Overview of current practices in data analysis for wood identification. A guide for the different timber tracking methods

Nele Schmitz, Hans Beeckman, Celine Blanc-Jolivet, Laura Boeschoten, Jez W.B. Braga, José Antonio Cabezas, Gilles Chaix, Simon Crameri, Victor Deklerck, Bernd Degen, et al.

GTTN
Global Timber Tracking Network

Overview of current practices in data analysis for wood identification

A guide for the different timber tracking methods

June 2020

GTTN
Global Timber
Tracking Network

Overview of current practices in data analysis for wood identification

A guide for the different timber tracking methods

June 2020

Editor:
Nele Schmitz

Authors*:
Hans Beeckman[1], Céline Blanc-Jolivet[2], Laura Boeschoten[3], Jez W.B. Braga[4], José Antonio Cabezas[5], Gilles Chaix[6,7,8], Simon Crameri[9], Bernd Degen[2], Victor Deklerck[1,10], Eleanor Dormontt[11], Edgard Espinoza[12], Peter Gasson[13], Volker Haag[14], Stephanie Helmling[14], Micha Horacek[15,16], Gerald Koch[14], Cady Lancaster[12], Frederic Lens[17], Andrew Lowe[11], Sandra Martínez-Jarquín[18], Justyna Anna Nowakowska[19], Andrea Olbrich[14], Kathelyn Paredes-Villanueva[20], Tereza C.M. Pastore[21], Tahiana Ramananantoandro[22], Andriambelo R. Razafimahatratra[22], Prabu Ravindran[23,24], Gareth Rees[25], Liz F. Soares[4], Niklas Tysklind[26], Mart Vlam[3], Charlie Watkinson[25], Elisabeth Wheeler[27,28], Robert Winkler[18], Alex C. Wiedenhoeft[23,24,29,30], Valentina Th. Zemke[14], Pieter Zuidema[3].

*Names are listed in alphabetical order.

Recommended citation:
Schmitz, N. (ed.), H. Beeckman, C. Blanc-Jolivet, L. Boeschoten, J.W.B. Braga, J.A. Cabezas, G. Chaix, S. Crameri, B. Degen, V. Deklerck, E. Dormontt, E. Espinoza, P. Gasson, V. Haag, S. Helmling, M. Horacek, G. Koch, C. Lancaster, F. Lens, A. Lowe, S. Martínez-Jarquín, J.A. Nowakowska, A. Olbrich, K. Paredes-Villanueva, T.C.M. Pastore, T. Ramananantoandro, A.R. Razafimahatratra, P. Ravindran, G. Rees, L.F. Soares, N. Tysklind, M. Vlam, C. Watkinson, E. Wheeler, R. Winkler, A.C. Wiedenhoeft, V.Th. Zemke, P. Zuidema (2020). Overview of current practices in data analysis for wood identification. A guide for the different timber tracking methods. Global Timber Tracking Network, GTTN secretariat, European Forest Institute and Thünen Institute. DOI: 10.13140/RG.2.2.21518.79689

Cover: José Bolaños
Project leader: Jo Van Brusselen
Pictures: Unless mentioned otherwise, all pictures in this document were provided by the authors.

The following document is an output from GTTN, which is coordinated by the European Forest Institute (EFI) and funded by the German Federal Ministry for Food and Agriculture (BMEL). The information expressed in this document are the views of the authors and do not necessarily represent those of the donor or of the European Forest Institute.

Affiliations:

1-Royal Museum for Central Africa, Tervuren, Belgium

2-Thünen Institute of Forest Genetics, Grosshansdorf, Germany

3-Wageningen University & Research, Wageningen, The Netherlands

4-University of Brasília, Brasília, Brazil

5-Forest Research Centre, National Institute for Agricultural and Food Research and Technology (INIA-CIFOR), Madrid, Spain

6-CIRAD, UMR AGAP, Montpellier, France

7-AGAP, Univ Montpellier, CIRAD, INRA, Montpellier SupAgro, Montpellier, France

8-ESALQ-USP, Wood Anatomy & Tree-Ring Lab, Piracicaba, Brazil

9-Institute of Integrative Biology, ETH Zurich, Zurich, Switzerland

10-Ghent University Woodlab, Ghent, Belgium

11-University of Adelaide, Adelaide, Australia

12-U.S. Fish & Wildlife Service, Ashland, USA

13-Royal Botanic Gardens, Kew, UK

14-Thünen Institute of Wood Research, Bergedorf, Germany

15-HBLFA Francisco-Josephinum, Wieselburg, Austria

16-Institute of Lithospheric Research, Vienna University, Vienna, Austria

17-Naturalis Biodiversity Center, Leiden, The Netherlands

18-Department of Biochemistry and Biotechnology, CINVESTAV Unidad Irapuato, Irapuato, Mexico

19-Cardinal Stefan Wyszynski University, Warsaw, Poland

20-Gabriel René Moreno University, Santa Cruz, Bolivia

21-Brazilian Forest Service, Brasília, Brazil

22-School of Agronomy, Department of Water and Forestry, University of Antananarivo, Antananarivo, Madagascar

23-USDA Forest Service, Madison, USA

24-University of Wisconsin, Madison, USA

25-Agroisolab UK Ltd, York, UK

26-National Institute of Agricultural Research (INRA), Kourou, Guyane Française

27-North Carolina State University, Raleigh, USA

28-North Carolina Museum of Natural Sciences, Raleigh, USA

29-Purdue University, West Lafayette, USA

30-São Paulo State University, São Paulo, Brazil

For more information about the authors see the GTTN project partner finder.

Table of Contents

List of Boxes

List of Tables

List of Figures

Rationale

Today we have five types of **timber tracking tools** available. Each has its own **strengths and limitations** (see the Timber Tracking Tool Infogram), but together they offer a broad range of methods that can assist us in identifying the botanical as well as the geographic origin (provenance) of most kinds of timber samples, even those smaller than 1 cm^3.

With this guide we want to provide an **overview of the current best-practice methods** used to analyse data derived from different wood identification methods, while presenting their respective strengths and limitations. We give advice on data analysis, from the development of reference data, through to the verification of identity and provenance of unknown samples against the reference database. We end with an **expert view on combining methods for wood identification** and discuss how timber identification possibilities could expand in the future.

The guide is meant for all researchers doing identification work to **support regular law enforcement measures, initiatives for transparent supply chains, and continued scientific research.** Undertaking forensic casework is qualitatively different from undertaking scientific research (UNODC 2016)[1]. To support court cases, the chain of custody of evidence material must be tracked and reference material must be valid. How to fulfil all requirements for forensic timber identification will not be discussed here. Instead, we refer to the best practice guide of the UNODC (2016) and the standards and guidelines of the Society for Wildlife Forensic Sciences.

We hope that this guide will be a first step towards the **harmonisation of data analysis procedures** within the global network of wood identification experts; and will facilitate collaborations, further innovations, and ensure reliable timber tracking methods.

DISCLAIMER: The recommendations in this guide are based on current research and the state of the science at the time of writing and will evolve.

REFERENCES:

UNODC (2016). Best Practice Guide for Forensic Timber Identification. New York, United Nations.

[1] For example, validation tests that are derived directly from published academic (non-forensic) literature are usually insufficient to derive legally compliant evidence (UNODC, 2016).

Abbreviations & Terminology

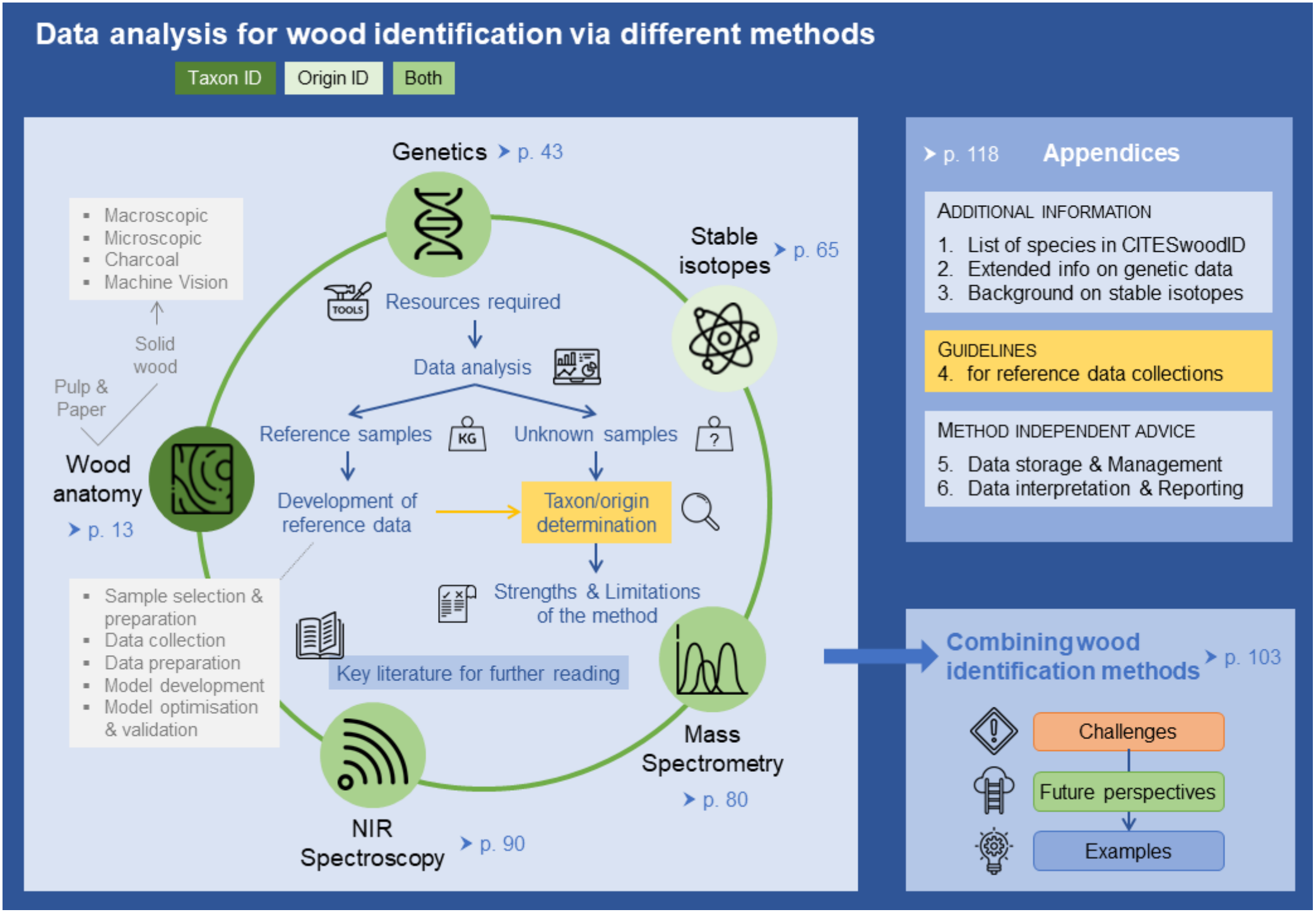
TERMINOLOGY

Provenance and geographic origin are used interchangeably in this guide as there is a preference for one word or the other depending on the scientific field, as well as the information source, but they have the same meaning.

AMU	Atomic Mass Unit
CA	Correspondence Analysis
CNN	Convolutional neural networks
cpDNA	chloroplast DNA
DAPC	Discriminant Analysis of Principal Components
DART TOFMS	Direct Analysis in Real Time (DART) Time-of-Flight Mass Spectrometry (TOF MS)
DL	Deep learning
FESEM	Field Emission Scanning Electron Microscope
HTS	High Throughput Sequencing
HWE	Hardy Weinberg Equilibrium
ID	Identification
KDA	Kernel Discriminant Analysis
LD	Linkage disequilibrium
LE	Linkage equilibrium
LDA	Linear Discriminant Analysis
LOOCV	Leave-one-out-cross-validation
LV	Latent Variables
MAF	Minor allele frequency
mtDNA	mitochondrial DNA
MV	Machine Vision

MVA	Multivariate Analysis
NFWFL	National Fish and Wildlife Forensic Laboratory
NIR	Near Infrared
NIRS	Near Infrared Spectroscopy
PARAFAC	Parallel Factor Analysis
PC	Principal Components
PCA	Principal Components Analysis
PLS-DA	Partial Least Squares for Discriminant Analysis
PLSR	Partial Least Squares Regression
3D-RLM	3D-Reflected Light Microscope
SEM	Scanning Electron Microscope
SIMCA	Soft Independent Modelling of Class Analogy
SNP	Single Nucleotide Polymorphism
SSR	Simple Sequence Repeats

Visual summary



1. Wood anatomy

Definition wood anatomical reference data: descriptions and/or illustrations of the macroscopic and/or microscopic features of the wood, preferably covering many samples from all major woody lineages.

Authors: Hans Beeckman, Peter Gasson, Volker Haag, Stephanie Helmling, Gerald Koch, Frederic Lens, Andrea Olbrich, Prabu Ravindran, Elisabeth Wheeler, Alex C. Wiedenhoeft, Valentina Th. Zemke

*Authors are in alphabetical order.

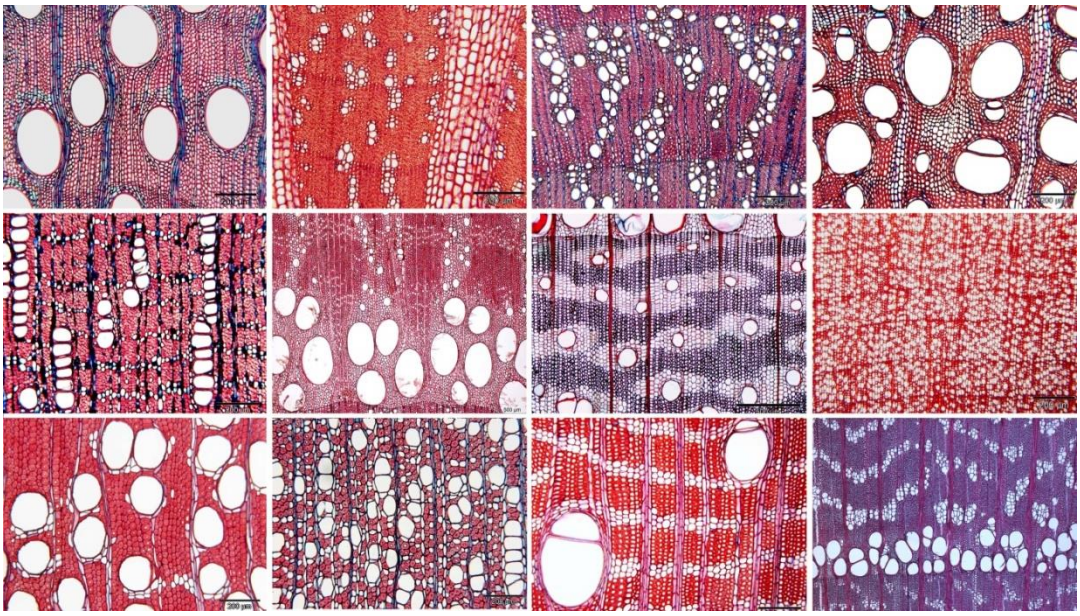


Fig. 1: Microsections showing the variety of the anatomy of hardwoods

1.1 RESOURCES REQUIRED FOR WOOD ANATOMICAL ANALYSIS

Access to reference material

The data for a wood anatomical analysis are the macroscopic and/or microscopic features of the wood. To analyse these data, they are compared to reference material. Therefore, access is needed to **physical reference collections**, ideally associated with herbarium vouchers, and/or to printed wood anatomical atlases (see Box 5) or electronic databases with/out interactive identification key (Table 1).

Table 1: Overview of digital reference databases used for the analysis of wood anatomical characteristics, for the purpose of wood identification.

Reference collection	Performance	Accessibility
InsideWood	Over 7000 descriptions of the microscopic anatomy of hardwoods. Worldwide in coverage, multiple-access identification key using the IAWA hardwood list of features[2]. Includes a searchable image database of over 45000 images.	Freely available on the internet
Commercial timbers	An interactive identification key with integrated database of the microscopic descriptions and illustrations for 404 hardwood taxa common in the international trade that occur in all major forest regions of the world.	Richter & Dallwitz (2000 onwards). Freely available on the internet
CITESwoodID	An interactive identification key with integrated database of macroscopic descriptions and illustrations for 44 CITES protected timbers (for the list of species see Appendix 1) and 34 look-a-like trade timbers.	Richter *et al.* (2005 onwards). Freely available on the internet
MacroHOLZdata	An interactive identification key with integrated database of macroscopic descriptions and illustrations for 130 commercial timbers/taxa (113 hardwoods, 15 softwoods, bamboo, and red palm).	Richter *et al.* (2002 onwards). Free of charge on request.

[2] IAWA Committee (1989)

Table 1 continued

Reference collection	Performance	Accessibility
Softwoods	An interactive identification key with integrated database of microscopic descriptions and illustrations for 53 coniferous taxa that occur in all major forest regions of the world.	Richter & Dallwitz (2016 onwards). Free of charge on request.
Microscopic Wood Anatomy of Central European species	Web-based identification key, which is a completely revised and updated version of Schweingruber (1990).	Freely available on the internet
Wood database of the Forestry & Forest Products Research Institute	An identification system of Japanese woods based on the IAWA list of hardwood identification, including a multiple entry key, and an image database.	Freely available on the internet
The Forest Species Database (FSD)	A database for automatic classification of forest species, composed of 2240 microscopic images from 112 different species from two groups (Hardwood and Softwood), 85 genera and 30 families.	Freely available on the internet
The Forest Species Database – Macroscopic (FSD-M)	A database for automatic classification of forest species, composed 2942 macroscopic images from 41 different forest species of the Brazilian flora.	Freely available on the internet

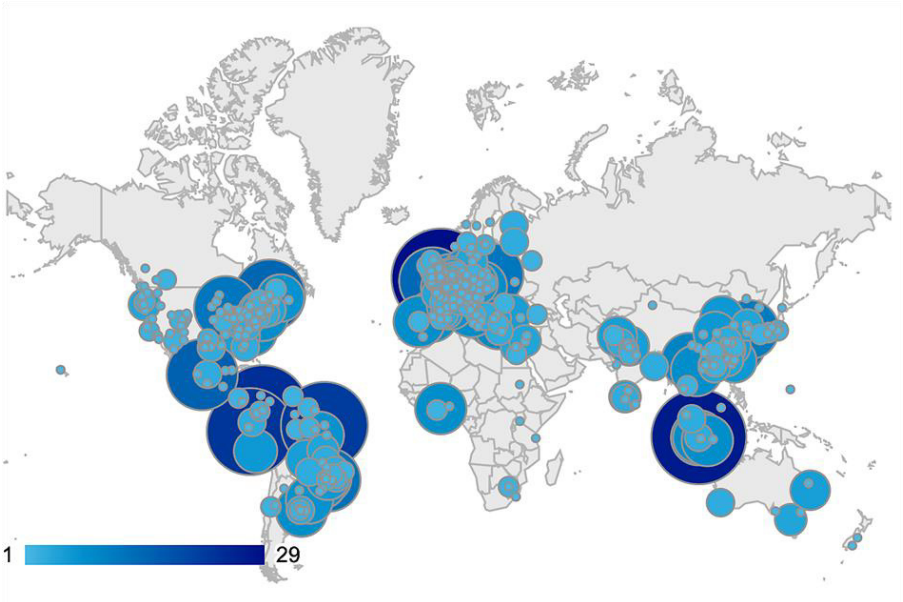


Fig. 2: The success story of the *InsideWood* database. The map shows the origin of the website visitors in March 2019. The size of the bubble and the intensity of its colour reflect the number of visitors from that location.

1.2 DATA ANALYSIS FOR TAXON IDENTIFICATION OF SOLID WOOD

1.2.1 WOOD ANATOMICAL ANALYSIS OF REFERENCE SAMPLES

1.2.1.1 DEVELOPMENT OF MACROSCOPIC REFERENCE DATA

Creating features lists

Macroscopic wood identification is based on observations in the three anatomical planes, as for microscopic wood identification. However, observations are made, with the unaided eye or with the help of a magnifying lens (recommended magnification 10-14x), after the transverse plane of the specimen is surfaced by sanding or with a utility knife[3]. The **transverse plane** offers the most useful diagnostic information about type, distribution, and arrangement of the wood cells as well as about growth-ring characteristics. However, in some taxa also the **longitudinal planes** can reveal key features (*e.g.* presence of a storied structure, size and composition of rays), especially the tangential plane.

Ruffinato *et al.* (2015) proposed a **list of features** for macroscopic wood identification.

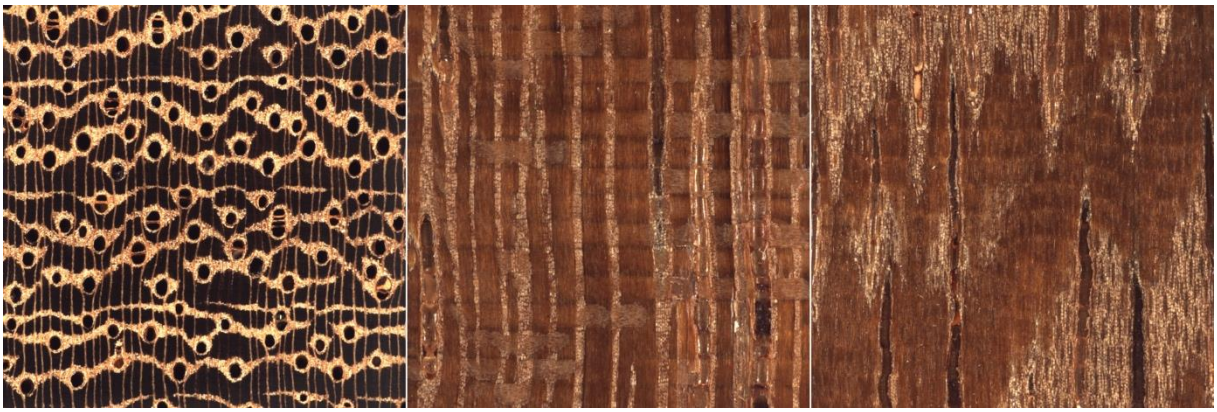


Fig. 3: XyloTron macrographs of *Martiodendron* sp. Transverse (left), radial (centre), and tangential (right) surfaces. Each image is 6,35mm x 6,35mm. Images courtesy of Richard Soares, Center for Wood Anatomy Research, Forest Products Laboratory, USA.

[3] A utility or craft knife with (high quality) disposable blades is recommended for end grain surfacing. The blade should always be moved in a sliding motion to produce clean cuts. With exceptionally soft woods a flexible razor blade produces better surfaces. Extremely hard woods often require several overlapping strokes with the knife to produce a clean surface large enough for initial evaluation of macroscopic features. Change blades frequently because the inevitable nicks will leave traces on the cut surface, which may obscure details of wood structure.

Developing reference data

See the checklist in Appendix 4 on how to ideally develop macro- and microscopic reference data.

1.2.1.2 DEVELOPMENT OF MICROSCOPIC REFERENCE DATA

Creating features lists

Microscopic wood identification is based on observations of a wood specimen in the **three anatomical planes**. The international standard to develop wood anatomical reference data for identification is to start from the **list of features** from the International Association of Wood Anatomists (**IAWA Committee** 1989), which is the basis for all species descriptions in the *InsideWood* Database, as well as for most databases based on microscopic features of angiosperms (see Table 1). A second list contains features useful for softwood identification (IAWA Committee 2004) and is used for the species descriptions in the *Softwoods* database (Richter & Dallwitz 2016 onwards).

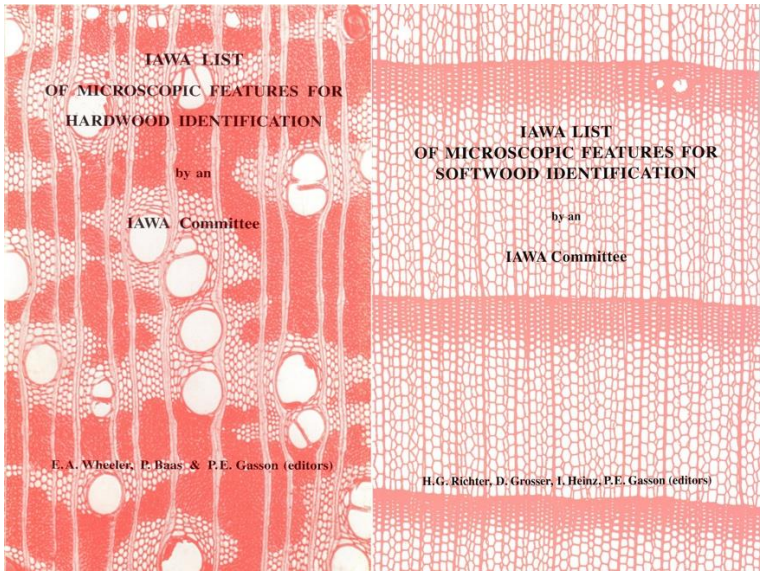


Fig. 4: Cover pages of the international standard guides used to develop wood anatomical reference data for identification of hardwood and softwood.

Most databases adhere to the IAWA lists of microscopic features for hardwood or softwood identification. However, these are not complete lists of features. One cannot, and should not expect to, simultaneously simplify, and fully capture the variability of wood using the IAWA lists. While they are a good starting point, the **IAWA lists should be modified to the scope of the reference database** (Table 1),

as the diagnostic power of anatomical characters varies between certain groups of woods.

Developing reference data

These features lists are then used to develop reference data and to describe the species in wood anatomical terms. For a checklist on how to build **the ideal reference database** for wood identification see Appendix 4.

BOX 1: USEFUL REFERENCES FOR SCIENTIFIC, COMMON & TRADE NAMES

When building reference databases it is essential to categorise your data under the correct name. However, what the correct name is, is not always so clear. Below are some useful references to help you determine the correct scientific name and how it is connected to common and trade names. In case of any doubt, it is highly recommended to seek the advice of a plant taxonomist.

