elata*, which is rarer in *P. angolensis* and almost lacking in *P. laxiflora*. The wood of *M. excelsa* is very unlikely to be confused with *P. elata* or *H. monophylla* under the microscope because none of the cells are storied, the rays are wider with a single row of upright cells at the margins, and each one often contains a single prismatic crystal. However, *M. excelsa* has been confused with *P. elata* based on morphological macroscopic wood features. InsideWood shows extensive anatomical information on *D. melanoxylon*. As stated before, using InsideWood with the standardized wood anatomical features of *P. elata* may lead to its identification as a *Dalbergia* species. Therefore, using the wood anatomical database alone could lead to an incorrect species identification.

1.3 | Using DART-TOFMS data for species identification

Direct Analysis in Real Time (DART) (see Cody et al[6]) Time-Of-Flight Mass Spectrometry (TOFMS) has shown promise in the analysis of wood and plants. Previous research using DART-TOFMS spectra was able to distinguish between two species of American oak (*Quercus alba* L. and *Quercus rubra* L.),[7] between four species of agarwood (*Aquilaria* spp.),[8,9] and between *Dalbergia* timbers from Africa, Madagascar and Asia.[10] Recent research has focused on the identification of plant species (*Mitragyna speciosa* (Korth.) Havil and *Datura*),[11] discrimination among selected CITES-protected *Araucariaceae*,[12] and differentiating coastal from inland populations of Douglas fir (*Pseudotsuga menziesii* (Mirb.) Franco) using Random Forest classification algorithms.[13] The main goal of this study is to determine if *Pericopsis elata* could be distinguished from the following species using DART-TOFMS: *P. angolensis*, *P. laxiflora*, *P. mooniana*, *M. excelsa*, *H. monophylla*, and *D. melanoxylon*. A second goal was to determine: (1) which classification technique, Kernel Discriminant Analysis (KDA) or Random Forest, performs better to separate these species; (2) if by using the variable (ions) importance lists retrieved from the Random Forest, the KDA could be improved; and (3) the lowest number of ions needed to separate the species.

2 | EXPERIMENTAL

2.1 | Materials

Heartwood samples of all *Pericopsis* species, *M. excelsa*, *H. monophylla*, and *D. melanoxylon*, were provided by different institutions. Table S1 (supporting information) lists the different samples with their geographic provenance, country of origin, and the source and number of specimens.

Species validation of the commercial timber samples was performed by comparing their mass spectra with those of curated xylaria (authenticated wood specimens collection) reference samples.

2.2 | DART TOFMS analysis

The specimens were analysed using a DART-SVP ion source (IonSense, Saugus, MA, USA) coupled to a AccuTOF 4G LC mass spectrometer (Jeol USA, Peabody, MA, USA). Heartwood slivers are placed directly in a stream of heated helium gas, produced by the DART ion source. Spectra were acquired in positive ion mode with the DART ion source parameters and mass spectrometer settings as defined in Evans et al,[12] McClure et al,[10] Lancaster and Espinoza,[8] and Espinoza et al.[9] The spectra were obtained over the mass range of *m/z* 50–700. The text files of the mass-calibrated, centroided mass spectra were exported using TSS Unity (Shrader Software Solutions, Inc., Grosse Pointe Park, MI, USA) data reduction software and used for further analysis.

2.3 | Specimen classification methods

A heatmap, showing the intensity of each ion-mass (*m/z* value) in the specimen (Figure 1), was created using the Mass Mountaineer Mass Spectral Interpretation Tools software (RBC Software, Peabody, MA, USA). Next, KDA was performed with the same Mass Mountaineer software package using a tolerance of 5 mDa and a 1% relative abundance threshold. KDA is a generalization of linear discriminant analysis (LDA) where the principal components are nonlinearly related to the input variables in the transformed space.[14] Each specimen is assigned to a class in the grouping variable (in this case species), and KDA then calculates the maximum separation between species classes in a training set. This is then mapped in the nonlinear higher-dimensional space.[12] KDA determines the species separation based on a subset of appropriate ions. The selected ions are those that are unique to one species, or which show higher intensity in one taxon but lower intensities in others. This process is simplified by a visual inspection of the heatmap. The model accuracy is assessed using leave-one-out-cross-validation (LOOCV, see Lever et al[15] and McClure et al[10]).