- Plants of the World online
- World Flora Online
- The International Plant Names Index
- Catalogue of life
- Germplasm Resources Information Network (also contains common names)
- Common Names Database of World Timbers
- Common and trade names of CITES listed tree species
- Common & trade names of commonly used species & their alternatives
- Tropicos
- Brazilian Flora 2020 project

1.2.1.3 DEVELOPMENT OF CHARCOAL REFERENCE DATA

Creating features lists

Wood species in charcoal can commonly be identified using the **IAWA features list** (IAWA Committee 1989). The use of this list, however, is limited by the fact that carbonization leads to dimensional changes in the structure of the material (often in

the range of 10-20% shrinkage). For example, Bodin *et al.* (2019) used 26 anatomical features and 112 feature states in their *CharKey* database for the identification of wood charcoals of French Guiana. Only the **visible qualitative anatomical characters** can be used. **Quantitative features**, such as vessel diameter and the width and height of rays (in dimensions, the number of cells remains the same), can only be used to a limited extent. These measured values in charcoal are significantly different than those measured in solid wood. Information on technical values, chemical tests, fluorescence, and extract colour are unavailable for charcoal.

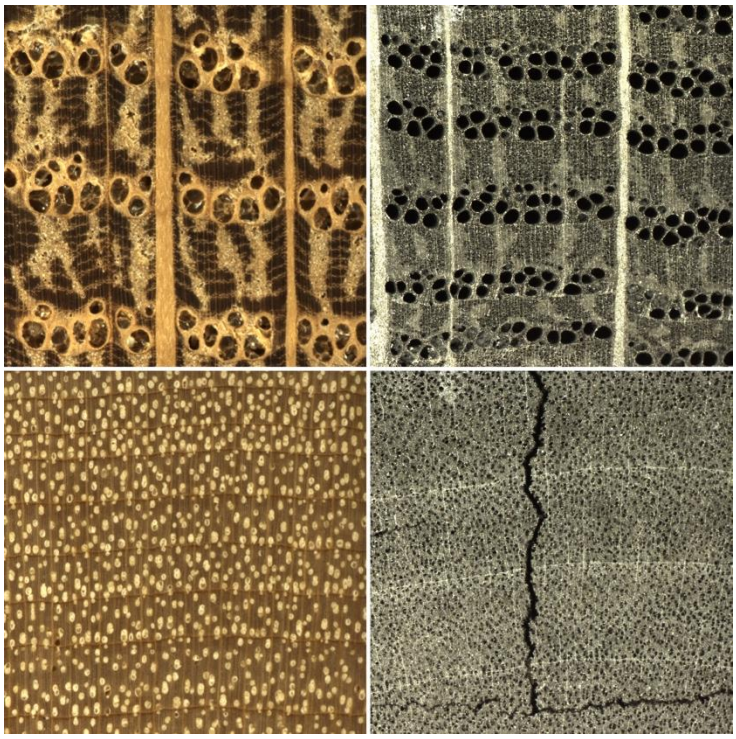


Fig. 5: Macrographs of wood before (left) and after (right) carbonisation, illustrating the changed structural dimensions in charcoal.

Developing reference images

To develop reference images for charcoal it is advisable to **carbonize verified or vouchered wood samples**. The wood samples have to be wrapped in aluminium foil and carbonized with a maximum temperature of 400 °C in a muffle furnace. Ideally, more than one sample is burnt, each of a **minimal sample size** of 25x25x15 mm^3. This would allow international sharing of reference samples between anthracologists.

Microscopic images of charcoal may vary in quality depending on the condition of the charcoal and the choice of technical equipment (Hubau *et al.* 2013). To produce reference images, it is advised to use dry solid charcoal with a wood structure that is not porous or melted and showing as few cracks as possible. Various high

magnification microscopes can be used (*e.g.* 3D-reflected light microscope, scanning electron microscope, field emission scanning electron microscope and X-ray computed tomography) but a 3D-reflected microscope using polarised light is ideally suited for charcoal identification (Zemke *et al.* 2020). This equipment allows a significantly improved visualization of anatomical structure in carbonized wood compared to the classical light microscope.

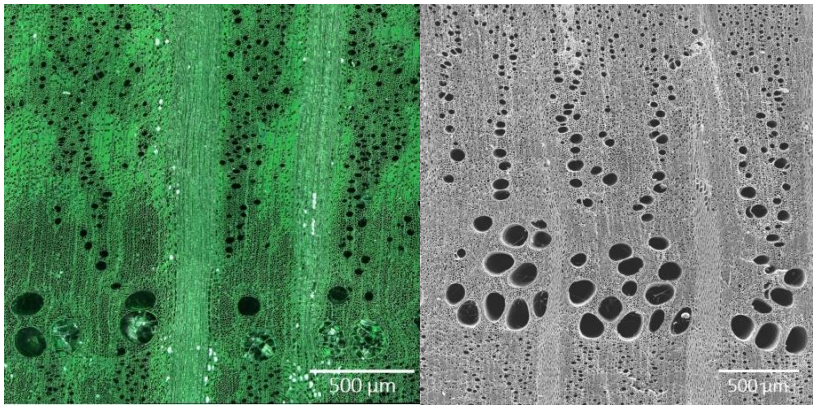


Fig. 6: Micrographs of the transverse surface of *Quercus* sp. made with a 3D-reflected light microscope (left) and a scanning electron microscope (right).

BOX 2: THE VALUE OF VOUCHERED *VS.* VERIFIED 'REAL-LIFE' SPECIMENS

Correct identification of reference samples is crucial, irrespective of whether the samples are linked to herbarium vouchers or not. The advantage of **vouchered specimens**, however, is the extra information that can be used by a plant taxonomist in support of species identification[4].

For identification requests that do not require species-level resolution, correctly identified **real-life specimens** can be more informative than vouchered reference samples, which can be of different taxa, provenances or product types than the ones regularly occurring in identification requests.

Conclusion, for species-resolution wood identification methods and methods that require fresh samples for the development of reference data, vouchered reference samples are the way to go. In addition, reference databases should be supplied with data from samples as they occur in identification requests.

[4] Specimens are more commonly misidentified at the species level than at higher taxonomic levels.

1.2.1.4 DEVELOPMENT OF REFERENCE DATA FOR MACHINE VISION (MV)

Sample selection and preparation

Field screening MV models are equivalent to trained field agents evaluating shipments. By and large such agents are not trained on vouchered wood specimens, but instead learn from commercial specimens of the woods in question. While correctly identified vouchered specimens are invaluable, their number is limited and slow to grow. More **critical for scaling MV technologies is access to wood anatomy/identification expertise** to assess the identification of non-vouchered specimens. This allows the use of non-vouchered specimens evaluated for shipment and permits building image datasets of sufficient size to **capture the commercially relevant biological variability** of wood anatomical features (see also Appendix 4). Another effective but logistically challenging option to rectify deficiencies in existing collections is via targeted field collections. See the GTTN sampling guide for more info.

To cover the intraspecific variation and to allow miss-identified samples to be recognised as such, at least **20 individuals per species** are needed.

Reference samples are prepared so that an image of wood, usually a transverse surface (*e.g.* Rosa *et al.* 2017; Tang *et al.* 2017; Ravindran *et al.* 2018) can be captured and processed by the MV system. **Reference slides** are, in order of priority[5], produced of transverse, tangential and radial planes. Both stained and unstained slides are prepared for each specimen. Using a standard stain can improve inter-specimen interpretability.

FUTURE VIEW: If all wood collections in the world would digitize their microsections with high resolution scanners, reference collections for machine vision could easily be made. Instead of starting from scratch finding wood samples, it would allow benefiting from the huge slide collections[6] that are already present in the world.

[5] Radial planes are often not that informative since the correct radial orientation of the section will have a great impact on the overall appearance of the radial section.

[6] Not all slides, however, will be of sufficient quality. Many historical thin sections are only useful to the trained human eye, machines might be misled by different thickness, orientation, stain, artefacts, … .

Image collection

In priority order the **transverse, tangential and radial surfaces** should be imaged. Image datasets captured from suitably prepared curated reference materials should have **sufficient resolution** to satisfy the following objectives:

- archival quality digitization of specimens sufficient to create *de facto* digital xylaria
- designation of archetypal specimen images for education purposes
- facilitate training and deployment of accurate MV classifiers for wood species identification

Model development

A field screening model targets a group of taxa for identification of anatomical features. Models for wood identification via MV methods can be subdivided according to the **classifier resolution**:

- Image (image classification to identify *e.g.* the species of wood)
- Pixel (image segmentation to locate and delineate individual vessels, rays, axial parenchyma patterns)

For **classifier development** modern MV methods can be used, such as deep learning (DL) with convolutional neural networks (CNN) that automatically learn highly discriminative features[7] from images, rather than using hand-designed filters.

Model validation

External validation is best accomplished by testing:

- ☑ under real-world conditions
- ☑ with multiple new specimen sets (from xylaria or from the field) never seen by the model
- ☑ across multiple pieces of the system hardware
- ☑ by a range of operators

[7]The early layers of a convolutional neural network have been shown to find filters that are similar to the manually engineered filters used in traditional 'shallow' techniques (Zeiler & Fergus 2014), but their correspondence to wood identification characters is typically unknown.

In the absence of such testing it would be unwise to assume that field performance of a deployed model will function at the same accuracy level as the accuracy reported on a test data set.

Determination of the necessary **threshold for field screening accuracy** should be determined by each country or enforcement entity according to relevant standards of jurisprudence.

1.2.2 WOOD ANATOMICAL ANALYSIS OF TEST SAMPLES

1.2.2.1 MACROSCOPIC ANATOMICAL ANALYSIS OF WOOD

After having developed a list of **structural features** to be observed, they can be used for wood identification with the help of comprehensive **reference material**, *e.g.* numerous timber atlases or computerized databases (Table 1).

CAUTIONARY NOTE: It is well-known that all, natural history collections have a proportion of incorrectly named specimens. Ideally, all reference collections should contain several replicates of each species, and all these replicates should ideally have correctly identified voucher material. If there is doubt about the accuracy of physical specimens, the use of a combination of physical and digital/printed reference material is recommended.

A **fresh end grain cut** is an absolute necessity as only a clean surface will reveal sufficient details of wood structure. Also be sure to turn the specimen at various angles under the light as the reflectivity of wood tissues varies. Moistening the wood surface with water or saliva can also improve the observation of anatomical features.

In most cases, an identification request contains **background information** (*e.g.* timber comes from a temperate European forest), allowing a significant restriction of potential taxa for the unknown sample. Such background information must always be interpreted with scepticism, as it can be incorrect. Conversely, it can greatly simplify the identification if the claimed provenance is clearly incorrect (*e.g.* if the supposed temperate specimen is clearly a tropical wood).

When observing an unknown sample, these are the **major questions:**

- Softwood or hardwood?

If a hardwood:

- What is the porosity type?
- What is the vessel clustering?
- Is there a conspicuous pattern of vessel arrangement?
- Are vessels easy to see (vessel diameter) and crowded or not (vessel frequency)?
- Is there a conspicuous pattern of axial parenchyma arrangement?
- Are rays wide? Are they of two sizes?
- Is storied structure present?
- Are there any observable secretory structures?

If a softwood:

- Are resin canals visible?
- Is there a distinct latewood band?
- What is colour and odour of the wood?

Once the key characteristics of the unknown sample are observed, **side-by-side observation of the unknown with known reference specimens** of suspected species is critical to determine the best identification

1.2.2.2 MICROSCOPIC ANATOMICAL ANALYSIS OF WOOD

Anatomical features are compared and interpreted step-by-step based on previous experiences of the practitioner. Various dichotomous and multi-access identification keys can be used among which *InsideWood* is one of the most used. Tips on using *InsideWood* for identifying an unknown wood sample can be found in Wheeler (2011) and on *InsideWood*'s Welcome Page and in the Links section. In addition, we can give the following advice:

- ☑ If the unknown specimen is large enough to hold in your hand, begin by making **macroscopic observations** (see §1.2.2.1 *Macroscopic anatomical analysis of wood*).
- ☑ Now, look at all relevant planes using **microsections**.
- ☑ **Choose 1, 2 or 3 characters you are certain about**[8] and do not allow mismatches and that you know from experience will narrow the search. Alternatively, you can check in *InsideWood* how good a character is by just searching on that one character and seeing how many matches there are.
- ☑ Gradually build up the description of your unknown until you have a **manageable shortlist of possible identities**. Be wary if you get only one or two matches, this rarely happens (*e.g.* the three characters "concentric phloem, solitary vessels, Sri Lanka" deliver only 1 match: *Salacia reticulata*).
- ☑ If **background information** is given in the enquiry, the shortlist can be compared against this information (*e.g.* a geographical location) but see the earlier caveat about trusting this information (§1.2.2.1 *Macroscopic anatomical analysis*).
- ☑ Once you have a shortlist you can compare the unknown with a combination of different **reference materials** (slides, atlases, databases ➤ see Table 1) such as the images on *InsideWood* and reference slides. As with all keys, knowledge of the terminology, appearance of the characters and some experience are needed.

Be aware that if using an interactive identification system, it is still imperative[9] that the identification be verified by reference to the literature and properly identified slides.

[8] Note that you have to be certain (i) that your unknown has/lacks the character, and just as importantly, (ii) that you can trust the coding of the character in the database (see Box 3).

[9] The former use of dichotomous keys taught us that even when getting down to a single name, this could be totally wrong (as learned after a bit more investigation).

BOX 3: NOBODY IS PERFECT, NOT EVEN REFERENCE MATERIALS

A wood identification can only be as good as the identifier and the reference collection(s) used. For almost all databases, not all descriptions are based on vouchered or properly identified wood samples (see also Box 1-2). In addition, species descriptions and images can vary due to different authors and/or different reference samples. Therefore, it is so important to **use a combination of reference materials** to include as much biological and human variability in the reference set as possible. Besides, most databases are incomplete and if the unknown taxon is not included in the reference collection being used, it is not possible to arrive at the correct identification.

A specific caveat of the *InsideWood* database and most systematic wood anatomy publications is that some hard to see IAWA features (*e.g.* perforated ray cells, disjunctive ray parenchyma, vessel-ray parenchyma pits restricted to marginal rows, crystals, silica, axial canals) are not visible in the photographs, not mentioned in the descriptions, or indicated as being absent in the taxon. Therefore, **use feature absence with great care** as 'absence' of features may not mean true absence in the species.

1.2.2.3 ANATOMICAL ANALYSIS OF CHARCOAL

Charcoal images are generally evaluated following the same procedures used to analyse microscopic wood images. However, there are some special features to consider when identifying wood species in charcoal.

☑ Make observations at **different magnification levels of the longitudinal planes** (radial and tangential).

Comparison with reference data from both solid wood and charcoal in atlases and databases (see Table 1):

☑ **Use quantitative characters with caution**. The carbonization process leads to a structural and dimensional shrinkage of the material. Quantitative values in the databases can only be used taking into account these structural changes.

☑ **Only use clearly identified characters** to classify family or genus to avoid mismatches.

☑ **When the provenance is known to be (sub)tropical**, certain individual characters can be analysed such as the presence of septate fibres, size, and arrangement of intervessel pits. Detailed images of different tissue types are then useful.

☑ **Use background information with care.** Charcoal with an archaeological context and reliable sourcing information usually originates from the area around the archaeological site. Recent charcoal, however, is traded worldwide via complex supply chains. Provenance information should therefore only be used if you or people you trust collected the samples.

Comparison with physical reference material is recommended from charcoal and/or solid wood and should be done **after delimiting the possible genus or species**. If there is no suitable reference material for comparing the charcoal sample, the exchange with other scientific wood[10] and/or charcoal collections should take place.

ⓘ **Over 50 charcoal collections exist worldwide** (Scheel-Ybert 2016). The largest is located at the Institute of Evolution Sciences of Montpellier in France. The wood collection at the Thünen-Institute in Germany also provides charcoal reference collections for wood species found in internationally traded charcoal and charcoal briquettes.

1.2.2.4 USING MACHINE VISION SOFTWARE FOR WOOD IDENTIFICATION

Preparation and image collection from the unknown sample as described in *§1.2.1.4 Development of reference data for machine vision.*

The MV model performs the **classification** tasks either in the on-site hardware (*e.g.* Ravindran *et al.* 2018) or through a cloud-based interface (*e.g.* Tang *et al.* 2017). **Processing of predictions** can be done by:

- ranking according to probability or similarity
- thresholding to some minimum confidence
- coupling with comparative images for human-machine-hybrid decision making

It should be dependent on the needs of the end-user, in the context of their legal and enforcement system, which of these approaches a system takes. Thus, there is no single 'best' approach.

[10] For a list of wood collections see the Index Xylariorum.

Validation of classification. The generalizability of models to test specimens has to be carefully verified for each model before trusting it to deliver accurate results in practice. To date, no published literature demonstrates or reports real-world performance of a machine vision system, so there is no scientific evidence that these systems work in the field.

Best practices to deal with overfitting, the phenomenon where model prediction accuracy on training data is much better than on the test data:

- ☑ Fully independent sets of samples used for training or for testing. A specimen should not contribute images to both the training data set and the test set.
- ☑ K-fold validation to ensure that the overall accuracies are not a result of a single, rare, artificially high-performing model.
- ☑ Field testing the system (i) on entirely new and validated specimens, (ii) by multiple users (iii) on several distinct pieces of hardware.

1.2.3 STRENGTHS & LIMITATIONS

1.2.3.1 MACROSCOPIC ANATOMICAL ANALYSIS OF WOOD

Strengths

Little equipment is needed. Therefore, it is the **fastest method** to do an evaluation, or sometimes even an accurate identification of the traded timber. In many cases the genus suggested by the trade name can be verified at the macroscopic level. This is however only true **for a trained wood anatomist**. Experience is even more important for macro- than for micro-scopic wood identification. To see the macroscopic features well, your eye should be trained, and the best training is daily practice identifying unknown specimens in comparison to reference materials, that is, identifying specimens in a xylarium.

Limitations

Macroscopic identification of **softwoods** is much more difficult than microscopic identification due to the considerably smaller number of characters available for observation.

Macroscopic characters are subject to **high variability** due to different growth conditions of the tree or history since harvesting (*e.g.* exposure to oxygen and UV radiation.

A NOTE ON COLOURS. Colour can be an important diagnostic feature for the identification of timbers, such as rosewoods and ebonies. Care is needed when using this feature, however, as there are many factors influencing the colour of heartwood[11], of stored samples, as well as of images in reference collections.

Macroscopic wood identification is a suitable **screening method**[12], a first filter for taxa. In some cases, a final identification can be made but this depends on the taxon and on the level of specificity requested. Especially for identification up to the species level, the method should be supplemented with other techniques that can provide species-level resolution. Also in the case of closely related (trade) timbers, the use of macroscopic characters will end with a choice of several likely matches, whose safe separation has to be left to microscopic analysis performed by any of the scientific institutions, which possess the necessary infrastructure and experienced staff.

Finally, there are limits on the **sample size**. The area of sample to be tested should be big enough to reveal useful structural details. For example, in the case of thin veneers (generally around 0.5 mm thick) typical structural patterns (such as vessel and axial parenchyma arrangement) cannot be fully recognized[13].

NOTE: Except in cases of tiny specimens, nearly all wood identifications that are based mostly on microscopic wood anatomy incorporate macroscopic characters and observations.

1.2.3.2 MICROSCOPIC ANATOMICAL ANALYSIS OF WOOD

Strengths

Microscopic wood anatomical identifications are based on the **generally used list of 163 anatomical characters** (51 categories), supplemented with a set of 58 non-anatomical descriptors (11 categories) (IAWA Committee 1989). This framework has facilitated the microscopic **descriptions of about 8700 timbers** (genus or species), currently available and documented in several computerized databases (Table 1). In addition, the anatomical descriptions of most references are supplemented with

[11] A range of teak samples can look quite different, while a trio of mahoganies (*Swietenia, Khaya, Entandrophragma*) can have exactly the same colour.

[12] Macroscopic screening is often combined with a wood density assessment.

[13] If the suspected identity of the sample is suggested, however, this can be checked against reference material. In this way even veneer of only 200 µm thick can be identified.

images. These databases and the **extensive reference collections**, of wood samples and microscope slides (see the Index Xylariorum), provide an essential reference material for the routine identification of internationally traded timbers and wood products. Microscopic analyses are routinely conducted for **use in official or legal cases**. In addition, microscopic analyses enable the **identification of the full set of wood products**, from solid wood, over thin veneer layers, down to wood chips and paper fibres.

BOX 4: SUCCESS STORY FROM A GTTN MEMBER

The Thünen Centre of Competence on the Provenance of Timber (Hamburg, Germany) has answered more than 4500 official requests (approx. 40 000 specimens) for microscopic wood identification since the implementation of the EUTR in Germany in 2013. Such inquiries mainly come from the timber trade and trade monitoring sectors (customs, conservationists) and increasingly from the authorities. Private consumers also show an increasing interest in knowing whether they have acquired a correctly named product.

Limitations

Laboratory equipment is required to prepare and observe microsections of the unknown wood sample to be identified.

The **minimum size of the specimen** required is usually 5-10 mm (cross-section and length), which allows sample preparation without an additional embedding procedure. Smaller specimens, *e.g.* individual veneer layers or fibre particles, can be embedded or mounted directly for the preparation of microscopic slides which requires more time and methodological effort. But even then, there are limits to how small a transverse area may be to reveal useful structural details. In the case of thin (200-500 µm[14]) veneers, for instance, typical structural patterns cannot be recognized in the transverse plane and need to be reconstructed from the corresponding patterns on longitudinal surfaces.

Some genera and many species have similar wood anatomy, so that sometimes a single description applies to a group of genera, an entire genus, or to species groups within a genus, *e.g. Quercus* spp. for the red oaks, white oaks, and evergreen oaks. In general, **it usually is only practical to identify an isolated piece of wood to genus**. Most descriptions are based on but a few samples, so likely do not represent the

[14] For example, hardwood plywood are mainly tropical face veneers of about 200 µm thick with core layers of lower-value woods like poplar or eucalypt.

entire species/genus variation. Moreover, trees are large organisms and single reference samples do not reflect the intra-tree variation.

1.2.3.3 ANATOMICAL ANALYSIS OF CHARCOAL

Strengths

Specific reference images and data for charcoal, developed by analysing high resolution images, provided by advanced microscopy techniques (see §1.2.1.3 *Development of charcoal reference data*) allows for **faster and more accurate charcoal identification**.

ⓘ **Specific charcoal references are not always needed**. For the most common identification questions, such as whether the charcoal is from temperate or tropical species no detailed anatomical analysis is required. With access to a reference wood or slide collection and some experience, you can compare what you see in wood to what you see in charcoal (keeping in mind that features are smaller in charcoal than in the original wood due to shrinkage).

Limitations

A successful evaluation of charcoal images depends on sample size and conservation status of the sample, as well as on the available equipment and references.

- **Fragments of 2-4 mm in cross-section and length** can be identified, provided that the images that can be produced from the sample clearly show all relevant anatomical features in the three planes (longitudinal, radial, and tangential).
- Carbonisation can lead to a variable **degree of fusion of cell elements**. For charcoal samples where structures are highly melted together, too many non-visible features can make it impossible to identify the wood species.
- **Access to a 3D-RLM, SEM or FESEM** is required for the development of charcoal reference data and for the testing of closely related species, tropical or uncommon woods.
- **A specific charcoal database** containing all relevant worldwide traded wood species for charcoal production is currently unavailable.

The creation of such databases has already begun at various scientific institutes (*e.g.* Institute of Evolution Sciences of Montpellier in France, Thünen Institute in Germany). Regional databases such as the *CharKey* for French Guiana (Bodin *et al.* 2019) already exist.

1.2.3.4 USING MACHINE VISION SOFTWARE FOR WOOD IDENTIFICATION

Strengths

Modern computer vision and machine learning techniques provide effective methods for **macroscopic** (*e.g.* Khalid *et. al.* 2008, Tang *et al.* 2017, Ravindran *et al.* 2018,) and **microscopic** machine vision (MV) identification (Rosa *et al.* 2017) of wood specimens, **potentially[15] to the species level**, although none of these papers report real-world testing results.

Ideally, a machine vision system requires no knowledge of wood anatomy. At present a system like the *XyloTron* requires some wood anatomical knowledge to orient the specimen correctly in the field of view, however, far **less training is required in wood anatomy for operation of MV tools** than for traditional field wood identification. When the promise of machine vision wood identification is realized, fewer trained people will be needed to deliver the same output and experts can focus their attention on identification problems that go beyond the capacity of a machine vision system.

In contrast to the development and evaluation of MV tools (see *Limitations* below), operating the software generally requires only the most rudimentary computer skills, with individual specimens identifiable via a single click or keystroke. MV models are **not subject to a range of human failings**, such as boredom, fatigue, lack of practice, subjectivity, and inconsistency over time. A single MV device could conceivably be operated 24 hours per day.

Limitations

The limitations for taxonomic precision for MV models are largely the same as those for macro- and micro-scopic wood anatomy, *i.e.* genus level identification. The caveat is that **in the absence of extensive field testing and verification, it is difficult to quantify model overfitting** and thus overestimating system accuracy. Overall, due

[15] Although only 77 species, it are all commercial species of the Democratic Republic of Congo that could be identified in Rosa *et al.* (2017).

to the biological variability of wood, wood identifications are as good as the set of reference samples used for the analysis.

The importance of **access to relevant scientific expertise to develop and evaluate MV tools** cannot be overstressed. Despite growing power, image-based MV wood identification remains an inherently human-mediated activity, relying on human judgement in model development and effective field deployment, as the human operator must make decisions to accept or reject the MV tool's results.

Model development for most MV methods requires **access to Graphics Processing Unit (GPU) computing resources** and specialized machine learning expertise, but these functions could be readily democratised using cloud-based computation resources.

For on-site, hardware-based MV processing, individual systems have to be updated with new models as they are developed, but while in use do not require an active data connection. **For cloud-based processing**, up-to-the-minute models can be employed but require an **active data connection with sufficient bandwidth** for real-time operation.

1.3 DATA ANALYSIS FOR TAXON IDENTIFICATION OF PULP, PAPER AND FIBREBOARD

1.3.1 DEVELOPMENT OF VESSEL ELEMENT REFERENCE DATA

In most cases, pulp, paper, and fibreboard are made of more than one timber species. Maceration of the pulp, paper or fibreboard hence gives isolated cells of a mixture of wood species. In fibre material made of hardwood the vessel elements are the cell type with the most structural characteristics. Therefore, **high-value descriptions** and **micrographic illustrations** of these cells are needed as references[16].

1.3.1.1 PREPARATION OF REFERENCE SLIDES

For the preparation of slides that can be stored permanently as references:

- ☑ **Take splits** of 1 cm length, 1 mm thickness and 1 cm in the radial direction to obtain a representative selection of vessel elements from successive growth increments.
- ☑ **Boil** the splits in distilled water and then **macerate** them in a solution of equal parts of glacial acetic acid (99%) and H_2O_2 (30%) at 60°C for 48h following the method of Franklin (1945)[17]. The woody tissue is broken up.
- ☑ **Separate the individual cell elements** by shaking the test tubes.
- ☑ In a plastic filter, **rinse** the resulting suspension with tap water and **stain** with Nigrosin (1%) (alcohol soluble).
- ☑ **Select the vessel elements** from the other cells in the pulp by needle using a reflected light microscope, place them on microscope slides, and embed in Euparal. Place the microscope slides under a flue to evaporate for 1 day and then heat at 50°C for 10 days.

1.3.1.2 PUBLISHING REFERENCES

To make the references available for all testing institutes worldwide a scientific publication is essential. Therefore, reference slides have to be analysed (for an example see Helmling *et al.* 2018) and documented with **micrographs using a**

[16] See also this presentation on the need for vessel atlases.

[17] For a detailed description see Helmling *et al.* (2016).

transmitted light microscope (bright field microscope) with measuring software[18] (or eyepiece grid or drawing tube with self-made calibration units). The observed structural characteristics should then be **verified with literature data** from various sources (such as *InsideWood*).

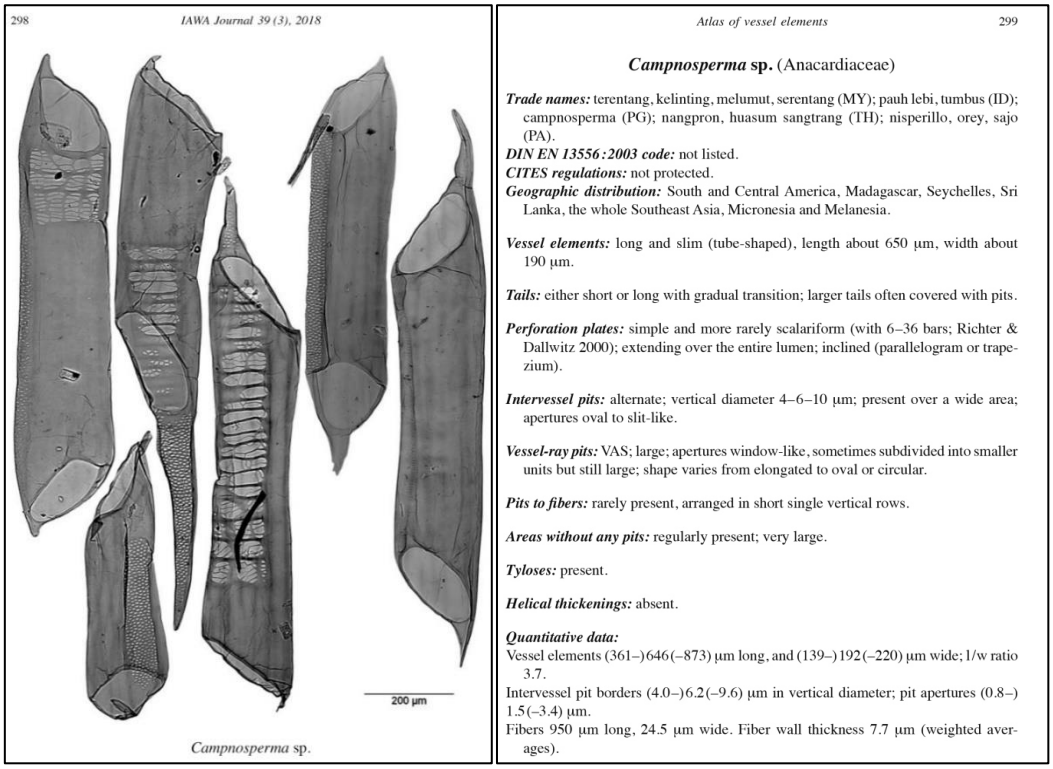


Fig. 7: Example of a reference for wood identification of pulp, paper and fibreboard showing *Campnosperma* sp. Photomicrographs of the vessel elements are produced with a light microscope. Each image of a vessel element is a superposition of up to 10 micrographs taken at different focus points with a light microscope (Olympus, BX51) (from Helmling *et al.* 2018).

[18] *E.g.* ImageJ is an open source software.

1.3.2 ANALYSIS OF TEST SAMPLES FROM FIBRE MATERIAL

CAUTION: Since it is common practice in industry to produce pulp, paper, and fibreboard of a **mixture of timber species** it is important to obtain a representative range of vessel elements and to identify each of them separately. Therefore, it is recommended to prepare and investigate enough material, which usually requires several microscopic slides.

1.3.2.1 PREPARATION OF FIBRE SAMPLES FOR IDENTIFICATION

To prepare the samples for investigation **fibres are separated into individual cells and stained**.

Paper

A piece of paper (1.5 cm^2) is put into a test tube filled with **distilled water to soften** the paper. After a few minutes, the pulp mass is rolled to a ball between the fingers and placed in a tube with water again. The test tube is shaken, and the fibre suspension poured on a filter.

Part of the material is then placed on 2 or 3 microscope slides. Staining of the fibrous mix is done with two drops of **Alexander and** one drop of **Herzberg solutions**[19], which allow differentiation between chemical pulp (violet) and mechanical wood pulp (yellow). Cover glasses are placed on the stained pulp, and slides should be investigated immediately, since the staining is not permanent. After this process cells with a high amount of lignin (mechanical pulp) are stained yellow, de-lignified cells (chemical pulp) are greyish blue (softwoods) or blue (hardwoods).

Fibreboard

A piece of the material is **boiled in water** and the resulting pulp placed on a filter.

For better contrast, the pulp is stained with **Nigrosin** (1%, alcohol soluble). After 10-15 minutes the material is rinsed with de-ionised water and mounted in glycerine (70-90%) on microscope slides.

[19] Preparation of the solutions is described in Harders-Steinhäuser (1974). Alexander solution consists of calcium nitrate tetrahydrate, Herzberg solution of zinc chloride and potassium iodide (Merck KGaA, Darmstadt, Germany).

1.3.2.2 COMPARING VESSEL ELEMENTS OF TEST SAMPLE WITH REFERENCE DATA

Restrict the list of potential candidates for identification of the unknown sample by looking at the main anatomical characteristics:

- Type of perforation plate
- Size of intervessel pits
- Vessel-ray pits[20]
- Helical thickenings
- Tyloses
- Vessel element length, width, and ratio of length to width

Compare the appearance of vessel elements to make a clear match with the reference. All the details of interest - like arrangement of the different pit types and shape of the pits - can only be provided by physical references or high-quality micrographs. A well-trained person with practice can directly distinguish the different commonly used timbers. In general, the anatomy of vessel elements of the timber species within a genus or in some cases within a family is so similar, that the identification of timbers in pulp, paper and fibreboard is restricted to genus or family level.

[20] Vessel-ray pits are either similar in size and shape to intervessel pits (All Pits Similar, APS) or different (Vessel-ray pits Apparently Simple, VAS) (Helmling *et al.* 2018).

1.3.3 STRENGTHS & LIMITATIONS

Strengths

Wood anatomy currently provides the only established method for the **identification of pulp, paper, and fibreboard components**, which are also subject to the controls of timber regulations.

The microscopic analysis of fibre material can identify cells, which fully correspond to scientifically described fibre references of timbers **up to family or genus level**. It can be substantiated if a declaration of the incorporated timbers in fibre materials is correct or if there are additional timbers represented.

Limitations

The number of useable microscopic features for identification is severely reduced in the macerated tissue of pulp, paper, and fibreboard, in comparison to the microscopic identification of solid wood blocks. The few anatomical characteristics of an unknown vessel element lead to many possible candidates. Therefore, **expertise is crucial** since the appearance of the unknown vessel element has to be recognized and matched with the reference material.

Fibre material is mostly made up of multiple timbers complicating their identification.

Identification to species level is not possible for pulp, paper, and fibreboard.

Scientifically described fibre references are still limited, especially for tropical timbers.

1.4 KEY LITERATURE FOR WOOD ANATOMICAL DATA ANALYSIS

BOX 5: NON-EXHAUSTIVE OVERVIEW OF PRINTED WOOD ANATOMICAL REFERENCE ATLASES

Akkemik, Ü. and B. Yaman (2012). Wood anatomy of eastern Mediterranean species. Kessel Publishing House.

Benkova, V.R. and F.H. Schweingruber (2004). Anatomy of Russian Woods. An atlas for the identification of trees, shrubs, dwarf shrubs and woody lianas from Russia. Birmensdorf, Swiss Federal of institutes for Forest, Snow and Landscape Research. Berne, Stuttgart, Vienna, 456 pp.

Bonilla, L.A.M.A., J. Barajas Morales and P.T. Lezama (2004). Anatomía de Maderas de México. Árboles y Arbustos del Matorral Xerófilo de Tehuacán. Publicaciones Especiales Del Instituto de Biología 19. Universidad Nacional Autónoma de México.

CITES (2002). CITES identification guide - tropical woods. Environment Canada. [pdf]

Crivellaro, A. and F.H. Schweingruber (2013). Atlas of Wood, Bark and Pith Anatomy of Eastern Mediterranean Trees and Shrubs with a Special Focus on Cyprus. Springer

Detienne, P. and P. Jacquet (1983). Atlas d'identification des bois de l'Amazonie et des regions voisines. CTFT.

Eom, Y.G. (2015). Wood anatomy of Korean species. Media Wood Ltd. 606pp.

García Esteban, L. (2002). Anatomía e identificación de maderas de coníferas a nivel de especie. Mundi Prensa. [including a CD]

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

Gregory, M. (1994). Bibliography of systematic wood anatomy of dicotyledons. IAWA Journal Supplement 1. 265 pp. ISBN 90-71236-22-6.

Harders-Steinhäuser, M. (1974). Faseratlas zur mikroskopischen Untersuchung von Zellstoffen und Papieren. Biberach/Riß: Güntter-Staib Verlag.

Helmling, S., A. Olbrich, I. Heinz and G. Koch (2018). Atlas of vessel elements. IAWA Journal 39(3): 249-352.

Ilvessalo-Pfäffli, M.-S. (1995). Fiber Atlas – Identification of Papermaking Fibers. Berlin, Heidelberg, Springer Series in Wood Science. Springer-Verlag.

Jiang, X. (2009). Atlas of Gymnosperms woods of China. 490pp. ISBN 978-7-03-026138-0.

Lemmens, R.H.M.J., I. Soerianegara and W.C. Wong (1995). Plant Resources of Southeast Asia 5(2). Timber trees: minor *Commercial timbers*. Backhuys Publishers, Leiden.

Lemmens, R.H.M.J., D. Louppe and A.A. Oteng-Amoako (Editors) (2012). Plant Resources of Tropical Africa 7(2). Timbers 2. PROTA Foundation, Wageningen, Netherlands/ CTA, Wageningen, Netherlands. 804 pp.

Louppe, D., A.A. Oteng-Amoako and M. Brink (Editors) (2008). Plant Resources of Tropical Africa 7(1). Timbers 1. PROTA Foundation, Wageningen, Netherlands/ Backhuys Publishers, Leiden, Netherlands / CTA, Wageningen, Netherlands. 704 pp.

Neumann, K., W. Schoch, P. Détienne and F.H. Schweingruber (2000). Woods of the Sahara and the Sahel. An anatomical atlas. Eidg. Forschungasanstat WSL, Birmendorf, Verlag Paul Haupt.

Ogata, K., T. Fujii, H. Abe and P. Baas (2008). Identification of the timbers of SE Asia & the Western Pacific. Kaiseisha Press, Japan, 400pp.

Schweingruber, F.H. (1990). Microscopic Wood Anatomy; Structural variability of stems and twigs in recent and subfossil woods from Central Europe. 3rd edition. Birmensdorf, Eidgenössische Forschungsanstalt WSL.

Soerianegara, I. and R.H.M.J. Lemmens (Editors). (1993). Plant Resources of Southeast Asia 5(1). Timber trees: major *Commercial timbers*. Pudoc Scientific Publishers, Wageningen.

Sonsin, J.A., P. Gasson, S.R. Machado, C. Caum and C.R. Marcati (2014). Atlas da diversidade de madeiras do cerrado paulista. Fundacao de Estudos e Pesquisas Agricolas e Florestais.

Sosef, M.S.M., L.T. Hong and S. Prawirohatmodjo (1998). Plant Resources of Southeast Asia 5(3). Timber trees: lesser-known *Commercial timbers*. Backhuys Publishers, Leiden.

Wheeler, E.A. and P. Baas (1998). Wood identification - A review. IAWA Journal 19(3): 241-264.

Bodin, St.C., R. Scheel-Ybert, J. Beauchêne, J-F. Molino and L. Bremond (2019). CharKey: An electronic identification key for wood charcoals of French Guiana. IAWA Journal 40(1): 1-17.

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.

Hafemann, L.G., l.S. Oliveira and P.R. Cavalin (2014). Forest species recognition using deep convolutional neural networks. In: International Conference on Pattern Recognition: 1103-1107.

Helmling, S., A. Olbrich, L. Tepe and G. Koch (2016). Qualitative and quantitative characteristics of macerated vessels of 23 mixed tropical hardwood (MTH) species: a data collection for the identification of wood species in pulp and paper. Holzforschung 70(9): 839-844.

Hermanson, J.C. and A.C. Wiedenhoeft (2011). A Brief Review of Machine Vision in the Context of Automated Wood Identification Systems. IAWA Journal 32(2): 233-250.

Hermanson, J. and A. Wiedenhoeft (2015). Data-driven wood anatomy: Using machine vision for wood identification (and beyond). Integrative and Comparative Biology 55:E273-E273 (poster abstract).

Hubau, W., J. Van den Bulcke, P. Kitin, L. Brabant, J. Van Acker and H. Beeckman (2013). Complementary Imaging Techniques for Charcoal Examination and Identification. IAWA Journal 34(2): 147-168.

IAWA Committee (1989). IAWA List of Microscopic Features for Hardwood Identification. IAWA Bulletin 10(3): 219-332.

IAWA Committee (H.G. Richter, D. Grosser, I. Heinz & P.E. Gasson, Eds) (2004). IAWA list of microscopic features for softwood identification. IAWA Journal 25: 1-70.

Khalid, M., E.L.Y. Lee, R. Yusof and M. Nadaraj (2008). Design of an intelligent wood species recognition system. International Journal of Simulation System, Science and Technology 9(3): 9-19.

Martins, J., L.S. Oliveira, Jr. A.S. Britto and R. Sabourin (2015). Forest species recognition based on dynamic classifier selection and dissimilarity feature vector representation. Mach. Vis. Appl. 26: 279-293.

Martins, J., L.S. Oliveira, S. Nisgoski and R. Sabourin (2013). A database for automatic classification of forest species. Mach. Vis. Appl. 24: 567-578.

Parham, R. and R. Gray (1982). The practical identification of wood pulp fibers. Atlanta, Tappi Press.

Ravindran, P., A. Costa, R. Soares and A.C. Wiedenhoeft (2018). Classification of CITES-listed and other neotropical Meliaceae wood images using convolutional neural networks. Plant methods 14(1): 25.

Richter, H.G. and M.J. Dallwitz (2000 onwards). *Commercial timbers*: descriptions, illustrations, identification, and information retrieval. In English, French, German, and Spanish. http://www.delta-intkey.com/wood/index.htm

Richter, H.G. and M.J. Dallwitz (2016 onwards). *Commercial timbers*: descriptions, illustrations, identification, and information retrieval. In English, French, German, Portuguese, and Spanish. http://delta-intkey.com

Richter, H.G., K. Gembruch and G. Koch (2005 onwards). In English, French, German, and Spanish. http://www.delta-intkey.com/citeswood/index.htm

Richter, H.G., M. Oelker and G. Koch (2002 onwards). *MacroHOLZdata* – Computer-aided macroscopic wood identification and information on properties and utilization of trade timbers. CD-ROM, Thuenen-Institut, Hamburg. In English, German and Spanish. Contact: gerald.koch@thuenen.de

Rosa da Silva, N., M. De Ridder, J.M. Baetens *et al.* (2017). Automated classification of wood transverse cross-section micro-imagery from 77 commercial Central-African timber species. Annals of Forest Science 74:30. https://doi.org/10.1007/s13595-017-0619-0

Ruffinatto, F., A. Crivellaro and A.C. Wiedenhoeft (2015). Review of Macroscopic Features for Hardwood and Softwood Identification and a Proposal for a New Character List. IAWA Journal 36(2): 208-241.

Scheel-Ybert, R. (2016). Charcoal collections of the world. IAWA Journal 37 (3):489-505.

Tang, X.J., T.H. Tay, N.A. Siam and S.C. Lim (2017). Rapid and Robust Automated Macroscopic Wood Identification System using Smartphone with Macro-lens. In Proceedings of PRWAC 2017.

Wheeler, E.A. (2011). *InsideWood* - a Web Resource for Hardwood Anatomy. IAWA Journal 32(2): 199-211.

Zeiler M.D. and R. Fergus (2014). Visualizing and Understanding Convolutional Networks. In: Fleet D., T. Pajdla, B. Schiele and T. Tuytelaars (eds). Computer Vision – ECCV 2014. ECCV 2014. Lecture Notes in Computer Science, vol 8689. Springer, Cham.

Zemke, V.TH., V. Haag and G. Koch (submitted to IAWA). Wood Identification of charcoal with 3D-reflected light microscopy.

2. Genetics

Definition genetic reference data: genotypes (sequence or molecular marker based) of georeferenced individuals, of a certain taxon (species, genus) or closely related taxa, covering most of the distribution range (taxon reference) or only a specific geographic area (provenance reference).

Authors: Céline Blanc-Jolivet, José Antonio Cabezas, Simon Crameri, Bernd Degen, Andrew Lowe, Melita Caroline Low, Justyna Anna Nowakowska, Niklas Tysklind

*Authors are in alphabetical order

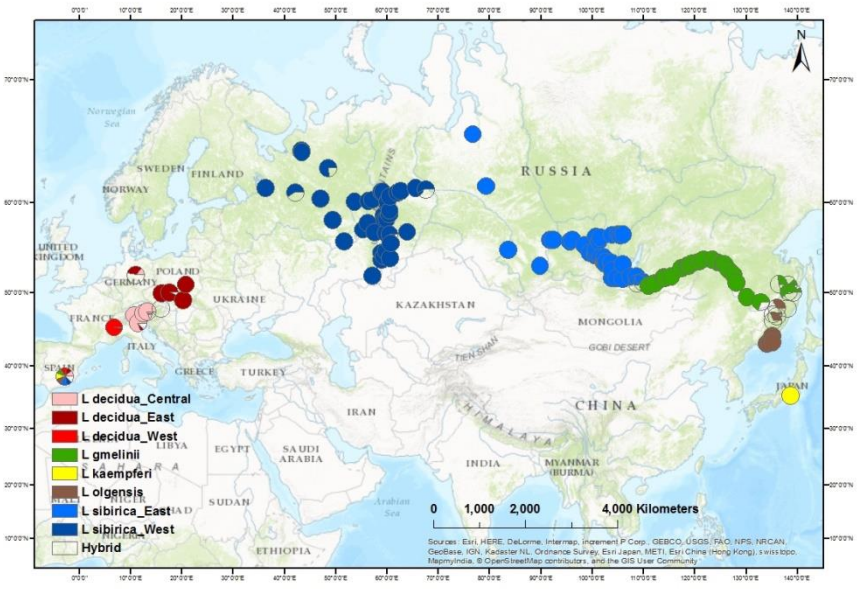


Fig. 8: Geographical distribution of the different genetic profiles of *Larix* spp.

2.1 RESOURCES REQUIRED FOR GENETIC ANALYSIS

Table 2: Overview of the most used software for species and/or provenance determination of wood using genetic methods.

Software (**R* package)	Performance	Species	Provenance	Accessibility
*adegenet**	exploration, basic statistics, clustering testing, assignment	x	x	Free software
*assignPop**	clustering testing, assignment	x	x	Free software
CLUMPP[21]	clustering	x	x	Free software
GDA-NT	exploration, basic statistics, assignment	x	x	Free software, on request
GenAlEx	exploration, clustering testing, basic statistics, assignment	x	x	Free software
GeneClass	clustering testing, assignment	x	x	Free software
*Geneland**	clustering, assignment		x	Free software
GeoAssign	assignment	x	x	Free software
Oncor	clustering testing, assignment	x	x	Free software
*poppr**	exploration, basic statistics	x	x	Free software
STRUCTURE	clustering, assignment	x	x	Free software

[21] *CLUMPP* does not do any group inference but takes already inferred groupings and does downstream analysis after grouping inference with population-genetic cluster analyses such as *Geneland* or *STRUCTURE*.

2.2 INTRODUCTION

2.2.1 BASICS OF POPULATION GENETICS

When using genetic methods in the identification of an unknown sample, we are **using the result of eco-evolutionary processes** such as mutation, genetic drift, migration and hybridisation, selection and adaptation, and speciation. Some basic knowledge of these evolutionary processes is essential prior to attempting analysis and interpretation of genetic data.

In the wild, individuals breed with other nearby individuals with which they are genetically compatible. Every individual, other than twins or clones, is genetically unique, although it shares much of its DNA with other individuals (in decreasing order of similarity) of the same family (*i.e.* parents and siblings), population, species, genus, order. The group of interbreeding individuals is known as **a population**, and it **can be recognised with genetic markers**, because the frequency of alleles in the population will conform to already established genetic models (*e.g.* Hardy-Weinberg equilibrium (HWE) and linkage disequilibrium (LD)). Individuals from different populations have the potential to interbreed but do so very infrequently, leading to differences in allele frequencies, and incongruences in the genetic models. These populations can further differentiate themselves, specialising in different environments, and/or changing their phenotype, and/or even losing the capacity of interbreeding between them, at which point, we humans apply the term 'species' to denote that we consider them different enough to be different entities.

It must be kept in mind, that **'species' and 'genera' are artificial human-made terms** that often do not agree with how biology works, and oftentimes we are working with taxa that are somewhere between different populations and different species, and a myriad of other human-made terms (*e.g.* subpopulations, subspecies, and species flocks, complexes, and rings) that further complicate things.

It must also be kept in mind that **different parts of the genome differentiate at different speed** throughout evolutionary processes, and genetic markers can be targeted to reveal the results of particular evolutionary processes (*e.g.* individuals, populations, species, genera). Therefore, a genetic marker that differentiates between two genera, will not be suitable to identify individuals.

When using population genetics, we try to identify groups of individuals that share some level of genetic information. By analysing different genetic markers or by moving the level of similarity up or down, we hope to identify populations, species, and genera. **The choice of genetic markers and the level of differentiation used for grouping individuals will, therefore, be dictated by the aim of the analysis**

(*e.g.* define populations for provenance testing or define species for species identification).

The peculiarities of the **plastid genomes** (mtDNA and cpDNA) mean that they evolve relatively slowly, making them ideal for distinguishing species and genera. Differences in plastid DNA can be detected by either sequencing sections of it (*e.g.* DNA barcodes), by identifying single nucleotide polymorphisms (SNPs) at specific positions (*e.g.* cpSNPs), or by studying size differences in microsatellite regions, also referred to as single sequence repeats (SSRs) and short tandem repeats (STRs). The characteristics that make plastid DNA good for genus and species identification, mean that they are generally poor at identifying individuals, populations, and closely related species.

Nuclear genomes evolve at very varied rates, but fast evolving loci, such as microsatellites, and certain SNPs, are ideal for identifying individuals. The fact that there are two copies of nuclear DNA (nDNA) mean that analyses and hypotheses based on heterozygosity (*i.e.* the condition of having two different alleles at a genomic position, known as a locus) can be used to identify interbreeding individuals (*i.e.* populations).

Populations of wild organisms are often geographically/spatially structured, which is precisely the feature we use to identify the provenance of an individual. The coverage and extent of a population will depend on its evolutionary history (*e.g.* colonisation after glaciations) and on the trees' pollen and seed dispersal capacity. Therefore, the **power of genetic tools to distinguish between two sampling locations will ultimately depend on the biology of the organism** under study and not on political boundaries.

BOX 6: GLOSSARY

DNA barcode reference database: a repository of plastid DNA barcodes (mtDNA, cpDNA) of georeferenced voucher specimens (see below) against which one can compare an unknown sample to identify the taxon it belongs to. Example: Barcode of life.

Genetic baseline reference database: a repository of organised genetic marker profiles (genotypes) of georeferenced voucher specimens of a certain taxon or closely related taxa. These genotypes indicate the population genetic structure, allowing the identification of the provenance of individuals.

Nuclear genome: all DNA that is found in the cellular nucleus of an organism. For most eukaryotes it is at least diploid and undergoes genetic recombination.

Plastid genome: all DNA that is found in cellular plastid organelles (*i.e.* mitochondria (mt) and chloroplasts (cp)). It is haploid and does not undergo recombination.

Population: a group of reproductively compatible conspecific individuals, that occurs in a specific geographic region and is genetically and/or spatially differentiated from other groups of conspecific individuals. The same population can be sampled in different locations. A single location can contain several populations (*e.g.* conspecific groups of individuals that reproduce at different times of the year, or groups of individuals of different species that are morphologically distinct or indistinct, which are then called cryptic species).

Population sample: a group of individuals collected in a single sampling location thought to be representative of the local population of a species. Not to be confused with a population. In population genetics a population sample usually contains 20 to 50 individuals and represents the unit of analysis.

Test sample: DNA extracted from a plant tissue of unknown taxonomic identity and/or provenance, whose profile must be interrogated against a database to validate its species assignment and/or geographical provenance.

Voucher specimen: plant sample of known provenance and accurate taxonomic classification.

2.2.2 CHOICE OF GENETIC MARKER AND METHODOLOGY

Most timber species can be identified with cost-effective and reliable methods, such as wood anatomy, down to the genus level. However, identifying closely related species among a genus is a difficult task. In such cases genetic methods can be powerful.