The results of the KDA were compared with those from the Random Forest method, which is implemented in the randomForest package[16] in R, which is a free software environment for statistical computing and graphics. All calculations were performed in RStudio (RStudio Team, 2015), an open source software for R. Spectral data were exported from Mass Mountaineer (tolerance of 250 mDa and

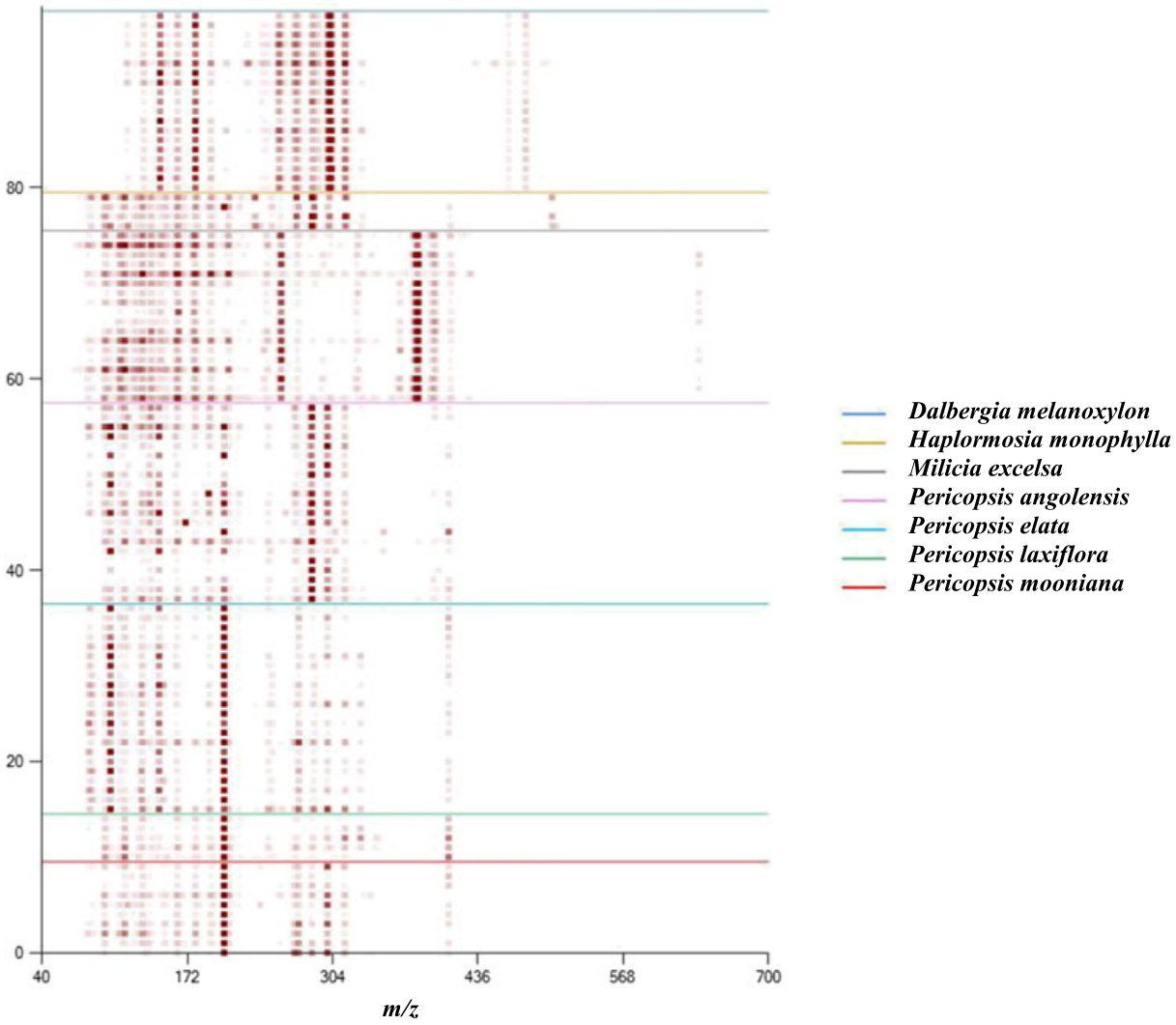

FIGURE 1 Heatmap of the ions present in the analysed specimens. The X-axis is the mass-to-charge ratio (*m/z*) of the molecules detected. The Y-axis indicates specimen number with its chemotype grouped per species. The intensity of the red squares in the heatmap correlates with the abundance of the ions in the specimens [Color figure can be viewed at wileyonlinelibrary.com]

1% threshold) to Microsoft Excel and imported into RStudio. A Random Forest is best described as a set of n regression or classification trees.[17,18] The dataset is randomly split into a training and validation dataset, in this case 80% and 20%, respectively. Each tree is constructed using a different subset of samples of the training dataset with the objective of classifying each sample to a class in the grouping variable (species). Each node in the tree is split using the best predictor variable, here ion relative abundance, among a randomly chosen subset of predictor variables.[16,19] In total, 10,000 classification trees were created to build the Random Forest with 50 randomly chosen ions at every node split. Model accuracy is determined by the out-of-the bag (OOB) principle. At each bootstrap iteration the samples that were not used in the training set are used to validate the current tree in that bootstrap iteration.[16] The overall OOB accuracy is reported as the estimation of the error rate, indicating the misclassification of samples. Instead of using the OOB classification error, we report the complement, or the Random Forest classification accuracy, to compare with the validation rate of LOOCV in KDA (OOB error + Random Forest classification accuracy = 1).