The selection of diagnostic genetic markers (targeting genetic sequences that differ between individuals, populations, species, genera, or families) is specific to:

- the scope of analysis: (i) differences between individuals or populations, or (ii) differences between plant genera/families
- the taxa of interest (*e.g. Aquilaria* agarwoods, *Dalbergia* rosewoods)

No single genetic marker allows the identification of all plant species (Li *et al.* 2015)**.** Many nuclear genetic loci are only present twice in diploid cells. Multi-copy nuclear genes, mitochondria, and chloroplasts are present in multiple copies per cell, which increases chances of successful analysis from low-quality or low-input samples, such as wood samples. Therefore, a **combination of plastid and nuclear sequences** meets different scopes of the analyses.

Discrimination of distantly related species or geographic provenances (*e.g.* plants belonging to different families or genera, often also below that level) can usually be performed by analysis of a few **chloroplast gene sequences** such as the *rbcL* and *matK* barcode genes, or in **combination with nuclear loci** with highly conserved flanking regions, such as *ITS*. In some cases (*e.g.* when populations diverged from each other a long time ago), plastid sequences correlate with geography and can be used to distinguish geographically separate populations of the same species. The level of differentiation between plastid genomes depends on factors such as population size, dispersal ability, and biogeographical history (Hollingsworth 2011). The **advantage of plastid sequences** is that they are present in high copy numbers per cell (Jiao *et al.* 2019), which increases sensitivity in low-quality wood samples, and that they can easily be compared across divergent taxa. Plant DNA barcodes are very well suited for ecological studies trying to identify ecosystem interactions. However, the relatively low species resolution of plastid sequences (*i.e.* DNA barcodes) in plants, especially in tropical tree species (Caron *et al.* 2019), preclude their general use for species identification of timber.

Discrimination among closely related species, populations of the same species, or individuals of the same population needs the analysis of ideally at least 10 microsatellites or 100 SNPs, depending on the evolutionary divergence among the units to be discriminated. Since evolutionary events, which lead to speciation, might

be either recent or not, and followed or not by migration/hybridisation events, it is advised to investigate genetic markers in both **nuclear and plastid genomes**.

The **advantage of nuclear markers** is that, if adequately sampled, they contain much more, independent and highly resolved, genomic information than chloroplast or mitochondrial markers. However, the highly variable markers that are useful at the population or individual level cannot be compared across divergent taxa. Besides, there are species with little or undetectable genetic variation, that show epigenetic variability, which allows the identification of stable epialleles that can be used for discrimination (Saéz-Laguna *et al.* 2014). However, epigenetic variability studies are not yet as generalised as genetic analysis, and thus, are out of the scope of this guide.

2.2.2.1 MICROSATELLITE MARKERS TO IDENTIFY SEIZED WOOD

Nuclear and plastid microsatellite markers, also known as simple sequence repeats (SSRs) and short tandem repeats (STRs), can be used in **illegal logging investigations**. The non-coding SSR sequences are present in all living organisms. They are neutral, highly polymorphic, and species-specific. A fast and reliable comparison can be made between **fragments of stolen fresh wood** (*i.e.* material of evidence) and **remaining, relatively fresh[22], tree stumps in the forest** (*i.e.* material of reference) using DNA profiles based on a few highly polymorphic nuclear (nSSR), mitochondrial (mtSSR), or chloroplast (cpSSR) DNA markers previously characterised for the targeted species.

Fig. 9: Sample taken from a remaining, relatively fresh, tree stump in the forest. It will be compared with fragments of stolen fresh wood, using microsatellite markers, in an illegal logging investigation.

[22] This method will become unreliable if the wood has been logged since more than 3 months.

This method has been successfully applied for the identification of some economically important coniferous and deciduous tree species, such as *Abies alba*, *Larix decidua*, *Larix kaempferi*, *Picea abies*, *Pinus sylvestris*, and *Acer platanoides, Alnus glutinosa*, *Betula pendula*, *Carpinus betulus, Fagus sylvatica*, *Fraxinus excelsior*, *Quercus petraea* and *Quercus robur*. Such DNA-based proof has been pivotal in **supporting decisions taken by several district courts** in Poland, as far as the identification of wood samples was proven by high probability (*ca.* 98-99%). More information can be found in: Dumolin-Lapègue *et al.* (1999), Nowakowska (2011), Nowakowska *et al.* (2015), Tereba *et al.* (2017).

2.2.2.2 SNPs MARKERS TO IDENTIFY SEIZED WOOD

Microsatellites are difficult to amplify from timber due to the low preservation quality of the DNA therein (Jolivet & Degen 2012). One way to overcome the difficulty of amplifying microsatellites from degraded DNA is to use Single Nucleotide Polymorphisms (SNPs) or INsertions/DELetions (INDELs). SNPs are single base substitutions in the genome, while INDELs are base insertions or deletions in the genome. Such markers contain on average less information on population structure than microsatellites, but are far more frequent, and are **easier to analyse from degraded DNA**. In addition, they can be studied using automated high-throughput platforms, that allow analysis of between a few and tens of thousands of SNPs. They can be either nuclear (nSNPs), mitochondrial (mtSNPs), or located in the chloroplast (cpSNPs), each with different properties. nSNPs are generally diploid (except in polyploid species), while plastid SNPs are haploid.

SNPs need to be developed *de novo* for each species or group of closely related species, which can be a lengthy process. However, once developed, they can be genotyped relatively easily. SNPs can be identified, for example, by low coverage sequencing of *e.g.* 10 individuals and mapping the reads to a reference genome to detect polymorphisms. If a reference genome for the species is not available, a low coverage draft genome can be generated by *de novo* assembly of the sequenced reads of one of the individuals sequenced. Using this methodology, **sets of SNPs have been developed for different valuable timber species** and have been used for (i) species identification in the genera *Hymenaea* (Chaves *et al.* 2018), *Dipteryx* (Honorio Coronado *et al.* 2019) and *Cedrela* (Paredes-Villanueva *et al.* 2019), (ii) provenance determination in *Carapa* spp. (Tysklind *et al.* 2019), *Dalbergia* spp. (Hartvig *et al.* 2020) and *Larix* spp. (Blanc-Jolivet *et al.* 2018), and (iii) individualisation of timber from bigleaf maple (Dormontt *et al.* 2020).

The remainder of this guide will focus on the use of SNPs for timber traceability.

2.3 DEVELOPMENT OF SNP GENETIC MARKERS

2.3.1 SAMPLING DESIGN

Given the costs of high throughput sequencing (HTS) methods, the number of individuals that can be included at the SNPs development stage is generally quite limited. Typically, less than 10 individuals are used, although costs are decreasing and up to 96 individuals could be included at this stage in future analyses. The **appropriate choice of individuals** to be included at this stage is critical to obtain genetic markers that achieve the desired effect (*e.g.* either species identification, or within species identification, or provenance determination). Depending on the genomic resources available for the target species/provenance, there are several approaches. Two decision trees are presented, one for species identification and another for identification of provenance (see Appendix 2, §7.2.1).

2.3.1.1 FOR SPECIES IDENTIFICATION

The sampling design should include individuals from all potential confounding species, such as congeners, to maximise the chances of identifying species-diagnostic SNPs. Several individuals per species (*e.g.* 2 or 3) should be included, and particularly, samples covering most of the focus species distribution range. This should avoid confounding effects of species identity and geography. The botanical identification of the individuals included here should be confirmed by a specialist of the taxa involved. Pictures of leaves, trunk, and flowers (if present) should be taken, and voucher specimens should be stored and, ideally, sent to herbaria (see also the General sampling guide for timber tracking).

2.3.1.2 FOR IDENTIFICATION OF PROVENANCE

The sampling design should aim to include individuals collected across the species range, including the most extreme locations, to maximise the chances of identifying SNPs that may show geographical differences in allele frequencies. It is desirable, to use more than one individual in locations where the species is abundant, or when a precise identification of provenance is to be done (*e.g.* concession level). Again, the species identification of individuals included here should be confirmed by a botanist and pictures and voucher specimens should be archived.

2.3.1.3 COMBINED APPROACH

Both types of markers, for species and provenance identification, may be developed simultaneously (i) by incorporating reference samples that include several congeneric species and (ii) by sampling locations spanning the whole genus distribution (Meyer-Sand *et al.* 2018; Honorio Coronado *et al.* 2019; Tysklind *et al.* 2019).

2.3.2 SNP DEVELOPMENT

The use of SNPs for species and provenance identification involves the identification of a sufficiently informative set of markers and the validation of those markers. For the development of candidate SNP markers, choose one or several HTS methods[23], which will allow you to identify substantial SNP variation among your samples in both types of genomes (*i.e.* nuclear and plastid). There might also be reference genomes available for some of your species of interest, or a closely related species, which might help you in the analysis (see also Appendix 2, §7.2.1). After bioinformatic analysis of the produced DNA sequences, calls for all individuals at the putative SNP loci are provided together with sequence alignments. **Loci for which one allele is fixed within a taxon and the other allele fixed, or almost fixed, in the other taxon** are good genetic marker candidates for discrimination among the two taxa (*i.e.* species identification). **Loci which are polymorphic within a species but have allele frequencies varying strongly among sampling locations** are good genetic marker candidates for the identification of provenance. The use of markers targeting unlinked loci, when possible, increases the discriminant power of the analysis.

2.3.3 SNP VALIDATION

The last step is to screen a larger number of individuals at the identified candidate loci to **validate the markers** using the genotyping method of your choice (*e.g.* Sanger sequencing, KASP, MassARRAY, microarrays, targeted sequencing, ...) (Honorio Coronado *et al.* 2020). We recommend using **several markers with redundant information** in the final set of loci, in case a rare variant is detected on a test sample.

[23] There are many HTS methods available on the market (*e.g.* Sucher *et al.* 2012, Türktas *et al.* 2015).

2.4 CONSTRUCTION OF A GENETIC BASELINE REFERENCE DATABASE

2.4.1 COLLECTION OF VOUCHER SPECIMENS AND REFERENCE SAMPLES

For the development of reference data, only use **fresh plant material**, collected as advised in the GTTN sampling guide. Only plant material from herbarium vouchers should be used for marker validation. Collect a set of reference samples covering **all species of (potential[24]) interest**. For each species, multiple individuals, **representative of the respective species distribution area or/and representing the different populations,** should be collected (see also Appendix 4). This will allow to identify diagnostic genetic loci at both levels, species and population (*i.e.* geographic provenance).

2.4.2 DATA CHECKING

It is of paramount importance to ensure the quality of the input data before doing analyses. The quality checking step is normally a reiterative process where the data goes through several rounds of data checking from several angles to **identify and quantify data of dubious quality and missing data**. How such data is dealt with will depend on whether the sources of such errors/missing data can be clearly identified, prior knowledge of population structure, loci characteristics, sample preservation history, and on the purposes of the database. The incorporation of incorrect data can grossly bias the reference database and flaw the individual assignment process. It is normally recommended to err on the side of caution and remove dubious data. The researcher should take informed decisions, be consistent, and **keep a record** of all data removed from the analysis.

See Appendix 2 for extended info on genetic data checking.

[24] Species of interest can change depending on remaining stands. Once a species has been extensively logged in one region, the focus might go to the next species that is qualitatively 'next in line' in that region, even if that species has not been of interest before.

BOX 7: EFFECT OF DATA CLEANING

Checking the data for technical errors is an essential step to obtain quality data, and the sources of dubious data should be identified. However, there are some situations where missing data could provide biologically meaningful information.

Balancing the sampling design might be recommended if there are underrepresented, true clusters (*i.e.* taxa, species, populations). The question is, however, how to select the individuals without losing potential cryptic groups. Uneven sampling is very frequent in tropical species due to uneven species abundance and sampling location accessibility.

There are different views on the need to **remove loci with a lot of missing data**. The advantage of removing them is improving the overall quality of the data. The advantage of keeping the complete dataset is that there might be informative polymorphisms in some clusters that may be missing in other clusters (*e.g.* null alleles that are specific to certain species/populations). This happens often when two related species are analysed with the same marker set or when two clusters are vastly different. The risk of removing such loci is the loss of distinguishing power among clusters.

General advice: Be consistent and keep a record of all data removed from the analysis.

2.4.3 DATA EXPLORATION

- ☑ Familiarise yourself with the data: number of individuals, samples, loci, and alleles; samples sizes, alleles per locus, *etc.* Explore the expected and observed heterozygosities, and the proportion of loci out of Hardy-Weinberg Equilibrium (HWE). Do not remove loci out of HWE.
- ☑ Identify data of dubious quality/origin and missing data.

Multivariate analysis (MVA) of genetic data (*e.g. adegenet*, see Table 2-3) is a good way to first approach a genetic dataset and **getting a quick glance at the structure and strength of signal** (*i.e.* population genetic structure) in the dataset. MVA does not make any genetic assumptions (*i.e.* HWE and LD) on the data and looks at the similarity in allele frequencies among individuals or sampling locations. It is not too computationally intensive and has a rich and flexible variety of graphical displays to represent the structure of the genetic diversity in the dataset. Given that it is based on MVA, *adegenet* offers a complementary analysis to other commonly used genetic analyses methods, such as *STRUCTURE* or *Geneland*.

See Appendix 2 for extended info on genetic analyses.

Table 3: Evaluation of some of the most used software that can be used for genetic data exploration.

Software	Advantages	Disadvantages
adegenet	Useful for data representation (genetic data and associated metadata), exploratory data analysis, and thorough genetic analysis and statistical modelling. Copes with both small datasets and large SNP datasets. MVA does not rely on genetic assumptions (*e.g.* HWE, LD) and can handle missing data by replacing missing values with observed means, so substandard quality genetic data may still produce interpretable and biologically meaningful results. Allows estimation of assignment probability of individuals to populations. Not computationally intensive. *R* scripts are reproducible and publishable.	Requires basic *R* and statistical knowledge.
GDA-NT	Locus quality function especially useful to look at locus performance (differentiation, diversity, F-value and correlation among genetic and geographic distances).	The Locus function is only available for test samples. Impractical data format.
poppr	Copes with large SNP datasets. Identifies and allows removal of (i) low genotyping success individuals and loci, (ii) uninformative loci, (iii) duplicated samples and/or clones. *R* scripts are reproducible and publishable.	Requires basic *R* and statistical knowledge.

2.4.4 ASSEMBLY AND ANALYSIS OF THE GENETIC BASELINE REFERENCE DATABASE

2.4.4.1 CLUSTERING DATA

Clustering (a.k.a. grouping) is done to determine which individuals belong to the same cluster (*i.e.* species group, species, or population, depending on the aim) or to determine admixture proportions (*e.g.* to identify hybrids of different species, or descendants from migrants from other populations). The analyses can be carried out for different **assumed numbers of clusters, K** (*i.e.* the estimated number of species or populations). There are **two types of statistical methods to assign samples to clusters**.

The first are methods based on population genetics models assuming HWE and LD (*e.g. STRUCTURE*, *Geneland*). There are different types of models (*e.g.* no admixture/admixture, correlation of allele frequencies or not, prior population information or not).

Knowledge on the biology of the organism is key here.

- The no admixture and uncorrelated allele frequencies model should be used if the allele frequencies are expected to be quite different (*e.g.* different congeneric species).
- The admixture model with correlated allele frequencies is recommended if closely related populations need to be detected (*e.g.* populations frequently exchanging migrants and/or if two clusters originated through isolation-by-distance).

Overall, it is recommended to start with the uncorrelated model and then assess the effects of changing to a correlated model. A combination of both may be applied in a nested mode: first identifying species with the no admixture and uncorrelated allele frequencies models, and then identifying populations within each of the species with the admixture and correlated allele frequencies.

The second type of statistical methods is genetic assumption-free methods (*e.g.* multivariate analyses, linear or kernel discriminant analysis, K nearest neighbour, multinomial models, partial least squares classification, decision trees such as recursive partitioning, neural networks, *random forest*, *etc.*).

2.4.4.2 IDENTIFYING THE NUMBER OF GROUPS

Choosing the optimal K-value is often misapplied (Janes *et al.* 2017). Various factors determine which clustering solutions are **biologically meaningful** and should be further inspected:

- The likelihoods of different clustering solutions.
- The level of hierarchy in the data.
- The scope of the analysis.

It is never advised to just interpret results for a single value of K. Also, wild organisms are rarely truly clustered into an exact number of populations, and the most biologically sensible interpretation of the data may be a range of values (*e.g.* from 7 to 10) to accommodate for isolation by distance and geneflow between populations.

The most widely used methods to compare the likelihoods are the deltaK (ΔK) method and the likelihood method implemented in *STRUCTURE Harvester* (Earl & vonHoldt 2012).

- The deltaK method is **useful to inspect the higher level of hierarchy** present in the data. The deltaK method might underestimate the true number of distinct populations if there is population structure hierarchy within a species.
- The likelihood method inspects the likelihood versus K plot, which often reaches a plateau at a certain number of K. Above that number of K, clustering solutions are likely not biologically meaningful (but can still be inspected). In **datasets with many closely related species and populations**, low-number clustering solutions (small K) will show the subdivision at a certain hierarchical level, such as among the closely related species, while not showing the fine scale structure among populations.

CAUTION: The distinction between populations of the same species and different species is subjective and case specific, since it depends on the taxonomy, the species concept applied for the taxa, and the purpose of the analysis (*e.g.* traceability, management, invasive species, conservation).

2.4.4.3 ESTIMATING BASIC POPULATION GENETICS STATISTICS

Once the species or populations (clusters) have been delimited by the above methods (DAPC, *STRUCTURE*, *Geneland*), basic population genetic statistics should be estimated, such as allele richness (AR), observed and expected Heterozygosity (H_O and H_E), deviations from HWE, and LD. These parameters and genetic differentiation among the populations (*e.g.* F_{ST}, G_{ST}, D_{Nei}, D_{Jost}, D_J) can then be estimated through a wide range of software (*e.g. GDA-NT*, *GenePoP*, *GenAlEx*, *DiveRsity* (*R* package)). Which estimator of differentiation to use will depend on the nature of the genetic marker used and the allele diversity in the system.

2.4.4.4 EVALUATING ASSIGNMENT POWER

Once the genetic baseline reference database has been defined, delimiting which individuals and population samples belong to which panmictic populations (clusters), the strength and reliability of the dataset needs to be assessed. The grouping is then tested for its power to correctly assign an individual to its species/population of provenance.

There are several **ways of doing this**, which normally involve:

- removing one or more individuals and/or loci from the baseline
- creating a new reference database without the removed individuals/loci
- assessing how many times the individual is correctly assigned to its species/population of provenance.

This procedure is automated in several programmes (*e.g. Geneclass2*, *Oncor*, *assignPop*). If the individuals are correctly reassigned to their original species/population, the reference database is powerful and accurate. The accuracy of individual assignment is specific to loci and populations. Some loci do not carry any relevant information, while others will be highly differentiated among clusters. Some populations will have unique genetic signals, while others will be quite similar to neighbouring populations. **How we define the clusters therefore impacts strongly on the testing** of the reference database and its performance for individual assignment.

A robust genetic baseline reference database should still perform well when we remove significant amounts of data from it (*e.g.* either individuals or loci).

2.4.4.5 SELECTION OF DIAGNOSTIC MARKERS

Several population genetics analysis programmes (*e.g. poppr*, *assignPop*) allow the identification and removal of **uninformative loci** at this stage. The cut-off value can be either set manually or set as a default (*i.e.* fewer than x variant observations in the whole dataset, or MAF < 0.01). It is worth running this analysis to check the proportion of uninformative loci in the dataset, even if the uninformative loci are not subsequently removed, to verify there are no problems with the genotyping.

Similarly, once the population genetic structure has been established, several population genetics analysis programmes (*e.g. assignPop*, *GDA-NT*, *adegenet*) allow the identification of the **most informative loci** (*i.e.* top 10 SNPs, top 25% SNPs) to detect such structure. These loci (a.k.a. **golden markers**) can then be incorporated into genotyping assays for the assignment of unknown provenance test samples.

2.5 ANALYSIS OF TEST SAMPLES

2.5.1 GENOTYPING TEST SAMPLES

Test samples always need to be analysed at least twice (*i.e.* two independent DNA extractions), better in triplicate, and it is often useful to conduct PCR with different amounts of DNA template. The consensus genotype will then be used in the assignment.

2.5.2 INDIVIDUAL ASSIGNMENT TO GENETIC BASELINE

A wide range of programmes can be used for the assignment of the test samples (*assignPop, Oncor, Geneclass2, adegenet, STRUCTURE, GeoAssign*), each based on different methodologies and characteristics of the data. The consensus genotype can be assigned to either species or populations of the validated reference database. Contrasting the results of several programmes, based on different hypothesis, can help strengthen the inferred species or provenance identification.

Two different case scenarios are possible: the sample is provided with a claimed provenance (*i.e.* a species or a geographical location), or the provenance is totally unknow.

When a claimed provenance is challenged, then only the likelihood and eventually exclusion probability of the test sample belonging to the claimed species/provenance should be reported. This can be tricky depending how the genetic baseline reference database is genetically structured. **Ideally, all potential provenances are genetically well differentiated and are clustered separately in the genetic baseline reference database**, therefore allowing assignment to a single provenance/cluster. However, if 1) the sampling size at the provenance used to construct the baseline is low, 2) several provenances are genetically close, and/or 3) the claimed provenance has a genetic discontinuity (*e.g.* two or more distinct genetic clusters are present at the same location), then the genetic baseline reference database might be based on the genetic clusters. In this case, the test individual should be tested against each of the genetic clusters, which are known to occur in the claimed provenance. **Using genetic clusters based on plastid markers, nuclear markers, as well as a combination of both, is desirable** as the genetic structure might differ among nuclear and plastid genomes. If several independent analyses on different type of markers provide the same answer, the test can be considered as secure. However, if both analyses provide different results, each case should be examined carefully together with quality indicators on the genetic baseline reference database for the specific markers used and the claimed provenance.

When test samples do not have a claimed provenance, then one has to compare the likelihoods of the test sample belonging to all genetic clusters in the genetic baseline reference database. **Some individuals may carry alleles that are very characteristic** of a certain genetic cluster, and we may get a high probability of belonging to that genetic cluster while the probability of belonging to other genetic clusters is low. In this case, we can report the provenance with a good degree of certitude. However, **many individuals will carry alleles that are common** across many genetic clusters, and for such individuals, we unfortunately will not be able to infer provenance with certitude.

2.6 STRENGTHS & LIMITATIONS

Strengths

- Level of resolution. Most timber species can be identified with cost-effective and reliable methods, such as wood anatomy, down to the genus level. However, **identifying closely related congeneric species** or **determining the geographic provenance of an individual tree** is typically not possible. In such cases, genetic methods can be extremely powerful, especially if high throughput sequencing data is available.
- Species delimitation. **New/unexpected species can be detected** during the construction phase of reference datasets. This is particularly relevant in species-rich but understudied regions such as the tropics.
- Conservation-relevant data. Owing to the large body of population biology theory, DNA data can be used to **estimate demographic parameters** such as population size, levels of genetic diversity, or levels of inbreeding or introgression with other species.

Limitations

- The absence of good quality DNA in the test samples is the main limitation of DNA-based wood identifications. Cambium (or leaves) is the most suitable source of good quality DNA (Nowakowska 2011) and is therefore used to develop reference data (see the GTTN sampling guide). However, except for recently seized wood material (see also §2.2.2.1 *Microsatellite markers to identify seized wood*), **most requests for species and provenance identification concern wood that is not fresh anymore**. When lower quality DNA is amplified, polymerase errors can result in artefacts (Dumolin-Lapègue *et al.* 1999; Asif & Cannon 2005). Moreover, the DNA amplification process can simply be inhibited by high amounts of residuals of polysaccharides and polyphenolic compounds (Tibbits *et al.* 2006).
- High costs, depending on the methods used, especially when only a single or a few samples need to be tested using high throughput genotyping technologies.
- Time-consuming analysis depending on the methods used.
- Evolutionary processes are at the base of the analysis of genetic markers, and the **signals found may not correspond to the desired resolution** (*e.g.* population shared across national borders, several biological species exploited under a single commercial name).

2.7 KEY LITERATURE FOR GENETIC DATA ANALYSIS

Asif, M.J. and C.H. Cannon (2005). DNA extraction from processed wood: a case study for the identification of an endangered timber species (Gonystylus bancanus). Plant Mol Biol Reporter 23: 185–192.

Blanc-Jolivet, C., Y. Yanbaev, B. Kersten et al. (2018). A set of SNP markers for timber tracking of *Larix* spp. in Europe and Russia. Forestry 91(5).

Caron, H., J.F. Molino, D. Sabatier *et al.* (2019). Chloroplast DNA variation in a hyperdiverse tropical tree community. Ecology and Evolution 9: 4897–4905.

Chaves, C., C. Blanc-Jolivet, A. Sebbenn *et al.* (2018). Nuclear and chloroplastic SNP markers for genetic studies of timber origin for Hymenaea trees. *Conservation Genetics Resources* 11: 329-331.

Chen, K.Y., E.A. Marschall, M.G. Sovic *et al.* (2018). *assignPop*: An R package for population assignment using genetic, non-genetic, or integrated data in a machine-learning framework. Methods in Ecology and Evolution 9(2).

Dormontt, E.E., D.I. Jardine, K.J. van Dijk *et al.* (2020). Forensic validation of a SNP and INDEL panel for individualisation of timber from bigleaf maple (*Acer macrophyllum* Pursch). Forensic Science International: Genetics 46.

Dumolin-Lapègue, S., M.H. Pemonge, L. Gielly, P. et al. (1999). Amplification of oak DNA from ancient and modern wood. Mol Ecol 8:2137–2140.

Earl, D.A., B.M. vonHoldt (2012). *STRUCTURE* HARVESTER: A website and program for visualizing *STRUCTURE* output and implementing the Evanno method. Conservation Genetics Resources 4: 359-361.

Guillot, G., F. Santos, A. Estoup (2008) Analysing georeferenced population genetics data with *Geneland*: a new algorithm to deal with null alleles and a friendly graphical user interface. Bioinformatics (Oxford, England) 24: 1406-7.

Hartvig, I., T. So, S. Changtragoon *et al.* (2020). Conservation genetics of the critically endangered Siamese rosewood (*Dalbergia cochinchinensis*): recommendations for management and sustainable use. Conservation Genetics.

Hollingsworth, P.M. (2011). Refining the DNA barcode for land plants. *Proceedings of the National Academy of Sciences* 108: 19451-19452.

Honorio Coronado, E.N., C. Blanc-Jolivet, M. Mader *et al.* (2019). Development of nuclear and plastid SNP markers for genetic studies of *Dipteryx* tree species in Amazonia. Conservation Genetics Resources 11:333–336.