[13] Before determining the performance of the Random Forest through the validation dataset, the classification error of the samples per species in the training set is given. This is a first indication of which species will be problematic. Finally, the performance of the Random Forest classification is determined using the validation dataset to test the model. Several measures for variable importance can also be assessed, which in this case indicate specific ions that are key for differentiating among species. The first measure is the Gini-index or Mean Decrease in Impurity (MDI), which is used to quantify the impurity in each node.[18] A second measure, based on permutation of the OOB data, is the Mean Decrease in Accuracy (MDA) and aims at improving the accuracy. The difference in prediction accuracy is a good indicator of variable importance.[18] A comprehensive review of MDI vs MDA can be found in Perrier's "Feature Importance in Random Forest".[20]

The lists of ions, ranked by variable-importance, were then used for KDA to determine if the Random Forest-generated ions give a higher classification accuracy than the empirical ion selection described above. Experiments were conducted with different numbers of ions (5, 10, 20, 30, 50, 100, 200, and 256) based on the importance values from the Random Forest.

2.4 | Model comparison

The KDA results were compared with those from the Random Forest classification under two different conditions. The KDA of the *Pericopsis* and the lookalike species was based on 65 ions. This experiment excluded *H. monophylla* because of the small sample size. The second KDA consisted of only *Pericopsis* taxa, and 248 ions were used for the classification. The species used and their respective sample sizes are listed in Table 1.

TABLE 1 Total sample number per species group

Group	Species	Samples
Pericopsis elata	*Pericopsis elata*	22
Pericopsis (others)	*Pericopsis angolensis*	21
	Pericopsis laxiflora	5
	Pericopsis mooniana	10
Dalbergia melanoxylon	*Dalbergia melanoxylon*	19
Milicia excelsa	*Milicia excelsa*	18
Sum		95

3 | RESULTS AND DISCUSSION

Figure 1 shows the heatmap for the different species analysed. The chemotype of *H. monophylla* is present in the heat map but was removed from the classification models due to the small sample size. However, this species could be separated from the other species based on visual inspection of the heatmap. As shown in Figure 1, the chemotypes for *H. monophylla*, *M. excelsa*, and *D. melanoxylon* are different from those of the *Pericopsis* species. In addition, some differences appear among the chemotypes of the *Pericopsis* taxa. For example, *P. laxiflora* seems to have more intense ions around *m/z* 409. *P. angolensis* can be differentiated by looking at the ions around *m/z* 285 and 299 and a lack of consistent intensity around *m/z* 205. The surprising find that the chemotype of *P. laxiflora* is dissimilar to the chemotype of *P. angolensis* erodes the support for the hypothesis that *P. laxiflora* could be a subspecies of *P. angolensis*.[2] However, a larger sample size of *P. laxiflora* is needed to statistically test this observation. Only *P. mooniana* has ions at *m/z* 260 and 273, and these are absent in the other *Pericopsis* species. Figures 2 and 3 show the KDA graphical representation of two different datasets. Figure 2 shows that the KDA classification algorithms clustered each of the *Pericopsis* species separately, whereas Figure 3 shows the separation of the protected *P. elata* from the other species.

The main goal of this study was to separate the CITES-listed *P. elata* from the other species in its genus and from its lookalikes. The LOOCV (KDA) was 95.79%, and the classification accuracy of the Random Forest was 96.05%, indicating that both KDA, with the empirically chosen ions, and Random Forest enabled us to correctly identify *P. elata* to a satisfactory level. Table 2 shows the confusion matrix, which summarizes the classification of the training dataset for the Random Forest. As can be seen, *P. elata* shows the lowest classification accuracy, but it was still high. Only two of the 16 samples are misclassified. Afterwards, the Random Forest is validated using the prediction data. Table 3 shows the results for the classification of the prediction data. In this example, the Random Forest classified all samples correctly. Next, we tried to differentiate between *Pericopsis* species. The LOOCV was 88.89% and the Random Forest accuracy was 93.75%. These results are, however, based on an unbalanced dataset, with only five samples for *P. laxiflora* compared with, for example, the 21 samples from *P. angolensis*. Although the final model performance is satisfactory, it might affect the model variability and the handling of misclassifications.[13] The overall classification accuracy is not significantly affected; however, this is often not an appropriate performance measure in learning extremely unbalanced data.[21] Using these small unbalanced datasets increases the risk of leaving a certain species out of the training dataset bootstrapping, skewing the model towards the more abundant species. Possible solutions are suggested

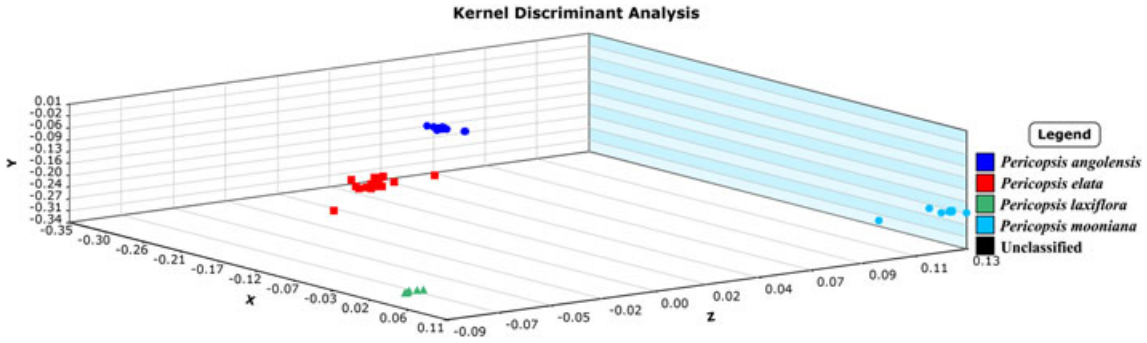

FIGURE 2 Graphical representation of the Kernel Discriminant Analysis of the four *Pericopsis* species, showing that the species segregate distinctly (LOOCV is 88.89%) [Color figure can be viewed at wileyonlinelibrary.com]