Honorio Coronado, E.N, C. Blanc-Jolivet, M. Mader *et al.* (2020). SNP markers as a successful molecular tool for assessing species identity and geographic origin of trees in the economically important South American legume genus *Dipteryx*. Journal of Heredity, accepted.

Jakobsson, M., N.A. Rosenberg (2007). *CLUMPP*: A cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics 23(14): 1801-6.

Janes, J.K., J.M. Miller, J.R. Dupuis *et al.* (2017). The K=2 conundrum. Molecular Ecology 26(14): 3594-3602.

Jiao, L.C., T. He, E.E. Dormontt *et al.* (2019). Applicability of chloroplast DNA barcodes for wood identification between Santalum album and its adulterants. Holzforschung 73(2): 209-218.

Jolivet, C., B. Degen (2012). Use of DNA fingerprints to control the origin of sapelli timber (*Entandrophragma cylindricum*) at the forest concession level in Cameroon. Forensic Science International: Genetics 6: 487-493.

Li, X., Y. Yang, R.J. Henry *et al.* (2015). Plant DNA barcoding: from gene to genome. Biological reviews of the Cambridge Philosophical Society 90(1): 157-66.

Meyer-Sand, B.R.V., C. Blanc-Jolivet, M. Mader *et al.* (2018). Development of a set of SNP markers for population genetics studies of Ipe (*Handroanthus* sp.), a valuable tree genus from Latin America. Conservation Genetics Resources 10: 779–781.

Nowakowska J.A. (2011). Application of DNA markers against illegal logging as a new tool for the Forest Guard Service. Folia Forestalia Polonica, series A 53(2): 142–149.

Nowakowska, J.A., T. Oszako, A. Tereba and A. Konecka (2015). Forest tree species traced with a DNA-based proof for illegal logging case in Poland. In Evolutionary biology: biodiversification from genotype to phenotype (pp. 373-388). Springer, Cham.

Paredes-Villanueva, K., C. Blanc-Jolivet, M. Mader *et al.* (2019) Nuclear and plastid SNP markers for tracing Cedrela timber in the tropics. Conservation Genetics Resources 12: 239-244

Piry, S., A. Alapetite, J.M. Cornuet *et al.* (2004). GENECLASS2: A software for genetic assignment and first-generation migrant detection. Journal of Heredity: 95: 536-539.

Pritchard, J.K., M. Stephens, P. Donnelly (2000). Inference of Population Structure Using Multilocus Genotype Data. Genetics 155(2): 945-959.

Saéz-Laguna, E., M.Á. Guevara, L.M. Diáz *et al.* (2014). Epigenetic variability in the genetically uniform forest tree species *Pinus pinea* L. PLoS ONE 9(8): e103145.

Sucher, N.J., J.R. Hennell and M.C. Carles (2012). DNA Fingerprinting, DNA Barcoding, and Next Generation Sequencing Technology in Plants. In: Sucher N., Hennell J., Carles M. (eds) Plant DNA Fingerprinting and Barcoding. Methods in Molecular Biology (Methods and Protocols), vol 862. Humana Press.

Tereba, A., S. Woodward, A. Konecka et al. (2017). Analysis of DNA profiles of ash (*Fraxinus excelsior* L.) to provide evidence of illegal logging. Wood Science and Technology 51: 1377–1387.

Tibbits, J.F.G., L.J. McManus, A.V. Spokevicius and G. Bossinger (2006). A rapid method for tissue collection and high throughput of genomic DNA from mature trees. Plant Mol Biol Reporter 24:1–11.

Türktaş, M., K.Y. Kurtoğlu, G. Dorado, B. et al. (2015). Sequencing of plant genomes? A review. Turkish Journal of Agriculture and Forestry 39(3): 361-376.

Tysklind, N., C. Blanc-Jolivet, M. Mader *et al.* (2019). Development of nuclear and plastid SNP and INDEL markers for population genetic studies and timber traceability of Carapa species. Conservation Genetics Resources 11: 337–339.

3. Stable isotopes

Definition isotopic reference data: values of the relative abundance of two stable isotopes, of one or several elements, that are characteristic for the wood of a specific taxon grown at a specific locality.

Authors: Laura Boeschoten, Micha Horacek, Kathelyn Paredes-Villanueva, Gareth Rees, Mart Vlam, Charlie Watkinson, Pieter Zuidema

*Authors are in alphabetical order.

Fig. 10: A view inside a stable isotope lab.

3.1 RESOURCES REQUIRED

Access to data sources

For isotope analyses international standard materials are required for correlation and comparison of data with (published) data from other laboratories. These **international standard materials** can be purchased from *e.g.*:

- the U.S Geological Survey (USGS, Reston-USA)
- the National Institute of Standards and Technology (NIST, Gaithersburg-USA)[25]
- the International Atomic Energy Agency (IAEA, Vienna-Austria)
- the European Commission's Joint Research Centre (JRC, Geel-Belgium)

In addition, to investigate the geographic origin (provenance) of a wood sample access to **specific databases** such as the ones below can be useful.

- Isotope data: Waterisotopes.org
- Environmental data: NASA earth science data, Climate Research Unit at the University of East Anglia

Software

Table 4: Overview of the software used for the analysis of stable isotope data for wood identification.

Software	Application	Accessibility
ARC GIS	Geographic information systems tools, isoscape visualisations and analysis, Voronoi diagrams, triangulated irregular networks, inverse distance weighting, Kriging, raster statistics, PCA	Commercial software
QGIS		Free software
ArDB (Elementar)	Isoscape visualisations, Linear Discriminant Analysis (LDA), Principle Component Analysis (PCA), K-Means Clustering, Calculated Columns (*R*), Quality Control Statistics	Commercial software
JMP	ANOVA	Commercial software
R	Can be used for any statistical computing or imaging task.	Free software
SPSS (IBM)	Discriminant analysis, PCA, multiple processing options.	Commercial software

[25] See also this project on High-Precision Isotopic Reference Materials.

3.2 USING STABLE ISOTOPES FOR ORIGIN IDENTIFICATION

3.2.1 DEVELOPING STABLE ISOTOPE REFERENCE DATA

BOX 8: WHAT DO WE MEAN WITH STABLE ISOTOPE REFERENCE DATA?

There is a difference between international standards and geographic reference data for isotopes. **International standards** are materials possessing exactly known isotope ratios of certain elements. **Reference data** are standardised by the international standard and are characteristic values of isotope ratios of certain elements for a geographical locality.

Stable isotope ratios for the elements H, C, N, O, S are usually reported in the conventional delta (δ) notation as a deviation from international standards, as described in this equation:

$$\delta = \frac{(heavy\ isotope/light\ isotope)_{Sample} - (heavy\ isotope/light\ isotope)_{Standard}}{(heavy\ isotope/light\ isotope)_{Standard}} \cdot 1000$$

Any unit cancels in this equation. However, to give a sense of scale, measurements are reported in permille (‰). Where isotope ratio values are positive, there is relatively more 'heavy isotope' in the sample than the standard, where values are negative there is relatively less 'heavy isotope' in the sample than the international standard.

The international standards used (see *§3.1 Resources required* for points of sale):

Stable isotope ratio	International standard used for analysis
D/H (hydrogen)	Vienna Standard Mean Ocean water (VSMOW2)
$^{13}C/^{12}C$ (carbon)	Vienna Pee Dee Belemnite (VPDB)
$^{15}N/^{14}N$ (nitrogen)	Atmospheric Nitrogen (N_{2Air})
$^{18}O/^{16}O$ (oxygen)	Vienna Standard Mean Ocean water (VSMOW2)
$^{34}S/^{32}S$ (sulphur)	Vienna Canyon Diablo Troilite (VCDT)
$^{87}Sr/^{86}Sr$ (strontium)	Seawater (NBS-987 standard)

3.2.1.1 SAMPLE COLLECTION

TIP: see also the GTTN sampling guide and Appendix 4.

☑ **Select the correct sample material** for the question at hand. Usually, wood material is to be selected (sapwood, heartwood). However, for the investigation of living or freshly cut trees, for example, other sample material (needles, leaves, bark, cambium) might be more suitable (see *e.g.* Horacek 2012). Always compare samples of the same material. If you have to mix, the difference in isotope fractionation between the different tissues needs to be known.

☑ **Prevent contamination** along the sampling process. For example, remove any (pencil) markings on the wood as they can interfere with the $\delta^{13}C$, or other isotope ratios.

☑ **Integrate spatial and temporal variation:**

Spatial heterogeneity in the isotope signal. If a species is light demanding, it may present a different $\delta^{13}C$ composition when samples are collected in an undisturbed forest or at a clearing. Variance in $\delta^{13}C$ may also be related to differences in vapour pressure, which is often a co-variate of radiation (Farquhar *et al.* 1989). Also, the $\delta^{18}O$ signal may differ depending on the primary source of water: precipitation, humidity, soil moisture or groundwater (Aggarwal *et al.* 2004).

➤ **Understand the ecophysiology of the taxon and sample at least 25 trees** per area of interest to include spatial variation.

Temporal heterogeneity in the isotope signal. Most isotopic values vary over time, between years and within a year. In addition, there are long term trends, such as the increasing CO_2 concentrations since industrialisation, which can affect the isotopic signal (Broadmeadow *et al.* 1992; McCarroll & Loader, 2004). It is therefore required to know the approximate date/age of the sample, if old wood (before industrialisation) is investigated, to correct for possible effects of the decrease in the amount of $\delta^{13}C$ in the atmosphere from -6.4‰ to *ca.* -8‰ since industrialisation (McCarroll & Loader 2004, van der Sleen *et al.* 2017).

➤ **Include as much information as possible** by sampling over a radial distance of at least 8 growth rings or 10 cm and by either mixing the wood material before measurement, or by processing each wood sample individually and later averaging the measurements.

BOX 9: HOW TO INTERPRET ISOTOPE VARIATION?

The size of the area in which the stable isotope ratios in the wood might be relatively constant depends on the variability of geographical factors such as topography, water resources, climate, and soil, to name just a few. Therefore, to ensure statistical analysis results are not over-interpreted, it is **worth considering how variable isotopes might be within a given eco/geo-system**. In trees, the carbon, oxygen and hydrogen stable isotope ratios vary both within single rings and between different trees of the same species on the same sampling site (Fig. 11). A review by Leavitt (2010) gives values for inter-stump and intra-ring variance for *Pinus* spp., *Phyllocladus* spp., *Abies* spp., *Quercus* spp., *Cryptomeria* spp., *Athrotaxis* spp., *Cedrus* spp., *Fagus* spp., and *Juniperus* spp. These values for stable isotope variance on any given site are also comparable with the review of tropical tree rings by van der Sleen *et al.* (2017). This suggests these values may potentially represent common values for site variance across different species, but these intra-species intra-site variations most certainly also depend on the specific site conditions. Very heterogeneous environments (*e.g.* strongly variable soil depth, shading) result in strong variations.

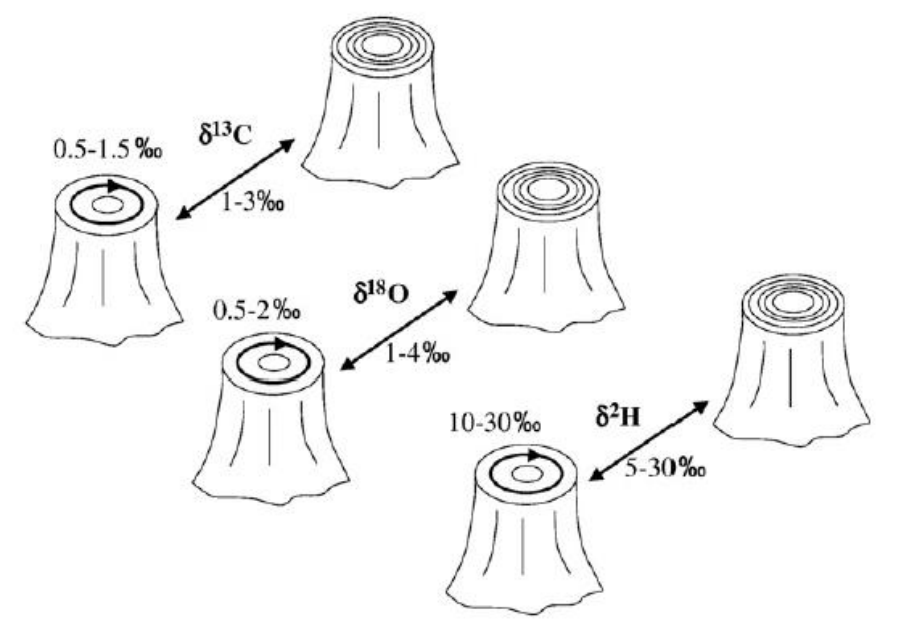


Fig. 11: Variability of stable isotope data in trees according to Leavitt (2010): "Summary of typical intra-tree (circumferential range of isotopic values indicated adjacent to stumps) and inter-tree (values adjacent to 2-sided arrows) variability for δ^2H, $\delta^{13}C$, and $\delta^{18}O$ as reported in the literature."

3.2.1.2 SAMPLE PREPARATION & ANALYSIS

When measuring the stable isotope profile of wood, it is important to **consider which elements' stable isotopes** might differ between the studied locations. The different element isotope ratios depend on different chemical, physical and biological processes (see Appendix 3), therefore some might be more informative than others in different situations (see Horacek *et al.* 2018). For example, certain stable isotopes are better suited for large-scale differentiation of provenance (*e.g.* H, C, O in western Central Africa for Tali timber, Vlam *et al.* 2018b), while others vary sometimes more at local scales (Sr, S, N in some cases) (English *et al.* 2001).

Wood comprises several components including alpha cellulose, lignin, amino acids, and water. These components are not equally distributed within a tree itself or between different species. Moreover, lignin and cellulose are produced by different metabolic processes resulting in slightly different isotopic profiles. Therefore, **cellulose extraction**[26] prior to hydrogen, oxygen and carbon isotope analysis is common (DeNiro & Epstein 1979, Rinne *et al.* 2005, Boner *et al.* 2007, Vlam *et al.* 2018). It is however more time and labour intensive and **whole-wood analysis** can provide results of the same quality (see Horacek *et al.* 2009, Gori *et al.* 2013). For the analysis of nitrogen and sulphur isotopes cellulose extraction needs to be combined with extra purification procedures to produce a measurable signal in the isotope ratio mass spectrometer. Alternatively, a bulk wood analysis needs to be performed.

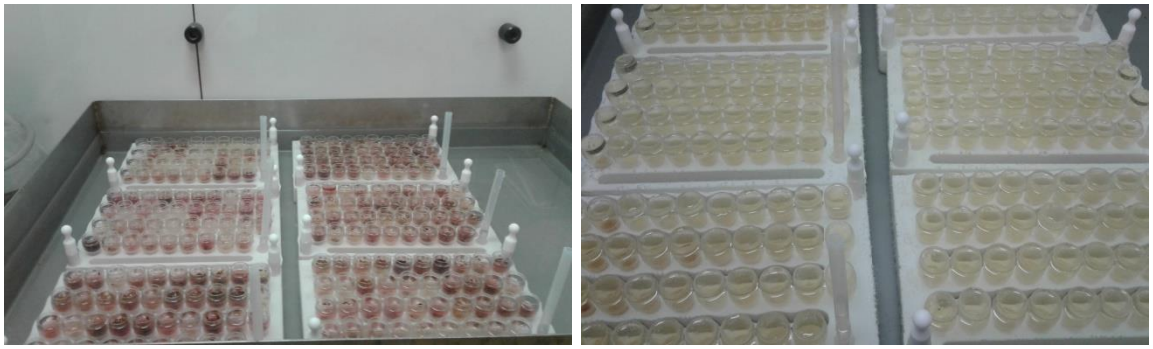


Fig. 12: Cellulose extraction of wood samples at the start (left) and the end (right) of the chemical treatment.

[26] There is a plethora of cellulose purification protocols available in literature, many for industrial applications. Researchers should take care to ensure their procedure is suitable for stable isotope analysis.

EXTRACTION OR NOT: Both methods deliver **good reference datasets**. However, cellulose stable isotope data are not easily compared to whole-wood stable isotope data. So, there is a strong preference for a reference database to contain data obtained from a single wood preparation method. If this is not possible, then at least the offset between cellulose-based and whole-wood based isotopic measurements should be quantified for the same samples.

Exchangeable hydrogen in wood needs to be accounted for or eliminated. Hydrogen in cellulose is known to exchange readily with hydrogen in atmospheric water. To maintain measurement precision, the proportion of exchangeable hydrogen in the sample should be accounted for or be eliminated. For timber cellulose, two commonly accepted approaches are nitration (Deniro *et al.* 1981) and water vapour equilibration (Sauer *et al.* 2009).

Avoid and generously remove **altered and moulding samples**, as microbial processes might alter the primary isotope pattern (see Horacek *et al.* 2018).

3.2.1.3 DATA PREPARATION

Before statistics are applied to help interpret data, it is always good practice to **understand what the analytical data means** and where the values come from. See Appendix 3 for more info.

Clean and homogenise stable isotope data by removing measurement errors and outliers that cannot be due to geographic origin (provenance). For example, if most samples were taken from medium sized trees, very small and very big trees can give aberrant data. Also, the height of the tree where samples were taken can result in data variation (Brienen *et al.* 2017).

Calculate mean values per tree from the duplicates (if using duplicate/triplicate samples per tree).

Check your data distribution and if not normal, use appropriate statistics. When data are normally distributed (Baxter 1999), a **(pre-)analysis**[27] can be done using scatterplots with confidence ellipsoids for discrimination among sites.

[27] This analysis can be done as an alternative to modelling, especially, if only a limited dataset is available.

3.2.1.4 VISUALISATION OF SPATIAL STABLE ISOTOPE PATTERNS VIA ISOSCAPES

WHEN TO USE? Isoscapes, a portmanteau of 'isotopic' and 'landscapes', are statistical maps of stable isotope ratios. The ambition of using isoscapes is to **look beyond what the reference data describes** to overcome the limitations of entirely data-driven tests such as Discriminant Analysis. Isoscape analysis is suitable for **large-scale differentiation** (*e.g.* continents, countries, and regions) when there is a wide distribution of reference samples in low density over the area of interest. To evaluate isotope information over short distances, the suitability of isoscapes depends on the available data. If the dataset is sufficiently detailed, it can also be used to evaluate **local variations**, such as variations in individual forest management units.

There are **as many ways to generate isoscapes as there are to generate maps**. Example methods can be regression of sample data with environmental data, Inverse Distance Weighting, Triangular Irregular Networks, Voronoi diagrams, and the various different forms of Kriging/co-Kriging. Some isoscapes are made by simply applying distance weighting to reference data, some are more complex models that incorporate climatic information with the reference data and some are entirely theoretically based models made using equations that are related to the physical fractionation of massive objects (*e.g.* Roden *et al.* 2000 created a theoretical model for hydrogen and oxygen isotope ratios in tree-ring cellulose).

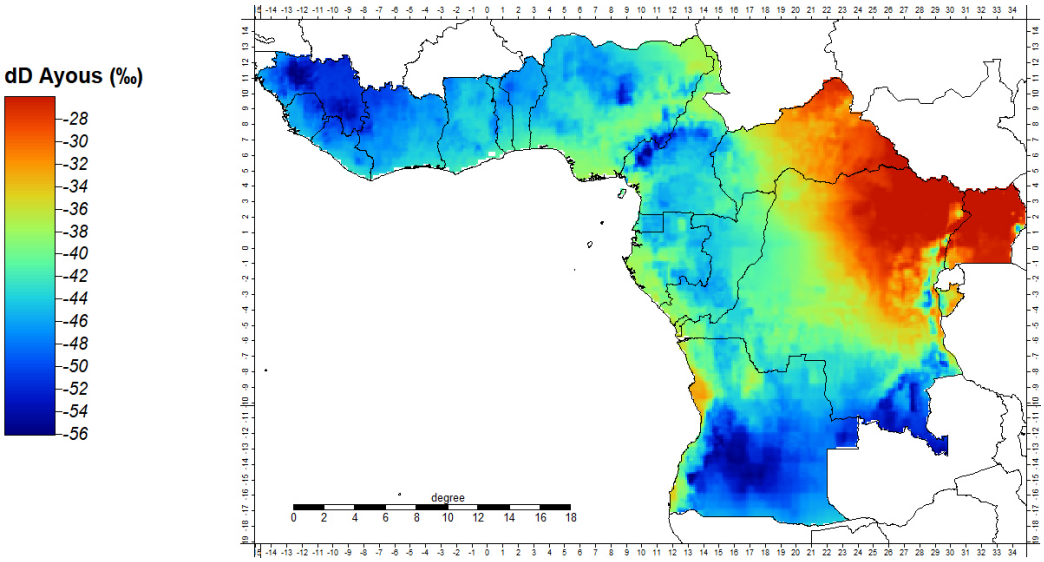


Fig. 13: Isoscape[28] of hydrogen isotope ratios (dD or D/H) of Ayous (*Triplochiton scleroxylon*) created using climate data and *Triplochiton* reference data analysed by the HBLFA Francisco-Josephinum, BLT Wieselburg, Austria (n=166). Degrees are degrees of longitude and lattitude.

3.2.1.5 MODEL DEVELOPMENT VIA DISCRIMINANT ANALYSIS

WHEN TO USE? Discriminant analysis is advantageous because it finds how to **maximise the differences** between the different stable isotope ratios of the different geographic provenances. Furthermore, it achieves this without losing the original information of the database. It is also a **time-served** method that has been used routinely to evaluate samples that have been involved in legal cases. However, only use multivariate methods **if more than two variables** (isotope ratios) reveal the provenance.

Discriminant analysis sometimes requires a minimum of 6 samples per group. A further **assumption** of the method is that samples geographically close together should have more similar isotope ratio values than samples that are geographically further away. However, this assumption does not always hold true.

[28] The model to create the isoscape was generated taking into account the relationship between D/H stable isotope ratios in the samples and D/H isotope ratios in mean annual hydrogen isotope precipitation data (waterisotopes.org) and a GTOPO30 Digital Elevation Model by means of a stepwise regression method that looks at the best predictors of any given isotope parameter from a list of about 324 different monthly-time-averaged climate grids/ 26 different climate parameters. These come from various sources such as NASA and the Climate Research Unit at the University of East Anglia (Watkinson *et al.*, in prep.).

Therefore, **data categorisation** must be carefully thought through. Typically, reference samples from provenances are assigned to different categories (*e.g.* country X, country Y, country Z or region A, region B, region C). It is good practice to attempt to break the categories of bigger countries down into smaller regions for better comparison with smaller countries. Another option is to assign reference samples to degree-cells. It should be noted that topographic elevation could affect isotopic signature and it is therefore good practice to record elevation of samples.

Kernel Discriminant Analysis (KDA) can be used to assess spatial clustering of the isotope data (for example with the *R* package *ks*). In this analysis the scaled mean stable isotope values per tree are used as test variables. Correct classification of each sample is then evaluated by comparing the percentage of correctly assigned samples per site to a prior probability of assignment per site based on random chance (see Vlam *et al.* 2018b for an example).

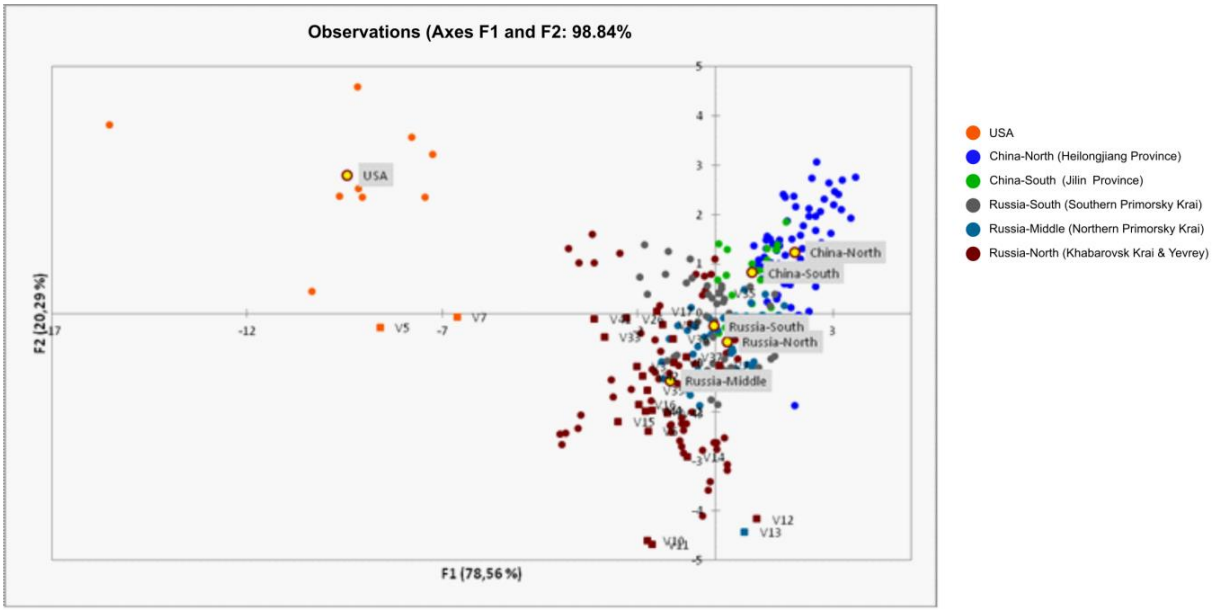


Fig. 14: Discriminate analysis of oak reference samples from Russia, China and USA using D/H, $^{18}O/^{16}O$, $^{13}C/^{12}C$ and $^{34}S/^{32}S$ showing results of 26 EIA test samples (squares) (EIA 2013).

3.2.1.6 MODEL VALIDATION

Leave one out cross validation (LOOCV) can be used to validate the model and calculate the success rate of assigning blind samples. Per site, n-1 is taken as a training dataset and the remaining ('removed') samples are assigned to one of the sites. The success rate of assigning blind samples is then tested in a 1000 bootstrap analysis.

3.2.2 ANALYSIS OF STABLE ISOTOPE DATA FROM TEST SAMPLES

3.2.2.1 SAMPLE & DATA PREPARATION

See §3.2.1 *Developing stable isotope reference data*.