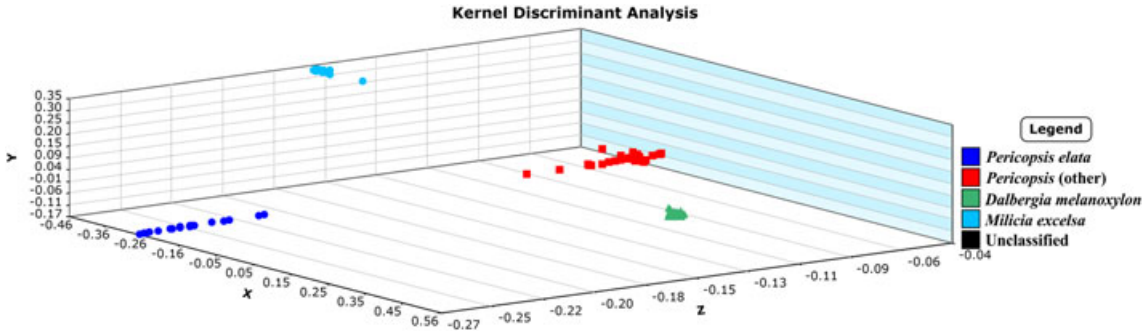

FIGURE 3 Graphical representation of the Kernel Discriminant Analysis showing that *Pericopsis elata* can be distinguished from *Pericopsis* (other), *Dalbergia melanoxylon* and *Milicia excelsa* (LOOCV is 95.79%) [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 2 Confusion matrix from the Random Forest of dataset 1 (all the species) based on the training set

	Dalbergia melanoxylon	*Milicia excelsa*	*Pericopsis* (other)	*Pericopsis elata*	Classification accuracy
Dalbergia melanoxylon	16				100.00
Milicia excelsa		13	1		92.86
Pericopsis (other)			30		100.00
Pericopsis elata			2	14	87.50

Note that the classification accuracy is shown and not the classification error

TABLE 3 Classification of the prediction set for dataset 1 (all the species)

	Dalbergia melanoxylon	*Milicia excelsa*	*Pericopsis* (other)	*Pericopsis elata*
Dalbergia melanoxylon	3			
Milicia excelsa		4		
Pericopsis (other)			6	
Pericopsis elata				6

by Chen et al.[21] This was, however, outside the scope of the current study and should be investigated further.

Supplementary goals of this study were to determine if empirically selected variables (ions) provided the same level of accuracy as the ions selected by the Random Forest algorithm and the minimum number of ions required to obtain the highest classification accuracy. Random Forest provides two ways of determining variable (i.e., ion) importance, the MDI and MDA. The Random Forest analysis produces a ranked list of variables (ions) that have the highest importance in separating classes, and, in this case, the most valuable ions for separating *P. elata* from the other lookalike species. This ranked list of ions was used to perform KDA, and the results were compared with those from the ions selected by Random Forest. Table 4 compares the accuracy of the two approaches using the calculated LOOCV. Table 4 also shows the results when different numbers of ions were used and the resulting LOOCV based on the most important ions according to MDI and MDA (ranging from 5 to 256 ions). It is clear that the highest LOOCV accuracy, for the lowest number of ions, was obtained with the 50 most important ions selected by the MDI algorithm of Random Forest.

TABLE 4 The LOOCV for the KDA performed using manually chosen ions, and based on the most important ions according to MDI and MDA, and the OOB estimation of error rate from the Random Forest for dataset 1 (all the species)

%	MDI	MDA
LOOCV – 5	70.53	70.53
LOOCV – 10	76.84	76.84
LOOCV – 20	94.74	94.74
LOOCV – 30	92.63	94.74
LOOCV – 40	94.74	93.68
LOOCV – 50	95.79	92.63
LOOCV – 100	94.74	95.79
LOOCV – 200	95.79	95.79
LOOCV – 256	95.79	95.79
LOOCV – manual	95.79	
OOB	3.95	
1 – OOB	96.05	

4 | CONCLUSIONS

We have demonstrated that the identification of *P. elata* can be accomplished using DART-TOFMS spectra. Although the heatmap of the *Pericopsis* taxa appears to be similar, the statistical post-processing of the spectra can be used to identify species with high accuracy. The chemotypes shown in the heatmap (Figure 1) of *M. excelsa*, *D. melanoxylon*, and *H. monophylla* are very different from those of the four other *Pericopsis* species, and minor differences in the chemotypes of the *Pericopsis* taxa can also be observed. The chemotype of *P. laxiflora* is dissimilar to the chemotype of *P. angolensis*. This observation does not support the hypothesis that *P. laxiflora* could be a subspecies of *P. angolensis*,[2] and a larger sample size of *P. laxiflora* will permit statistical testing of this.

Taxa classifications using KDA and Random Forest algorithms have similar and satisfactory success rates, showing that both methods can be used to determine the species identity. A recurring challenge when performing KDA is to determine which variables to use, as suboptimal choices may lead to overfitting and bias through arbitrary selection of discriminant ions. However, Random Forest algorithms are based on dataset training and multiple bootstrap iterations, rather than analyst judgment, and do incorporate all ions detected among samples. The results from the Random Forest classification analysis are therefore more objective. It was, however, interesting to note that KDA through manual selection of ions provides similar classification accuracy to Random Forest. We show that the variable importance measures included in the Random Forest can, however, aid in the ion choice for KDA. For this case, we observed that only 50 ions were needed to achieve the best accuracy. We conclude that, in addition to wood anatomy, timber samples that have questionable origin can be analyzed by DART-TOFMS and the resulting spectra can be evaluated using either KDA (LOOCV 95.79%) or Random Forest (96.05%).

Ultimately, DART-TOFMS and post-processing analysis of the spectra provide robust identifications of timbers when traditional wood anatomical methods are indecisive, or if the required expertise is unavailable, and these tools provide an additional approach for combatting illegal timber trade. The success of this method naturally depends on the availability of samples to ensure balanced datasets. The authors would like to take this opportunity to invite xylaria throughout the world to share their vast collections.

ACKNOWLEDGEMENTS

This research was conducted under the HerbaXylaRedd belspo-project (Brain..be – code: BR/143/A3/HERBAXYLAREDD). The authors would like to thank Stijn Willen (Laboratory for Wood Technology, UGent), Gabriela D. Chavarria, Erin McClure-Price and Pamela J. McClure (National Fish & Wildlife Forensic Lab, Oregon USA) for their help with the sample preparation and thank Bonnie Yates for the careful editing of the manuscript. The findings and conclusions in the article are those of the authors and do not necessarily represent the views of the U.S. Fish and Wildlife Service.