3.2.2.2 DISCRIMINANT ANALYSIS WITH THE TEST SAMPLE

When **evaluating the probability scores**[29] for the sample, several reasons should be considered that could be the cause of a low probability:

- The sample is legitimate, but the database does not represent the location the sample is from. The statistic is then biased towards assigning the sample to the 'most similar' provenance.
  - Solution: collect more reference samples.
- The provenance probabilities of the sample are split across several similar categories/countries.
  - Solution: collect more reference samples and/or include additional stable isotopes if possible.
- The sample is an unusual sample for the category, sharing a similar signature only with few samples.
  - Solution: collect enough reference samples to cover the variability of the region.
- The sample is genuinely not from the declared category/provenance.

More **general recommendations for data interpretation** can be found in Appendix 6.

For **model optimisation**, unknown/test samples can be compared using all reference data in the model or using only a selection (see under §3.2.1).

[29] The percentage score from the test is more like a score of similarity than a probability "the sample originated from provenance x".

3.2.2.3 USING ISOSCAPES TO INTERPRET SAMPLE DATA

As isoscapes are interpolated maps, every pixel within the map (also known as a 'grid') contains a value. Statistical tools can be applied to these values in the same way as statistical tools can normally be applied. A **search for values on the maps that fulfil a criterion** (*e.g.* the 95% confidence interval range) can be done. These regions can then be overlaid to **narrow down where the isotope signature of a sample is possible within the limitations of the model**. This is useful because it is possible to represent an analytical result on a map rather than as a percentage or a dot on a graph.

More geographical information than just the isoscapes can be considered when evaluating a signature. For example, the provenance can be narrowed down by also considering the natural distribution of the tree population, where the timber is typically commercially harvested or where it is illegally harvested. In addition, it is useful if additional maps are produced to show the uncertainty in the mean value shown in the isoscape.

Fig. 15: Illustrative example of an isoscape result. Regions in Spain where the signature of a sample (declared as spruce from Spain) satisfied the 95% confidence interval of the isoscapes for D/H, $^{18}O/^{16}O$, $^{13}C/^{12}C$ and $^{34}S/^{32}S$.

3.2.3 STRENGTHS & LIMITATIONS

Strengths

- Only a **small amount of sample** is needed for the analyses (min. 5 g).
- Different stable isotopes can be combined to increase the **power and accuracy** in assignment tests. This is useful as a signature of different stable isotopes is determined by a different set of climatological, chemical, and biological processes.
- The measurement of stable isotope abundance by mass spectrometers is well documented, **can be done in many labs** and is relatively well standardized.

Limitations

- While **testing the provenance ID of derived wood products**, such as plywood, is possible (by analysing the individual layers), the investigation of chipboard or MDF (medium density fibreboard) currently is not possible or would be extremely laborious.
- The stable isotope method is not a 'fingerprinting' method in its classical sense. **No individual isotopic 'fingerprint' exists** and potentially samples of different geographic provenance can have similar isotope patterns. Therefore, it is advised to evaluate the declared provenance instead of trying to determine the provenance from the isotopic signature. For the same reason, it is also advised to use as large a group of stable isotope parameters as is economically reasonable to improve the provenance classification of unknown test samples.

For a discussion on isotope variation at different scales and hence the opportunities for **spatial differentiation** based on isotopic composition see Horacek *et al.* (2018) and Vlam *et al.* (2018b).

3.3 KEY LITERATURE FOR STABLE ISOTOPE DATA ANALYSIS

Aggarwal, P.K., K. Fröhlich, K.M. Kulkarni and L.L. Gourcy (2004). Stable isotope evidence for moisture sources in the Asian summer monsoon under present and past climate regimes. Geophysical Research Letters 31(8).

Boner, M. and H. Förstel (2004). Stable isotope variation as a tool to trace the authenticity of beef. Analytical and Bioanalytical Chemistry 378: 301-310.

Boner, M., Th. Somner, C. Erven and H. Förstel (2007). Stable isotopes as a tool to trace back the origin of wood. Proceedings of the international workshop "Fingerprinting methods for the identification of timber origins", October 8-9, Bonn/Germany.

Brienen, R.J.W., E. Gloor, S. Clerici, R. Newton, L. Arppe, A. Boom, S. Bottrell, M. Callaghan, T. Heaton, S. Helama, G. Helle, M.J. Leng, K. Mielikainen, M. Oinonen and M. Timonen (2017). Tree height strongly affects estimates of water-use efficiency responses to climate and CO2 using isotopes. Nature Communications 8: 288.

Broadmeadow, M.S.J., H. Griffiths, C. Maxwell and A.M. Borland (1992). The Carbon Isotope Ratio of Plant Organic Material Reflects Temporal and Spatial Variations in Co2 within Tropical Forest Formations in Trinidad. Oecologia 89(3): 435-441.

Deniro, M.J. and S. Epstein (1979). Relationship between oxygen isotope ratios of terrestrial plant cellulose, carbon dioxide and water. Science 204: 51-53.

Deniro, M. J. (1981). The Effects of Different Methods of Preparing Cellulose Nitrate on the Determination of the D-H Ratios of Non-Exchangeable Hydrogen of Cellulose. Earth and Planetary Science Letters 54(2): 177-185.

EIA, Environmental Investigation Agency (2013). Liquidating the Forests: Hardwood flooring, organized crime, and the World's last Siberian Tigers.

English, N.B., J.L. Betancourt, J.S. Dean and J. Quade (2001). Strontium isotopes reveal distant sources of architectural timber in Chaco Canyon, New Mexico. Proceedings of the National Academy of Sciences of the United States of America 98(21): 11891-11896.

Farquhar, G.D., J.R. Ehleringer and K.T. Hubick (1989). Carbon Isotope Discrimination and Photosynthesis. Annu. Rev. Plant Physiol. Plant Mol. Biol. 40:503-37.

Gori, Y., R. Wehrens, M. Greule, F. Keppler, L. Ziller, N. La Porta and F. Camin (2013). Carbon, hydrogen and oxygen stable isotope ratios of whole wood, cellulose and lignin methoxyl groups of *Picea abies* as climate proxies. Rapid Communications in Mass Spectrometry 27(1): 265-275.

Heaton, K., S.D. Kelly, J. Hoogewerff and M. Woolfe (2008). Verifying the geographical origin of beef: The application of multi-element isotope and trace element analysis. Food Chemistry 107: 506-515.

Horacek, M., M. Jakusch and H. Krehan (2009). Control of origin of larch wood: discrimination between European (Austrian) and Siberian origin by stable isotope analysis. Rapid Commun. Mass Spectrom. 23: 3688–3692.

Horacek, M. (2012). Christmas tree production in Europe: Control of declared geographical origin by stable isotope analysis – a pilot study. Die Bodenkultur 63(1): 35-41.

Horacek, M., G. Rees, M. Boner and J. Zahnen (2018). Comment on: Developing forensic tools for an African timber: […], by Vlam et al., 2018. Biological Conservation 226: 333-334.

Leavitt, S (2010). Tree-ring C–H–O isotope variability and sampling. Science of the Total Environment 408: 5244–5253.

McCarroll, D. and N.J. Loader (2004). Stable isotopes in tree rings. Quaternary Science Reviews 23(7-8): 771-801.

Rinne, K.T., T. Boettger, N.J. Loader, I. Robertson, V.R. Switsur and J.S. Waterhouse (2005). On the purification of alpha-cellulose from resinous wood for stable isotope (H, C and O) analysis. Chemical Geology 222(1-2): 75-82.

Roden, J.S., G. Lin and J.R. A Ehleringer (2000). Mechanistic model for interpretation of hydrogen and oxygen isotope ratios in tree-ring cellulose. Geochimica et Cosmochimica Acta 64(1):21–35.

Sauer, P.E., A. Schimmelmann, A.L. Sessions and K. Topalov (2009). Simplified batch equilibration for D/H determination of non-exchangeable hydrogen in solid organic material. Rapid Communications in Mass Spectrometry 23(7): 949-956.

Seal, R.R. II (2006). Sulfur Isotope Geochemistry of Sulfide Minerals. USGS Staff - Published Research. 345. https://digitalcommons.unl.edu/usgsstaffpub/345

Siegenthaler, U. (1979). Stable Hydrogen and Oxygen Isotopes in the Water Cycle. Lectures in Isotope Geology pp 264-273 Springer-Verlag Berlin. DOI: 10.1007/978-3-642-67161-6_22

Thode, H.G., J. Monster and H.B. Dunford (1961). Sulphur isotope geochemistry. Geochimica et Cosmochimica Acta 25:159-174. Pergamon Press Ltd. Printed in Northern Ireland.

van der Sleen, P., P.A. Zuidema and T.L. Pons (2017). Stable isotopes in tropical tree rings: theory, methods and applications. Functional Ecology 31(9): 1674-1689.

Vlam, M., A. Boom, G.A. de Groot and P.A. Zuidema (2018a). Tropical timber tracing and stable isotopes: A response to Horacek et al. Biological Conservation 226: 335-336.

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

4. DART TOF Mass Spectrometry

Definition DART TOFMS reference data: the chemical fingerprint of a wood sliver (and by extent species/provenance) based on the complete set of small chemical molecules found within the sample.

Authors: Victor Deklerck, Edgard Espinoza, Cady Lancaster, Sandra Martínez-Jarquín, Kathelyn Paredes-Villanueva, Robert Winkler

*Authors are in alphabetical order.

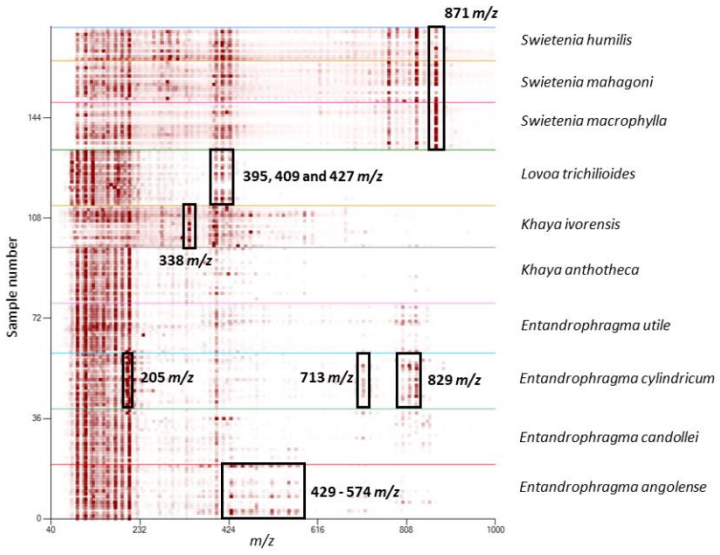


Fig. 16: Heatmap showing the presence of the ions for the different specimens per species.

4.1 RESOURCES REQUIRED

Access to reference material

The largest database currently available for DART TOFMS is the **Forensic Spectra of Trees Database©** (ForeST Database©). This database is held and curated by the National Fish and Wildlife Forensic Laboratory, NFWFL (Ashland, OR, USA). ForeST© contains thousands of species from hundreds of genera. The focus of the database is for species identification of CITES-listed species, lookalikes of the CITES-listed, and commercially significant timber species, especially from tropical regions. **A softwood specific database** is currently being refined and provides genus-level identification.

Reference applicability

For use of the ForeST Database©, NFWFL recommends that it be used in conjunction with **JEOL-line TOF MS**. Due to the chemical complexity of wood products, academic investigation of species separation and classification can use a variety of instrument parameters. **Use of the ForeST Database© requires tuning instruments to the database**. Contact NFWFL for more details on instrument validation.

NOTE: DART TOFMS is not the only mass spectrometer that can be used for timber identification, it is however, the most developed one with the most developed database and adjoined software.

Software

Table 5: Overview of the software used for wood identification via DART TOFMS.

Software	Performance	Accessibility
R	Can be used for any statistical computing task. Computer programming knowledge needed. Customization easily implementable.	Free software
Mass Mountaineer	Purpose-built software for DART TOFMS with all standard statistical packages included. More labour-intensive customization possibilities for other research explorations. Recommended for use in forensic science.	Commercial software
NIST Search Software	The National Institute of Standards and Technology (NIST) search software provides some of the most advanced search algorithms for qualitative analysis. Spectra can be quickly compared.	Commercial software

4.2 USING DART TOFMS FOR TAXON OR PROVENANCE IDENTIFICATION

4.2.1 DEVELOPMENT OF DART TOFMS REFERENCE DATA

NOTE: DART TOFMS is not regularly used for provenance identification at the time of this publication. However, although the variation in the results is subject of further investigations, the protocol below can be used as a start for provenance identification as well (Espinoza *et al.* 2014, Finch *et al.* 2017, Paredes-Villanueva *et al.* 2018).

4.2.1.1 SAMPLE SELECTION & PREPARATION

Collect samples of the species and/or geographic region (provenance) needed to be characterised and include **equivalent sample sizes per species/provenance**. Use only the **heartwood** (see GTTN sampling guide) and the non-contaminated parts of the wood (see Appendix 4). However, it is still good practice to also collect **reference material of known or suspected contaminants** (such as varnish, oil, coating). Molecular ion peaks of known contaminants can be subtracted from the mass spectra and contaminated wood samples can be salvaged in post-processing.

4.2.1.2 SPECTRA COLLECTION

Collect the spectra as per the instructions of the mass spectrometer instrument. **The mass spectrometers are calibrated** with a reference solution of known masses (for example polyethylene glycol solution). If the ForeST Database© is to be used, the DART TOFMS should be calibrated to the database (see §4.1 > *Reference applicability*).

For **archiving and sharing the mass spectrum**, it is important to declare the provenance of the reference material. One easy solution is to label the individual spectrum file names with explicit data that can describe the metadata easily. NFWFL uses the following strategy: GenusSpecies_AnalysisLabAccession_SourceAccession. For example:

File name: DalbergiaNigra_WD123456_MADw1234

Species: *Dalbergia nigra*

NFWFL accession number: WD123456

Original catalogue number: Forest Products Lab ID for Madison Wood Collection (MADw) 1234.

4.2.1.3 DATA PREPARATION

Before developing a statistical model there is a need to remove outliers from training sets. Hawkins (1980) describes 'an outlier' as *an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism*. **Cross-check the spectra with spectra already in the database and with spectra of known contaminants**.

- Heatmaps allow for a simultaneous comparison of all the spectra data and are ideal for identifying outliers (*e.g.* in *R* or *Mass Mountaineer*).
- Average spectra can be created (*e.g.* in *R*) for each species by combining all available spectra for this species. From this super spectrum outliers can be identified.
- Be cautious of contaminants. Use heatmaps of the known contaminants and cross-reference to the wood sample data set to avoid selecting contaminate molecular ion peaks in model training.

Before removing samples from the dataset, outliers should be further evaluated. One way to evaluate spurious spectra is to determine if suspected outlier specimens have a different and distinct chemotype from other spectra of the same species. Another way is to check if the intensity is similar for all the *m/z* ions. Samples from the same species will show similar intensities of the molecules detected.

Possible reasons for outliers are:

- The reference database does not yet represent the chemical intra-variability of the species or provenance[30].
- Mislabelled or misidentified samples.
- The measurement was not done properly.
- The tissue type was different from the other specimens (*e.g.* sapwood *vs.* heartwood)
- There is a contaminant present on the sample.

[30] When is an outlier not an outlier? Whichever approach you take to determine this, it is key to know your data and your research area well. For more info on data interpretation see Appendix 6.

It is recommended that outlier samples be re-analysed by DART TOFMS to be able to exclude the possibility of an erroneous measurement.

4.2.1.4 MODEL DEVELOPMENT

- ☑ Only use reference spectra which are in consensus.
- ☑ Hold out reference spectra to check later for model overfitting and model validation.
- ☑ Build statistical models for species/provenance classification using (i) the same number of spectra for each species/provenance included, and (ii) a suitable kind of statistical analysis. What is suitable will depend on the species/provenance (group) being investigated.
  - Multivariate statistics such as Principal Component Analysis (PCA), Kernel Discriminant Analysis (KDA) or Discriminant Analysis of Principal Components (DAPC) with *Mass Mountaineer* are **recommended for forensic analysis** because this software easily produces quantitative results and provides probability estimates, which are used in court to describe certainty of the analysis.
  - *Random forest* (Deklerck *et al.* 2017, Finch *et al.* 2017, Paredes-Villanueva *et al.* 2018, Deklerck *et al.* 2019) has the advantage that it **can be used when there are less than 4 classes**.
  - PAM clustering or Adaptive Boosting are two robust classifiers for discrimination between **two groups**.

The application of DART TOFMS to wood identification is a relatively new technique (since 2012), and therefore model development options are not limited to the above list since new mathematical tools are continuously being explored (such as dynamic time warping and deep learning algorithms).

4.2.1.5 MODEL OPTIMISATION & VALIDATION

Before constructing the final model, it is proposed to **screen the spectral data pre-processing parameter settings for optimal classification accuracy**. Depending on the classification algorithm this can be done as follows:

- Reduce the number of variables (combinations of ions) by carefully selecting components in a PCA.
- Select ions for model building using Fisher Ratio analysis or by careful visual examination of heatmaps.
- Mass tolerance for binning and relative abundance cut-off threshold settings (pre-processing parameters) and number of ions (classification parameter) can be screened in an automated way, when using the *random forest* algorithm or any algorithm that is data frame dependent and does not work with individual text files for model-building, as described in Deklerck *et al.* (2019).

Test the model for overfitting and classification accuracy by removing reference spectra (leave one out cross validation, LOOCV) and by analysing unused reference spectra that were not used in the model development (sometimes called 'hold-out set' or 'validation set'). Model parameters are optimised to obtain (i) the highest LOOCV possible, and (ii) accurate classification of the hold-out specimens.

4.2.2 ANALYSIS OF DART TOFMS DATA FOR TEST SAMPLES

NOTE: The **purpose of this protocol** is to provide a procedure to analyse and identify specimens of wood by comparing them with a curated database of known species or a specific set of reference samples. The intent of this method is not to identify all the compounds found in the wood samples, but to infer from specific ions present that a given wood sample did or did not originate from a known species. An experienced analyst may occasionally need to vary the procedure to accommodate a particular sample. The different data analysis steps:

- **Wood anatomical analysis** of the unknown sample to determine the genus.
- **Collection of spectra** as described in §4.2.1.2.
- **Library search for preliminary classification**. Determine if the unknown spectra match a species held in the ForeST reference database. The top library hits are then used to create the subsequent heatmaps and multivariate statistical models. This step is implemented in *Mass Mountaineer*.
- **Creation of a heatmap and evaluation of the unknown**. Frequently the library search will indicate that two or three species have spectra similar to the unknown. These taxa should be used to create the heatmap and to determine if the chemotypes for each species selected are similar to that of the unknown spectra.
- **Selection of the variables (ions) for the multivariate modelling**. Use the full set of curated reference spectra available for the species/provenance of interest as the training set. For *random forest*, variable selection is automatically included in the set-up and can be optimized (see §4.2.1.5). This is similar for other machine learning techniques or classification algorithms in *R*.
- **Taxonomic assignment of the unknown** once a model performs satisfactorily. When using *Mass Mountaineer*, the software has a utility to automatically classify the unknown spectra against the multivariate model and the program provides a probable estimate of accuracy. For *random forest*, first build the model according to §4.2.1.4, then use the full reference set as training and the unknowns as test set.
- **Final classification decision** is obtained when (i) the library results, (ii) the heatmap, and (iii) the multivariate analysis all agree with each other.
- **Classification validation** as described in §4.2.1.5.

4.2.3 STRENGTHS & LIMITATIONS

Strengths

- **Very small sample** is needed (wood slivers).
- **Versatile sampling**. As only a wood sliver is needed, samples can be easily collected from a range of products (guitars, watches, logs, …).
- Analyses can be done at **low cost**. Once you have the mass spectrometer, running costs are low and no other equipment, expensive chemicals or software is required for wood identification.
- **Tailor-made software** for data analyses (*Mass Mountaineer*, see Table 5) comes together with the purchase of a JEOL DART TOFMS.
- Analyses can be performed in few minutes, allowing for **high-throughput** screening. Once models are built for certain (groups of) species/provenances wood identification can be done within a short time.
- A **curated reference database** (ForeST Database©) of about 2000 species, representing 630 genera is currently available (Nov. 2019).

Limitations

- **Extreme physical or chemical processes** in the wood (caused by *e.g.* temperature, micro-organisms) could alter the DART MS profile.
- **Possible interference of chemical contaminants** such as glue and packaging.
- **Wood panels and plywood** might be difficult to identify due to the adhesives used for the manufacturing. The mix of species used for their fabrication might also be a limitation.
- **Some tree species** present few signals in DART MS and therefore provide insufficient data for evaluation.
- **Sap- and heart-wood can show different profiles**. As the current database is based on heartwood only (being the best identifier and the most used in wooden objects), caution is therefore needed when an unknown sample might contain sapwood.

4.3 KEY LITERATURE FOR DART TOFMS DATA ANALYSIS

Baxter, M.J. (1999). On the Multivariate Normality of Data Arising from Lead Isotope Fields. Journal of Archaeological Science 26(1): 117–124.

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

Cody, R.B., A.J. Dane, B. Dawson-Andoh, E.O. Adedipe and K. Nkansah (2012)."Rapid classification of White Oak (*Quercus alba*) and Northern Red Oak (*Quercus rubra*) by using pyrolysis direct analysis in real time (DART (TM)) and time-of-flight mass spectrometry. Journal of Analytical and Applied Pyrolysis 95: 134-137.

Deklerck, V., K. Finch, P. Gasson, J. Van den Bulcke, J. Van Acker, H. Beeckman and E. Espinoza (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 31(19): 1582-1588.

Deklerck, V., C. Lancaster, J. Van Acker, E. Espinoza, J. Van den Bulcke and H. Beeckman (under review). Chemical fingerprinting of wood sampled along a pith-to-bark gradient allows individual distinction and provenance identification. Forestry.

Deklerck, V., T. Mortier, N. Goeders, R.B. Cody, W. Waegeman, E. Espinoza, J. Van Acker, J. Van den Bulcke and H. Beeckman (2019). A protocol for automated timber species identification using metabolome profiling. Wood Science and Technology https://doi.org/10.1007/s00226-019-01111-1.

Espinoza, E.O., C.A. Lancaster, N.M. Kreitals, M. Hata, R.B. Cody and R.A. Blanchette (2014). Distinguishing wild from cultivated agarwood (*Aquilaria* spp.) using direct analysis in real time and time of-flight mass spectrometry. Rapid Communications in Mass Spectrometry 28(3): 281-289.

Espinoza, E.O., M.C. Wiemann, J. Barajas-Morales, G.D. Chavarria and P.J. McClure (2015). Forensic Analysis of Cites-Protected *Dalbergia* Timber from the Americas. IAWA Journal 36(3): 311-325.

Evans, P.D., I.A. Mundo, M.C. Wiemann, G.D. Chavarria, P.J. McClure, D. Voin and E.O. Espinoza (2017). Identification of selected CITES-protected Araucariaceae using DART TOFMS. IAWA Journal 38(2): 266-S3.

Finch, K., E. Espinoza, F.A. Jones and R. Cronn (2017). Source Identification of Western Oregon Douglas-Fir Wood Cores Using Mass Spectrometry and Random Forest Classification. Applications in Plant Sciences 5(5): apps.1600158.

D., Hawkins (1980). Identification of Outliers. Chapman and Hall, London.

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

Lancaster, C. and E. Espinoza (2012). Evaluating agarwood products for 2-(2-phenylethyl) chromones using direct analysis in real time time-of-flight mass spectrometry. Rapid Communications in Mass Spectrometry 26(23): 2649-2656.

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

Paredes-Villanueva, K. (2018). Tropical timber forensics: a multi-methods approach to tracing Bolivian *Cedrela* (Doctoral dissertation, Wageningen: Wageningen University).

Paredes-Villanueva, K., E. Espinoza, J. Ottenburghs, M.G. Sterken, F. Bongers and P.A. Zuidema (2018). Chemical differentiation of Bolivian *Cedrela* species as a tool to trace illegal timber trade. Forestry: An International Journal of Forest Research 91(5): 603-613.

5. NIR spectroscopy

Definition NIR spectroscopy reference data: NIR spectra database, in reflectance/transmittance or absorbance, collected under specific conditions, characterising the wood of a specific taxon, grown in a specific geographical area.

Authors: Jez W.B. Braga, Gilles Chaix, Tereza C.M. Pastore, Tahiana Ramananantoandro, Andriambelo R. Razafimahatratra, Liz F. Soares

*Authors are in alphabetical order.

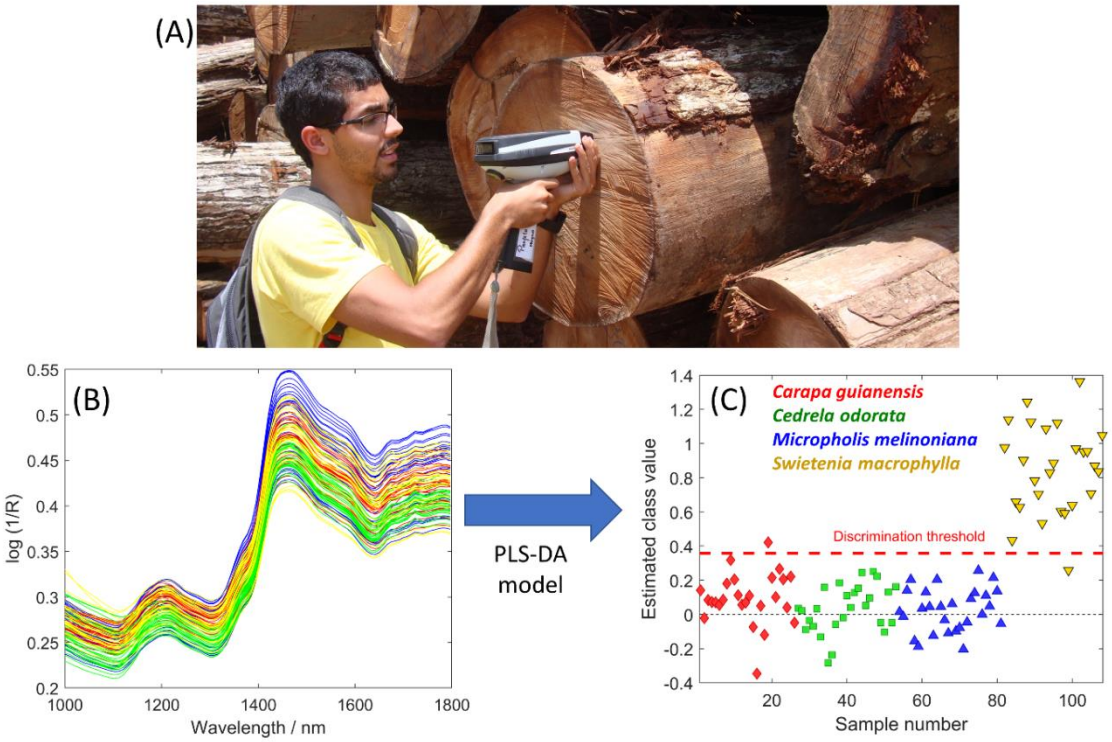


Fig. 17: Collecting NIR spectra with a portable spectrometer (A). NIR database (B). Discrimination with a PLS-DA model (C).