ORCID

V. Deklerck http://orcid.org/0000-0003-4880-5943

REFERENCES

1. Convention on International Trade in Endangered Species of Wild Fauna and Flora. Appendices I, II and III. Available: https://cites.org/sites/default/files/notif/E-Notif-2016-068-A.pdf (accessed December 12, 2016).
2. Prota 7(1): Timbers. In: Louppe D, Oteng-Amoaki AA, Brink M, eds. *Plant Resources of Tropical Africa*. Wageningen: Backhuys Publishers; 2008.
3. PROSEA 5(1): Plant resources of South-East Asia 5. *Pericopsis mooniana*. In: Soerianegra I, Lemmens RHMJ, eds. *Timber Trees: Major Commercial Timbers*. Wageningen: Pudoc Scientific Publishers; 1993:342-345.
4. InsideWood (2004 – onwards). Available: http://insidewood.lib.ncsu.edu/search (accessed November 25, 2016).
5. Wheeler EA, Baas P, Gasson E. IAWA list of microscopic features for hardwood identification. *IAWA Bull.* 1989;10:219-332.
6. Cody RB, Laramee JA, Durst HD. Versatile new ion source for the analysis of materials in open air under ambient conditions. *Anal Chem.* 2005;77:2297-2302.
7. Cody RB, Dane AJ, Dawson-Andoh B, Adepipe EO, Nkansah K. Rapid classification of white oak (*Quercus alba*) and northern red oak (*Quercus rubra*) by using pyrolysis direct analysis in real time (DART™) and time-of-flight mass spectrometry. *J Anal Appl Pyrol.* 2012;95:134-137.
8. Lancaster C, Espinoza E. Evaluating agarwood products for 2-(2-phenylethyl) chromones using direct analysis in real time time-of-flight mass spectrometry. *Rapid Commun Mass Spectrom.* 2012;26:2649-2656.
9. Espinoza EO, Lancaster CA, Kreitals NM, Hata M, Cody RB, Blanchette RA. Distinguishing wild from cultivated agarwood (*Aquilaria* spp.) using direct analysis in real time (DART™) and time-of-flight mass spectrometry. *Rapid Commun Mass Spectrom.* 2014;28:281-289.
10. McClure PJ, Chavarria GD, Espinoza E. Metabolic chemotypes of CITES protected *Dalbergia* timbers from Africa, Madagascar, and Asia. *Rapid Commun Mass Spectrom.* 2015;29(9):783-788.
11. Lesiak AD, Musah RA. Rapid high-throughput species identification of botanical material using Direct Analysis in Real Time high resolution mass spectrometry. *J Vis Exp.* 2016;116:1-11. e54197.
12. Evans PD, Mundo IA, Wiemann MC, et al. Identification of selected CITES-protected Araucariaceae using DART TOFMS. *IAWA J.* 2017;38(2):266-281. https://doi.org/10.1163/22941932-20170171.

13. Finch K, Espinoza E, Jones FA, Cronn R. Source identification of western Oregon Douglas-fir wood cores using mass spectrometry and random forest classification. *Appl Plant Sci*. 2017;5(5):1-12. https://doi.org/10.3732/apps.1600158.
14. Baudat G, Anouar F. Generalized discriminant analysis using a kernel approach. *Neural Comput*. 2000;12:2385-2404.
15. Lever J, Krzywinski M, Althman N. Model selection and overfitting. *Nat Methods*. 2016;13(9):703-704.
16. Liaw A, Wiener M. Classification and regression by randomForest. *R News*. 2002;2(3):18-22.
17. Breiman L, Friedman J, Stone CJ, Olshen RA. *Classification and regression trees*. Chapman & Hall; 1984.
18. Strobl C, Malley J, Tutz G. An introduction to recursive paritioning: rationale, application, and characteristics of classification and regression trees, bagging and random forests. *Psychol Methods*. 2009;14(4):323-348.
19. Breiman L. Random Forests. *Machine Learning*. 2001;45(1):5-32.
20. Perrier A. Feature Importance Random Forest; 2015. Available: http://alexperrier.github.io/jekyll/update/2015/08/27/feature-importance-random-forests-gini-accuracy.html.
21. Chen C, Liaw A, Breiman L. *Using random forest to learn imbalanced data*. Berkeley: University of California; 2004:1-12 Available: http://statistics.berkeley.edu/sites/default/files/tech-reports/666.pdf.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Deklerck V, Finch K, Gasson P, et al. 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 Commun Mass Spectrom*. 2017;31:1582–1588. https://org/10.1002/rcm.7939