5.1 RESOURCES REQUIRED

Access to reference material

Until now, there are no open-access international common databases of NIR spectra of wood available. The **difficulties in sharing NIR spectra databases** are due to various aspects, such as:

- **Spectrometer hardware.** Different spectrometers can have a different type of sensor, detector, light source, spectral resolution, spectral range.
- **Spectrometer software.** Each manufacturer of bench or portable equipment uses specific software to control the equipment, measure the spectra and for statistical data processing (different data file type).
- **Spectrometer settings.** The settings used to produce spectra (integration time, scan number, initial spectral resolution), reflectance/transmittance or absorbance mode might differ between users of the equipment.
- **Sample condition**. Wood samples used for analysis might differ in material characteristics (solid wood or ground wood, grain size if ground wood, moisture content, temperature, ligneous plane)

As a result, discriminating between spectra collected from different devices or under different environmental conditions might result in erroneous wood identification. The differences between devices might be larger than the differences between species, especially if sample measurement conditions are not standardized. The sharing or standardization of NIR databases require application of specific calibration transfer approaches, which are detailed in the chemometric literature (Wang 1991, Honorato 2007).

Fig. 18: Collecting NIR spectra with different portable spectro-meters.

Software

Table 6: Overview of some of the software that could be used for NIR spectroscopy wood identification.

Software	Application	Accessibility
Chemflow	Pattern recognition, classification, quantitative analysis.	Free online software but need internet access.
Grams (Thermo Fisher)	Pattern recognition, classification, quantitative analysis. Linked to Bruker, BUCHI, FOSS or Thermo Fisher NIR and FTIR spectrometers.	Commercial software, delivered with Bruker, BUCHI, FOSS or Thermo Fisher equipment.
NIRCal (BUCHI)		
OPUS (Bruker)		
WinIsi (FOSS)		
MATLAB	Can be used for any statistical computing or imaging task. Computer programming knowledge needed.	Commercial software
OCTAVE		Free software
R		Free software
Scilab		Free software
PLS_ToolBox/Solo	Pattern recognition, classification, quantitative analysis, curve resolution, optimization, experimental design, calibration transfer.	Commercial software
The Unscrambler		Commercial software
SIMCA	Pattern recognition, classification, quantitative analysis.	Commercial software
XLSTAT		Commercial software
STATISTICA	Pattern recognition, classification, quantitative analysis, experimental design.	Commercial software

5.2 USING NIR SPECTROSCOPY FOR TAXON OR PROVENANCE IDENTIFICATION

5.2.1 DEVELOPMENT OF NIR SPECTROSCOPIC REFERENCE DATA

5.2.1.1 SAMPLE SELECTION & PREPARATION

The NIR spectrum is a response that varies with (1) the physical characteristics of the wood, (2) moisture content, (3) the concentration of the chemical components present in the wood, (4) the distribution of these components over the wood surface. All factors that can influence these variables can hence also influence the NIR spectra. Therefore, careful selection and preparation of reference samples is needed (see the GTTN sampling guide and Appendix 4). The **roughness of the sample surface** should be controlled and be homogeneous and clean (*e.g.* by using number 80 sandpaper and a dry brush).

After these first considerations, take approximately two-thirds of the total samples for **training the model**. For the **external validation of the model**, use the remaining one-third. The external validation set of samples should be entirely independent of those samples used to build the model (ASTM 2017). Both sets of samples should include **all species/provenances considered** and in sufficient sample numbers to **cover the variability** within the species/provenances, with at least 20 samples per species and per provenance (Agelet & Hurburgh Jr 2010). The split of the samples between training and validation sets can be performed randomly or applying specific algorithms (*e.g.* Duplex algorithm) (Westad & Marini 2015).

5.2.1.2 SPECTRA COLLECTION

After careful sample selection and preparation (see above), spectra are collected from **solid or ground wood** (*e.g.* by using a grinder of 500µ or 1mm) using a benchtop or handheld NIR spectrometer. The choice of sample preparation (solid or ground) should be made prior to the spectral collection and the same sample preparation procedure should be followed for all samples.

When measuring spectra, the **training set and validation set should stay entirely independent**. All replicate spectra of one sample must be used either in the training set, or in the external set to validate and not be mixed. However, it is not necessary to proceed by groups of species or provenances. **Measurement order is best mixed**, proceeding with samples of different species/provenances chosen in a random way.

Collect at least **3 diffuse reflectance spectra** for each of **3 different spots** selected on the wood surface. A comprehensive model for wood identification includes spectra obtained from **all three wood anatomical planes** (longitudinal, radial, and transverse) (Braga *et al.* 2011, Pastore *et al.* 2011, Bergo *et al.* 2016, Soares *et al.* 2017). If it is not possible to sample all planes it is recommended to collect spectra on the longitudinal-radial (RT) surface because it is more representative of the wood (several growth rings might be sampled). The transverse plane can also be used if this surface is more accessible. However, when a specific plane was selected for the method development, the application of the method must be restricted to this plane.

Make sure that all spectra are collected under the **same environmental conditions**, in the model development phase as well as in the validation phase. As a minimum, moisture content and temperature of the sample should be controlled. It is advised to use samples with a moisture value[31] below 15%, ideally 12%, using a climatic chamber set at 20°C and 65% relative humidity (if the analysis will be carried out under laboratory conditions).

NOTE. The data analysis steps of data cleaning, model development and model validation are given here below in separate chapters for the sake of clarity. In practice, however, the data cleaning and model development steps are alternatingly iterated until the optimal model for the identification of the species/provenance of interest is found. Only after the training model is set, the validation data is used to evaluate the performance of the method.

5.2.1.3 DATA CLEANING

NIR spectroscopy is an indirect wood identification method. Spectra cannot be compared directly between samples as the NIR region is composed of highly overlapping spectral bands resulting in wood spectra that are similar between samples. Therefore, multivariate analysis is used (*e.g.* PCA, PLSR) to decrease the number of variables, the collinearity, and to work on independent variables (Principal Components, Latent Variables). **First, the spectral information needs to be calibrated, in a standardised way**, before samples can be compared based on the chemometric model.

[31] Water in the wood will displace the baseline of the spectrum and the intensity of the absorption bands (due to O-H bonds).

☑ Verify the raw NIR spectra and **evaluate the need for pre-processing** to minimize additive and multiplicative effects or to exclude uninformative or noisy regions.

☑ Find the most appropriate pre-processing method to **correct baseline shifts and other instrumental variations**. One of the most applied methods is first and second derivative pre-processing based on Savitzky-Golay smoothing (by identifying the optimum window length of the filter). However, other methods can also be tested like detrending, standard normal variate or vector normalization, mean or class centering, multiplicative scattering correction or their combination (Rinnan *et al.* 2009).

☑ Verify the raw or pre-processed NIR spectra to **eliminate outliers**. Spectral exclusion from the training set is important to remove possible measurement errors or spectra from samples presenting some characteristics that differ from most of the samples that define a species/provenance. Whenever possible, outlier samples should be verified to identify the reason for this anomalous behaviour.

How to decide what is anomalous:

- Have a look at spectra profiles as some extreme outliers might be identified visually.
- Exclude all samples that present Hotelling T^2 and Q residuals of PCA larger than the predefined confidence limits for these parameters (*i.e.* 95% or 99% of confidence), as illustrated in Fig. 19.

☑ Select a spectral range to **minimize spectral noise**. An improvement in the results could be achieved by the selection of a spectral range (i) related to the difference between the species/provenances or, (ii) unrelated with the chemical information of the wood (Schwanninger *et al.* 2011). The selection of the spectral range can be performed by the knowledge of the specific regions for wood components, based on model parameters (*i.e.* regression coefficients or variable importance in the projection (VIP) scores (Chong & Jun 2005)) or automatic search algorithms (*i.e.* Genetic Algorithm, Successive Projection Algorithm (SPA), Orthogonal Projection Algorithm (OPS), among others) (Araújo *et al.* 2001, Teófilo *et al.* 2009). Regions located at the limits of the spectra (beginning and end) are non-informative as they are noisy and should therefore be removed.

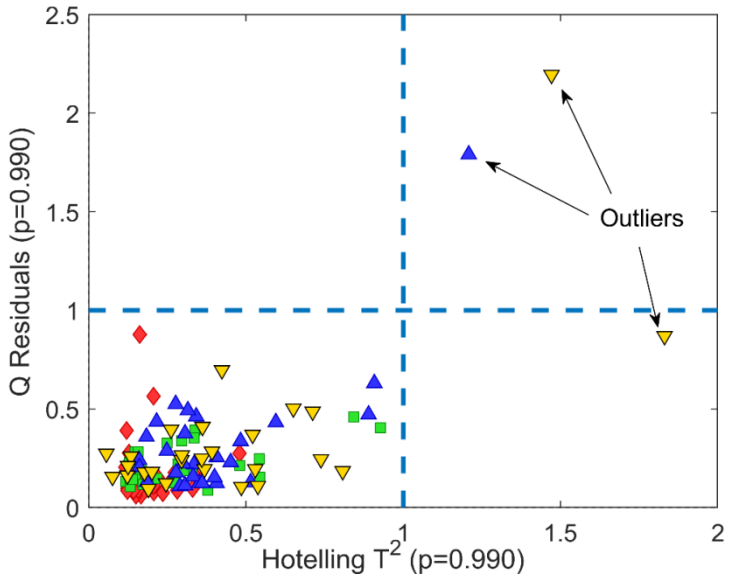


Fig. 19: Illustration of a Hotelling T^2 versus Q residuals graph for outlier identification in a PCA or PLS-DA model. (---) 99% confidence limits.

Based on our experience of the chemical heterogeneity of wood samples, it is not necessary to average the replicates of the spectra taken from different spots of a sample. The **spectra taken from different spots** help to increase the representativeness of the data set for the species or provenance of interest (Soares *et al.* 2017, Silva *et al.* 2018, Snel *et al.* 2018).

IMPORTANT NOTE: The same criteria used in this data cleaning step for outlier removal, should also be used in the validation or application step to decide if a sample belongs to the population used for model development and to **minimise misidentifications due to an unsuitable model** (Soares *et al.* 2017, Silva *et al.* 2018, Snel *et al.* 2018). The check for outliers in the validation or test set is performed to prevent that an outlier sample will be analysed in the models. Results obtained for outliers cannot be trusted.

5.2.1.4 MODEL DEVELOPMENT

Wood identification using NIR spectroscopy can be based on discrimination or classification models.

- A discrimination model is a binary forecast model that is built using samples of the species to be identified (target classes) as well as samples from all species that are visually or anatomically similar and might be analysed in the application of the model (non-target classes, used for validation or to model the characteristics of the samples/species not belonging to the target classes). The most applied models are Partial Least Squares for Discriminant Analysis (PLS-DA) or Linear Discriminant Analysis (LDA), performed by most software (see Table 6).
- A classification model is a one class classifier, which is usually developed using only samples of the species of interest (one or more) and validated with independent samples of this species and of all visually or anatomically similar species (look-alike species). The most applied approach is Soft Independent Modelling of Class Analogy (SIMCA), which is present in most of the software (see Table 6). With this approach, independent SIMCA models must be developed when there is more than one species of interest.

Next to these, there is a variety of other approaches such as machine learning, support vector machine, PLS-DA weighted, local PLS-DA, PARAFAC (Balabin *et al.* 2010).

To find the optimal model for the identification of the species/provenance of interest, a **careful selection of the number of principal components** (PC) **or latent variables** (LV) should be performed. The selection should be based on the cross-validation results (see further below §5.2.1.5) for the training set **to avoid low efficiency rates and overfitting the model**. Low efficiency rates are obtained when a lower number of PC, in comparison with the optimal value, is selected. When a higher number of PC is selected the model is overfitted. Usually, cross-validation tests using a representative number of samples, allow a good estimation of the correct number of principal components or latent variables (Soares *et al.* 2017).

After data optimisation and model development using a discrimination approach, a **discrimination threshold** is estimated based on the dispersion of the estimated class values obtained for the training set. This threshold is estimated according the Bayes' theorem to minimize the occurrence of false positive and negative discrimination errors. For classification models, the classification is performed by the estimation of **confidence limits** that delimitate a region for the species of interest in the space defined by the principal components of the SIMCA model.

5.2.1.5 MODEL VALIDATION

IMPORTANT NOTE: Generally, for model validation, an **independent set of samples** is used that was not used for construction of the model. It is essential to **keep the same experimental conditions** as those used in the model development phase. Preparation of samples to build the model and to validate the model should be the same (Soares *et al.* 2017, Silva *et al.* 2018, Snel *et al.* 2018).

- Cross-validation tests can be used to identify the **best pre-processing method**, as well as to identify the **number of model parameters**, *i.e.* latent variables or principal components (Soares *et al.* 2017, Chaix *et al.* 2018, Silva *et al.* 2018, Snel *et al.* 2018). Repeated cross-validation is more accurate, since the groups used in cross-validation are generated randomly. The leave-one-out method can also be used but it takes more time if there are many samples and several pre-processing methods to be tested.
- Double cross-validation tests can be used to evaluate the **robustness of the model** (Filzmoser *et al.* 2009). Double cross-validation requires, however, many samples.
- Quality parameters can be calculated, based on the results obtained with the independent set of validation samples. The most common ones are the **efficiency rate** (percentage of correct identifications) and the rates of **false positive and negative errors**.
- Periodic performance checks of models should be undertaken with samples of known identity to **verify stability and accuracy** of the results. The performance of the model can, for example, decrease **after maintenance of the spectrometer, or after extra reference spectra have been added** to the model (adding more species/provenance variability). Therefore, models must periodically be updated and adjusted.

5.2.2 ANALYSIS OF NIR SPECTROSCOPIC DATA FROM TEST SAMPLES

The different data analysis steps

- **Samples are prepared and spectra collected** as described in §5.2.1.
- **Spectra are pre-processed** using the same method as used during the establishment of the model.
- **Spectra are analysed** in the PLS-DA, LDA or SIMCA model and the species/provenance **identification of the sample** is given by the software. For PLS-DA or LDA methods, a test sample is only identified as belonging to a specific class (species or provenance) if the sample is not an outlier and its estimated class value is larger than the discrimination threshold for that class. The estimated class value must be lower than the discrimination threshold for all other predefined classes of the model. For SIMCA, samples are classified as belonging to a specific class if they are within the confidence region of this class. Otherwise, they could be identified as outliers or inconclusive results (when samples are within the confidence region of multiple classes) (Soares *et al.* 2017, Silva *et al.* 2018, Snel *et al.* 2018).
- **If a significant number of test samples of the predefined classes were identified as outliers by the existing models**, this would indicate a lack of representativeness of the reference data used for model development or that the new samples were measured in different experimental/environmental conditions. In case of lack of representativeness, extra reference samples should be collected, added to the previous database and the model be recalculated and validated.

5.2.3 STRENGTHS & LIMITATIONS

Strengths

- Sample preparation **time** is minimal.
- Few or no **consumables** needed.
- Once models are built, a wood sample can be identified in a **few seconds** and deep scientific background from the users is not needed in routine analysis.
- The analysis can be done under **field conditions**, provided that the training samples were measured under field conditions as well.
- Some portable NIRS instruments associated with PLS-DA models can show an **efficiency rate** of larger than 90% for wood discrimination in the field (Soares *et al.* 2017).

Limitations

- Depending on the specifications of the spectrometer, an **initial investment** of around 1000 to 75000 US$ is needed to buy the NIR spectrometer. Relatively cheap spectrometers exist but with narrow spectral range and low spectral resolution. The choice of the instrument depends on the objectives of the study.
- The results might be influenced by the specific environmental and sample conditions at the moment of spectra collection. Establishment of **equal experimental conditions for reference and test samples** is of utmost importance.
- The **transposition (model transfer)** of a statistical model for wood identification using NIR spectra from one machine to another may require standard samples and mathematical standardization procedures (Wang 1991, Honorato 2007). This is needed even if the type of equipment is the same (bench or portable) and even when the equipment is from the same manufacturer and model. That is because the calibration settings of the machine might be different in different labs.

5.3 KEY LITERATURE FOR NIRS DATA ANALYSIS

Agelet, L.E. and C.R.Jr. Hurburgh (2010). A Tutorial on Near Infrared Spectroscopy and Its Calibration. Critical Reviews in Analytical Chemistry 40: 246-260.

Araújo, M.C.U., T.C.B. Saldanha, R.K.H. Galvão, T. Yoneyama, H.C. Chame and V. Visani (2001). The Successive Projections Algorithm for variable selection in spectroscopic multicomponent. Chemometrics and Intelligent Laboratory Systems 57: 65-73.

ASTM E1655-17 (2017). Standard Practices for Infrared Multivariate Quantitative Analysis, ASTM International, West Conshohocken, PA, www.astm.org.

Balabin, R.M., R.Z. Safieva and E.I. Lomakina (2010). Gasoline classification using near infrared (NIR) spectroscopy data: Comparison of multivariate techniques. Analytica Chimica Acta 671: 27–35.

Bergo, M.C., T.C. Pastore, V.T. Coradin, A.C. Wiedenhoeft and J.W. Braga (2016). NIRS Identification of *Swietenia macrophylla* is Robust Across Specimens from 27 Countries. IAWA Journal 37(3): 420-430.

Braga, J.W.B., T.C.M. Pastore, V.T.R. Coradin, J.A.A. Camargos and A.R. da Silva (2011). The Use of near Infrared Spectroscopy to Identify Solid Wood Specimens of *Swietenia macrophylla* (Cites Appendix Ii). IAWA Journal 32(2): 285-296.

Chaix, G., T. Giordanengo, V. Segura, N. Mourey, B. Charrier and J.P. Charpentier (2018). Near infrared spectroscopy, a new tool to characterize wood for use by the cooperage industry. In: Stevanovic Tatjana (ed.). Chemistry of lignocellulosics: current trends. Boca Raton: CRC Press, 42-65.

Chong, I.-G. and C.-H. Jun (2005). Performance of some variable selection methods when multicollinearity is present. Chemometrics and Intelligent Laboratory Systems 78: 103–112.

Dardenne, P. (2010). Some considerations about NIR spectroscopy: Closing speech at NIR-2009. NIR news 21, 8–10.

Filzmoser, P., B. Liebmann and K. Varmuza (2009). Repeated double cross-validation. Journal of Chemometrics 23: 160–171.

Honorato, F.A., M.N. Martins, R.K.H. Galvao, B. Barros Neto and M.F. Pimentel (2007). Calibration transfer in mutivariate methods. Química Nova 30: 1301-1312.

Pastore, T.C.M., J.W.B. Braga, V.T.R. Coradin, W.L.E. Magalhaes, E.Y.A. Okino, J.A.A. Camargos, G.I.B. de Muniz, O.A. Bressan and F. Davrieux (2011). Near infrared spectroscopy (NIRS) as a potential tool for monitoring trade of similar woods: Discrimination of true mahogany, cedar, andiroba, and curupixa. Holzforschung 65(1): 73-80.

Rinnan, A., F.W.J van den Berg and S.B. Engelsen (2009). Review of the Most Common pre-Processing Techniques for Near-Infrared Spectra. TrAC Trends in Analytical Chemistry 28(10): 1201-1222.

Schwanninger, M., J.C. Rodrigues and K. Fackler (2011). A review of band assignments in near infrared spectra of wood and wood components. Journal of Near Infrared Spectroscopy 19: 287-308.

Silva, D.C., T.C.M. Pastore, L.F. Soares, F.A.S. de Barros, M.C.J. Bergo, V.T.H. Coradin, A.B. Gontijo, M.H. Sosa, C.B. Chacon and J.W.B. Braga (2018). Determination of the country of origin of true mahogany (*Swietenia macrophylla* King) wood in five Latin American countries using handheld NIR devices and multivariate data analysis. Holzforschung 72(7): 521-530.

Snel, F.A., J.W. Braga, D. da Silva, A.C. Wiedenhoeft, A. Costa, R. Soares, V.T. Coradin and T.C. Pastore (2018). Potential field-deployable NIRS identification of seven *Dalbergia* species listed by CITES. Wood Science and Technology 52(5): 1411-1427.

Soares, L.F., D.C. da Silva, M. C. J. Bergo, V.T.R. Coradin, J.W.B. Braga and T.C.M. Pastore (2017). Evaluation of a Nir Handheld Device and PLS-DA for Discrimination of Six Similar Amazonian Wood Species. Química Nova 40(4): 418-426.

Teófilo, R.F., J.P.A. Martins and M.M.C. Ferreira (2009). Sorting variables by using informative vectors as a strategy for feature selection in multivariate regression. Journal of Chemometrics 23: 32–48.

Wang, Y., D.J. Veltkamp and B.R. Kowalski (1991). Multivariate instrument standardization. Analytical Chemistry 63 (23): 2750-2756.

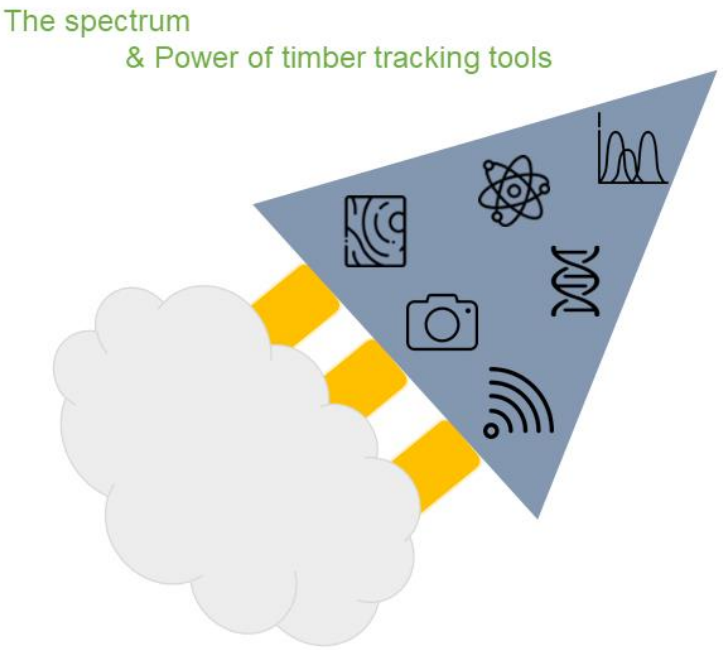
Westad, F. and F. Marini (2015). Validation of chemometric models - A tutorial. Analytica Chimica Acta 893: 14-24.

Williams, P., P. Dardenne and P. Flinn (2017). Tutorial: Items to be included in a report on a near infrared spectroscopy project. Journal of Near Infrared Spectroscopy 25, 85–90.

6. An expert view on the combination of provenancing methods

Authors: Bernd Degen, Victor Deklerck, Micha Horacek, Kathelyn Paredes-Villanueva, Niklas Tysklind, Charlie Watkinson

*Authors are in alphabetical order.



6.1 CURRENT CHALLENGES & FUTURE PERSPECTIVES

NOTE: For taxon identification, methods are already combined as standard. The taxon needs to be identified or verified before the provenance of a wood sample can be analysed via genetics, stable isotopes, DART TOFMS, or NIR spectroscopy. This chapter focuses on combining different methods to identify or check the provenance of wood.

Not all wood identification methods perform optimally under all conditions (*cfr.* the Timber Tracking Tool Infogram). In addition, the biological processes that lead to variation in the anatomical, genetic, and chemical characteristics of wood are spatially heterogeneous and ruled by different processes. This offers the opportunity to combine different wood identification methods to reduce the limitations and enforce the strengths of the single methods. In other cases, only one method might be suitable to do the wood identification. Once similarly extensive reference databases are built for all methods, guidelines can be developed to help judge **which method or combination of methods is most suitable for a given case**. For now, we start with evaluating the different challenges encountered when we want to combine different timber tracking methods and give some future views on potential mitigation measures.

Table 7: The challenges of combining different timber tracking methods and potential mitigation measures.

Challenge	Future perspective
Sampling	
Different sampling designs, database limitations.	If different methods are used, the analyses should be carried out on subsamples of the same sample set. Sampling campaigns should hence collect material to be analysed by all combined methods (see also the GTTN sampling guide).
The importance of variables (*e.g.* age and size of trees, sampling scale and intensity, variation of environmental conditions) differs across methods, requiring different ways of sampling.	The sampling strategy for samples that will be used by several methods should consider all method-specific sample requirements.

Table 7 continued

Challenge	Future perspective
Data analysis	
Different types of data (units, data format, …). The first challenge will be the harmonisation within each method.	Analytical methods could be used that can cope with these differences (classification trees, machine learning, MVA, …). In future, a standard form should be produced that exactly defines how to fill in the respective data into a database.
Different sources of bias, success rate, resolution.	To enable method combinations exhaustive reporting on used statistics and on calculations of assignment success will be crucial. In-depth analyses of the different sources of bias should be made. Their impacts on assignment rates could then be modelled.
Different quality of reference data (linked to sampling intensity and/or strength of the spatial structure in the data).	Self-assignment tests are one way to judge on the quality/performance of the data. Quality requirements will need to be formulated and fulfilled to warrant the quality of a dataset and database.
Interpretation & Reporting	
Different wording and definitions of terms.	Harmonisation of terminology.
Weighing results of different methods on the same material when conclusions go into different directions.	The task of the scientist is to report his/her findings using scientific criteria, unambiguously supporting the scientist´s conclusions. To develop such criteria, which will assist decision makers, the mechanisms behind the conflicting characters need to be understood, allowing judgement of the reliability of the different methods.
Different reporting formats.	Harmonised reporting templates indicating whether same reference samples were used or not, and with enough flexibility to cater for various end-users, countries, legal systems.

Theoretically, there are **different ways to combine wood identification methods**:

- By transforming data in a common data type to be able to analyse all at once.
- By keeping data in different formats but applying a common analysis approach and combining the results (*e.g.* two nearest neighbour analyses)
- By sharing and including method specific sample information in a common data interpretation.
- By selective application of the best performing method for a given wood identification request, combining the methods to provide a broad spectrum of services of optimal quality.

At this moment, the generation of a common data type seems impossible, or at least unsuitable. The standardisation would lead to a loss of information (*e.g.* it is possible to transform continuous data into categorical data, but they will not be as powerful as before). However, below two examples are given of how methods can still be combined.

6.2 CONCRETE EXAMPLES OF HOW METHODS COULD BE COMBINED

6.2.1 USING MAPS AS THE INTERFACE FOR GEOGRAPHIC ORIGIN ASSIGNMENT

The difference in the way the data from the different timber tracking methods behave means the measurements do not endear themselves to a common statistical method for geographic origin verification. However, it may be possible to have a single statistical evaluation that can incorporate all data if Geographic Information Systems are used.

Initially a marker or set of markers must be defined for any given species/genus. Their presence or absence can be geographically related either by the environmental conditions the tree is growing in or by how populations spread over time. The presence or absence of the marker in the sample can be represented as a 1 for presence or a 0 for absence in reference samples. These binary ways of classifying the data can then be used to generate Voronoi diagrams in a Geographic Information Systems platform (*e.g.* QGIS, ArcGIS) which will act as the statistical surface for testing unknown samples.

Voronoi diagrams are a way of clustering data in a distance weighted format, however, only the boundary of the data is distance weighted between samples, not the value. This method lends itself to being suitable to map out data that varies in a binary manner when no other values can exist other than presence or absence.

Table 8: Example of how to write the marker data in binary for creating the Voronoi diagram, this method should be applicable to both SNPs or mass/charge ratios from DART TOFMS.

Nr	Latitude	Longitude	Family Name	Genus	Marker 1	Marker 2
1	44.640478	-72.922832	Sapindaceae	Acer	1	0
2	44.627043	-73.031589	Sapindaceae	Acer	1	0
3	44.627043	-73.031589	Sapindaceae	Acer	0	1
4	44.606542	-72.809634	Sapindaceae	Acer	0	1
5	44.606542	-72.809634	Sapindaceae	Acer	1	0

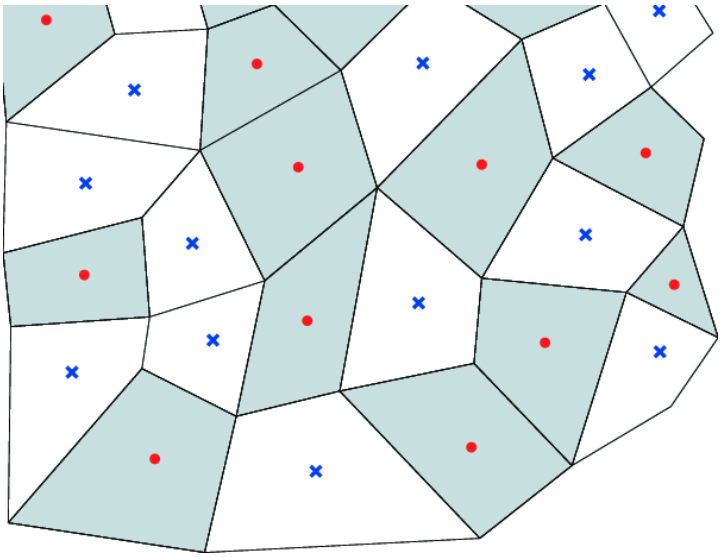


Fig. 20: Example of a binary Voronoi diagram with red circles indicating samples with a value of 0 and blue crosses indicating samples with a value of 1.

Once Voronoi diagrams have been created for each marker in the reference dataset, it will be possible to query the Voronoi diagram statistical surfaces for areas that meet the criteria of the unknown sample. These queries can be overlaid to narrow down the provenance of the reference sample to only include areas where the markers seen in the unknown sample have been seen before in the reference dataset, creating a single geographic shape.

This result can be coupled with the isoscape results from stable isotope analysis. As it is illogical that SNP markers or chemometric markers follow exactly the same geographic patterns as stable isotope ratios, it seems reasonable that using the data from these methods together could produce a highly focussed and resolved geographic origin assignment result that would be much greater than any of the various methods could achieve independently. Furthermore, Voronoi diagram mapping could be a useful tool to apply to DART TOFMS and DNA markers to further investigate patterns in the data and search for natural co-variates for further prediction.

6.2.2 USING THE SOFTWARE *GEOASSIGN* TO COMBINE GENETIC & STABLE ISOTOPE DATA

6.2.2.1 BACKGROUND INFORMATION

In the frame of the ITTO-project "Development and implementation of a species identification and timber tracking system with DNA fingerprints and isotopes in Africa" (Degen & Bouda 2015, Degen *et al.* 2016), reference data and blind test data of Iroko (*Milicia excelsa, M. regia*), sampled in seven African countries, were used to **test the potential advantages of a combined data analysis**.

From **seven African countries** (Cameroon, Democratic Republic of Congo, Gabon, Ghana, Ivory Coast, Kenya and the Republic of the Congo) a total of 1920 samples were sampled for genetic analyses and 474 reference samples for stable isotope analyses. **55 SNP loci** from 1480 individuals served as genetic reference data (for more info see Dainou *et al.* 2016, Blanc-Jolivet *et al.* 2017). For the stable isotope reference data, biogenic main elements isotopes (**D/H, $^{13}C/^{12}C$, $^{15}N/^{14}N$, $^{18}O/^{16}O$**) and the dominantly geologically influenced stable isotopes (**$^{34}S/^{32}S$, $^{87}Sr/^{86}Sr$**) were analysed (AgroIsolab, Germany).

BOX 10: GENERAL REQUIREMENTS FOR BLIND TESTS

➤ The frame conditions of the test must be **transparent** and documented.

➤ The test must **not favour** (*e.g.* sampling from the same spots as where reference samples were taken), **nor disadvantage** (*e.g.* using exclusively old samples, or samples from outside the reference sampling area) any particular method or participating party.

➤ The **same number of samples and sample locations** should have been investigated for the reference data set by all methods to be tested. If different sample numbers/locations were required to describe a geographic region, this should be reported.

➤ The precise wording for the **test results** needs to be agreed on beforehand. The answers have to be clear and unambiguous to warrant an unequivocal assessment of the results. Answers should be given in three categories: (1) 'Right', (2) 'Wrong', and (3) 'No Result'. Samples of the 'No Result' category should not be included in the calculation of the success rate but be reported separately.

[Adapted from Horacek *et al.* 2018, supplemental material.]

6.2.2.2 ANALYSIS METHOD TO COMBINE DATA

For the **assignment of individuals**, a new approach using a **nearest neighbour classification** of individuals to provenance has been applied (Degen *et al.* 2017). All computations were done by the online tool *GeoAssign*. This approach computes pairwise distance measures between all individual reference data and the test data.

- Only individuals with less than 20% of missing data (non-successful SNP-genotyping or failed measurements of isotopes) were part of the calculations.
- For the metric isotope data, the program made a z-transformation of the different elements and calculated then the city-block-distance among the isotope profile of the different individuals.
- For the SNP data in genetics it computed the genetic distance of Gregorius among the multi-locus genotypes of the different individuals (Gregorius 1984).
- Then the reference samples were ordered from the smallest to the largest distance to the test individual. A small distance means similar genetic composition or similar stable isotope profile of reference sample and test individual.

If the declared country of origin was correct, then we expected a higher proportion of reference samples from the declared country of origin among the **5% most similar reference samples** (number of nearest neighbours). **An index I was computed** for all present countries in the reference samples. The index I varied between -1 and +1.

- If the index was 0, the proportion of similar reference samples in the defined neighbourhood was exactly as high as randomly expected according to the sample intensity in the country.
- If the index was positive, there was an excess of similar reference samples compared to the sample proportion.
- If the index was negative, there was a deficit of similar reference samples compared to the sample proportion.

Arithmetic means of the index I of both methods for each case were calculated for the combined data analysis.

6.2.2.3 SELF-ASSIGNMENT SUCCESS OF BOTH METHODS

As an indicator of power of the reference data, we computed the proportion of correctly assigned individuals in self-assignment tests (Cornuet *et al.* 1999). Here the individuals of the reference data were self-classified to the sampled groups using the leave-one-out approach (Efron 1983). **Both the genetic and stable isotope data sets had a relatively weak success rate** of the self-assignment test. This is due to cross border areas of similar genetic or stable isotope composition. The spatial structure of stable isotope and genetic patterns do not fit very well to the country borders. **Strong differences between both methods were observed between countries** (Table 9).

Table 9: Results of the self-assignment test of the Iroko-reference data.

Country	% Correct self-assignment (5%-percentile)	
	Stable isotopes	Genetics
Cameroon	59	44
Congo (Brazzaville)	18	42
Ivory coast	77	98
Democratic Republic of Congo	69	47
Gabon	62	74
Ghana	66	0
Kenya	83	75
Mean	**62**	**54**

6.2.2.4 ASSIGNMENT SUCCESS OF METHOD COMBINATION STRATEGIES

The performance criterion for the data analysis was the number of mistakes on the claims of the blind test (number of cases with right declarations classified as false, and number of cases of undetected false declarations). Claims for blind test samples were accepted if the index I of the declared country had the highest positive value. If not, the claim was rejected.

Findings of the different data analysis strategies:

- Application of both methods independently on the blind test samples lead to 6 mistakes for the stable isotopes and 10 mistakes for the genetics (Fig. 21 a).
- Application of only the method with the best reference data (highest success-rate of self-assignment) came up with a total of 7 mistakes (Fig. 21 b).
- Application of arithmetic means of index I resulted in 4 mistakes (Fig. 21 c).

This example suggests that **a complete combination of methods and data by applying mean indices does improve the precision compared to the best stand-alone method**. But the selective application of methods according to self-assignment success per country was not better than when methods were not combined.

Fig. 21 a-c: Results of the assignment of blind test samples. (a) independent by each method, (b) by method with best reference data, (c) by means of Index-I of both methods. Index-I results in green are cases with correct decisions, in red are cases with wrong decisions, in bold are countries with the highest index-I (most likely provenance within the reference data) the assignment result. Cam Cameroon, CBr Congo Brazzaville, IC Ivory Coast, DRC Democratic Republic of Congo, Gab Gabon, Gha Ghana, Ken Kenya, CAR Republic of Central Africa.

			Stable Isotopes								Genetics							
				Index-I								Index						
ID_Sample	Claim	Solution	Decision	Cam	CBr	IC	DRC	Gab	Gha	Ken	Decision	Cam	CBr	IC	DRC	Gab	Gha	Ken
G2S_O_M1	Gha	Gha	accept	0.211	-0.333	0.000	-1.000	-1.000	**0.421**	0.050	reject	-1.000	-1.000	**0.515**	-0.950	-1.000	0.423	-1.000
G2S_O_M4	Gab	CBr	reject	**0.105**	0.056	-1.000	0.000	0.059	-0.500	-1.000	accept	-0.450	0.136	-1.000	0.028	**0.160**	-1.000	-1.000
G2S_O_M5	DRC	CBr	reject	**0.211**	0.000	-1.000	0.000	0.000	-0.500	-1.000	accept	-0.400	**0.150**	-1.000	0.057	0.100	-1.000	-1.000
G2S_O_M6	Cam	CAR	accept	**0.368**	0.111	-1.000	-0.571	-1.000	0.000	0.050	reject	-0.294	0.060	-1.000	0.086	**0.134**	-1.000	-1.000
G2S_O_M7	Gab	CAR	reject	**0.211**	0.000	-1.000	-0.143	-0.250	-0.500	0.050	accept	-0.588	0.106	-1.000	-0.182	**0.303**	-1.000	-1.000
G2S_O_M10	Gab	DRC	reject	**0.316**	0.056	-1.000	-0.286	-0.750	0.053	-1.000	reject	-0.333	**0.225**	-1.000	-0.125	0.085	-1.000	-0.833
G2S_O_M12	DRC	DRC	accept	0.053	0.111	-1.000	**0.143**	-0.500	0.000	-1.000	accept	0.083	-0.538	-0.800	**0.345**	-0.385	-1.000	-1.000
G2S_O_M14	Cam	Gab	accept	**0.579**	0.000	-1.000	-0.286	-1.000	-1.000	-1.000	accept	**0.119**	0.014	-1.000	-0.083	0.100	-1.000	-1.000
G2S_O_M15	Gab	Cam	reject	**0.421**	0.000	-1.000	-0.857	-0.250	0.105	-1.000	reject	**0.431**	-0.462	-1.000	-0.250	-0.154	-1.000	-1.000
G2S_O_M20	Cam	Cam	accept	**0.632**	0.000	-1.000	-0.571	-1.000	-0.500	-1.000	accept	**0.203**	0.000	-0.800	-0.476	0.129	-0.500	-1.000
BT_2014_518	Gab	CBr	reject	0.158	**0.167**	-1.000	0.000	-0.750	0.000	-1.000	reject	0.014	-0.200	-1.000	**0.381**	-0.467	-1.000	-1.000
BT_2014_526	Gha	Gha	reject	**0.421**	0.000	-1.000	-0.429	-0.500	0.000	-1.000	reject	-0.933	-1.000	**0.515**	-1.000	-0.923	0.408	-1.000
BT_2014_527	DRC	DRC	accept	0.105	0.167	-1.000	**0.143**	-0.500	-1.000	-1.000	reject	**0.121**	-0.231	-1.000	0.113	0.033	-1.000	-1.000
BT_2014_549	Ken	DRC	reject	0.053	-0.333	-1.000	**0.500**	-0.500	-1.000	-1.000	reject	0.017	**0.148**	-1.000	0.132	-0.385	-1.000	-1.000
BT_2014_554	Gab	Cam	reject	**0.474**	0.111	-1.000	-0.571	-1.000	-0.500	0.000	reject	**0.194**	-0.438	-1.000	-0.040	0.120	-1.000	-1.000
BT_2014_566	Cam	DRC	accept	**0.263**	-0.667	-1.000	-0.286	0.176	-0.500	-1.000	reject	-0.235	**0.290**	-1.000	-0.625	0.186	-1.000	-1.000
BT_2014_569	DRC	DRC	accept	0.053	0.111	-1.000	**0.143**	-0.500	0.000	-1.000	reject	0.068	0.000	-1.000	-0.190	**0.194**	-1.000	-1.000
BT_2014_596	Gab	DRC	reject	**0.526**	-0.667	-1.000	-1.000	-0.500	0.211	-1.000	reject	-0.563	**0.328**	-1.000	-0.455	0.172	-1.000	-1.000
BT_2014_599	Gha	Gha	reject	**0.158**	0.056	-1.000	0.143	-0.750	0.000	-1.000	reject	-0.933	-1.000	**0.551**	-0.905	-0.923	**0.347**	-1.000
BT_2014_600	DRC	DRC	reject	-0.500	0.000	-1.000	0.143	**0.235**	-1.000	-1.000	reject	-0.733	0.183	-1.000	-0.900	**0.500**	-1.000	-1.000

Fig. 20 b

			Stable Isotopes								Genetics							
				Index-I								Index-I						
ID_Sample	**Claim**	**Solution**	**Decision**	**Cam**	**CBr**	**IC**	**DRC**	**Gab**	**Gha**	**Ken**	**Decision**	**Cam**	**CBr**	**IC**	**DRC**	**Gab**	**Gha**	**Ken**
G2S_O_M1	Gha	Gha	accept	0.211	-0.333	0.000	-1.000	-1.000	**0.421**	0.050								
G2S_O_M4	Gab	CBr									accept	-0.450	0.136	-1.000	0.028	**0.160**	-1.000	-1.000
G2S_O_M5	DRC	CBr	reject	**0.211**	0.000	-1.000	0.000	0.000	-0.500	-1.000								
G2S_O_M6	Cam	CAR	accept	**0.368**	0.111	-1.000	-0.571	-1.000	0.000	0.050								
G2S_O_M7	Gab	CAR									accept	-0.588	0.106	-1.000	-0.182	**0.303**	-1.000	-1.000
G2S_O_M10	Gab	DRC									reject	-0.333	**0.225**	-1.000	-0.125	0.085	-1.000	-0.833
G2S_O_M12	DRC	DRC	accept	0.053	0.111	-1.000	**0.143**	-0.500	0.000	-1.000								
G2S_O_M14	Cam	Gab	accept	**0.579**	0.000	-1.000	-0.286	-1.000	-1.000	-1.000								
G2S_O_M15	Gab	Cam									reject	**0.431**	-0.462	-1.000	-0.250	-0.154	-1.000	-1.000
G2S_O_M20	Cam	Cam	accept	**0.632**	0.000	-1.000	-0.571	-1.000	-0.500	-1.000								
BT_2014_518	Gab	CBr									reject	0.014	-0.200	-1.000	**0.381**	-0.467	-1.000	-1.000
BT_2014_526	Gha	Gha	reject	**0.421**	0.000	-1.000	-0.429	-0.500	0.000	-1.000								
BT_2014_527	DRC	DRC	reject	0.105	**0.167**	-1.000	0.143	-0.500	-1.000	-1.000								
BT_2014_549	Ken	DRC	reject	0.053	-0.333	-1.000	**0.500**	-0.500	-1.000	-1.000								
BT_2014_554	Gab	Cam									reject	**0.194**	-0.438	-1.000	-0.040	0.120	-1.000	-1.000
BT_2014_566	Cam	DRC	accept	**0.263**	-0.667	-1.000	-0.286	0.176	-0.500	-1.000								
BT_2014_569	DRC	DRC	accept	0.053	0.111	-1.000	**0.143**	-0.500	0.000	-1.000								
BT_2014_596	Gab	DRC									reject	-0.563	**0.328**	-1.000	-0.455	0.172	-1.000	-1.000
BT_2014_599	Gha	Gha	reject	**0.158**	0.056	-1.000	0.143	-0.750	0.000	-1.000								
BT_2014_600	DRC	DRC	reject	-0.500	0.000	-1.000	0.143	**0.235**	-1.000	-1.000								

Fig. 21 c

			Mean stable isotopes + genetics							
				Index I						
ID_Sample	Claim	Solution	Decision	Cam	CBr	IC	DRC	Gab	Gha	Ken
G2S_O_M1	Gha	Gha	accept	-0.395	-0.667	0.258	-0.975	-1.000	**0.422**	-0.475
G2S_O_M4	Gab	CBr	accept	-0.173	0.096	-1.000	0.014	**0.110**	-0.750	-1.000
G2S_O_M5	DRC	CBr	reject	-0.095	**0.075**	-1.000	0.029	0.050	-0.750	-1.000
G2S_O_M6	Cam	CAR	reject	0.037	**0.086**	-1.000	-0.243	-0.433	-0.500	-0.475
G2S_O_M7	Gab	CAR	reject	-0.189	**0.053**	-1.000	-0.163	0.027	-0.750	-0.475
G2S_O_M10	Gab	DRC	reject	-0.009	**0.141**	-1.000	-0.206	-0.333	-0.474	-0.917
G2S_O_M12	DRC	DRC	accept	0.068	-0.214	-0.900	**0.244**	-0.443	-0.500	-1.000
G2S_O_M14	Cam	Gab	accept	**0.349**	0.007	-1.000	-0.185	-0.450	-1.000	-1.000
G2S_O_M15	Gab	Cam	reject	**0.426**	-0.231	-1.000	-0.554	-0.202	-0.448	-1.000
G2S_O_M20	Cam	Cam	accept	**0.418**	0.000	-0.900	-0.524	-0.436	-0.500	-1.000
BT_2014_518	Gab	CBr	reject	0.086	-0.017	-1.000	**0.191**	-0.609	-0.500	-1.000
BT_2014_526	Gha	Gha	accept	-0.256	-0.500	-0.243	-0.715	-0.712	**0.204**	-1.000
BT_2014_527	DRC	DRC	accept	0.113	-0.032	-1.000	**0.128**	-0.234	-1.000	-1.000
BT_2014_549	Ken	DRC	reject	0.035	-0.093	-1.000	**0.316**	-0.443	-1.000	-1.000
BT_2014_554	Gab	Cam	reject	**0.334**	-0.164	-1.000	-0.306	-0.440	-0.750	-0.500
BT_2014_566	Cam	DRC	reject	0.014	-0.189	-1.000	-0.456	**0.181**	-0.750	-1.000
BT_2014_569	DRC	DRC	reject	**0.061**	0.056	-1.000	-0.024	-0.153	-0.500	-1.000
BT_2014_596	Gab	DRC	reject	**-0.019**	-0.170	-1.000	-0.728	-0.164	-0.395	-1.000
BT_2014_599	Gha	Gha	accept	-0.388	-0.472	-0.225	-0.381	-0.837	**0.174**	-1.000
BT_2014_600	DRC	DRC	reject	-0.617	0.092	-1.000	-0.379	**0.368**	-1.000	-1.000

6.2.3 USING *GENELAND* AND *ADEGENET* TO COMBINE GENETIC AND PHENOTYPIC DATA

Geneland and *Adegenet* offer a **common analytical environment for the different wood identification tools** (Jombart *et al.* 2010, Guillot *et al.* 2012).

- *Geneland* is a geographically explicit genetic analysis programme that allows the inclusion of phenotypic data (*e.g.* wood anatomy, isotopic, and chemical composition) into the detection of clusters and the assignment of individuals. By comparing the geographical structure obtained with (i) only genetic markers, versus (ii) both genetic and phenotypic markers, we can assess the impact phenotypic markers have in allowing us to detect spatial structure and improve individual assignment.
- *Adegenet* transforms genotypic data into data that can be analysed in an interpretable manner through multivariate analyses (MVA), which can be applied to any numerical type of data (*e.g.* phenotypic data such as morphological characteristics, or microchemistry). Qualitative and categorical types of data can be carefully transformed into numerical data.

The different structures obtained with the different markers independently (*i.e.* genetic, isotopic, chemical) can be compared with each other, which can help us **understand the processes leading to the heterogeneity in the geographical structure of the signals** produced by each type of marker.

Finally, by applying different weights to the different types of markers in different species, we could **explore the relative contribution of each marker set to the most exact provenance assignment**.

6.2.4 KEY LITERATURE FOR METHOD COMBINATIONS

Blanc-Jolivet, C., B. Kersten, K. Dainou, O. Hardy, E. Guichoux, A. Delcamp and B. Degen (2017). Development of nuclear SNP markers for genetic tracking of Iroko, Milicia excelsa and Milicia regia. Conservation Genetics Resources 9(4): 531-533.

Cornuet, J.M., S. Piry, G. Luikart, A. Estoup and M. Solignac (1999). New methods employing multilocus genotypes to select or exclude populations as origins of individuals. Genetics 153(4): 1989-2000.

Dainou, K., C. Blanc-Jolivet, B. Degen, P. Kimani, D. Ndiade-Bourobou, A.S.L. Donkpegan, F. Tosso, E. Kaymak, N. Bourland, J.L. Doucet and O.J. Hardy (2016). Revealing hidden species diversity in closely related species using nuclear SNPs, SSRs and DNA sequences - a case study in the tree genus Milicia. Bmc Evolutionary Biology 16.

Degen, B., C. Blanc-Jolivet, K. Stierand and E. Gillet (2017). A nearest neighbour approach by genetic distance to the assignment of individual trees to geographic origin. Forensic Science International-Genetics 27: 132-141.

Degen, B. and H. Bouda (2015). Verifying timber in Africa. ITTO Trop Forest Update 24(1):8-10.

Degen, B., H.-N. Bouda, C. Blanc-Jolivet (2016). Development and implementation of a species identification and timber tracking system with DNA fingerprints and isotopes in Africa. ITTO project PD 620/11 Rev.1 (M)

Efron, B. (1983). Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation. Journal of the American Statistical Association 78(382): 316-331.

Gregorius, H.R. (1984). A unique genetic distance. Biometrical Journal 26(1): 13-18.

Guillot, G., S. Renaud, R. Ledevin, J. Michaux and J. Claude (2012). A Unifying Model for the Analysis of Phenotypic, Genetic, and Geographic Data. Systematic Biology 61(6): 897-911.

Horacek, M., G. Rees, M. Boner and J. Zahnen (2018). Comment on: Developing forensic tools for an African timber: [...], by Vlam *et al.* 2018. Biological Conservation 226: 333-334.

Jombart, T., S. Devillard and F. Balloux (2010). Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genetics 11: 94.

Paredes-Villanueva, K. (2018). Tropical timber forensics: a multi-methods approach to tracing Bolivian *Cedrela* (Doctoral dissertation, Wageningen: Wageningen University).

7. Appendices

7.1 APPENDIX 1: LIST OF CITES-PROTECTED TRADE TIMBERS IN THE DATABASE *CITESwoodID*

Botanical name	Trade name	CITES Appendix	Footnote
Aniba rosaeodora	Pau rosa	II	#12
Aquilaria spp., *Gyrinops* spp.	Eaglewood	II	#14
Araucaria araucana	Pehuén, pino araucano	I	*
Bulnesia sarmientoi	Palo santo	II	#11
Paubrasilia echinata (syn. *Caesalpinia echinata*)	Pau Brasil, Fernambuc	II	#10
Cedrela odorata	Cedro	II	#5
Dalbergia nigra	Brazilian Rosewood	I	*
Dalbergia spp.	Rosewood (in general)	II	#15
Diospyros spp.	Ébène de Madagascar	II	#5
Dipteryx oleifera	Almendro	III	no
Fitzroya cupressoides	Alerce	I	*
Fraxinus mandshurica	Manchurian ash	III	#5
Gonystylus spp.	Ramin	II	#4
Guaiacum spp.	Lignum vitae	II	#2
Guibourtia spp.§	Bubinga, kevazingo	II	#15
Osyris quadripartita	East African sandalwood	II	#2
Pericopsis elata	Afrormosia, Kokrodua	II	#17
Pilgerodendron uviferum	Ciprés de las Guaitecas	I	*
Pinus koraiensis	Korean pine	III	#5

Platymiscium pleiostachyum	Granadillo	II	#4
Podocarpus spp. §	Podo, Maniu	I and III	#1
Pterocarpus erinaceus	Vène, kosso	II	#1
Pterocarpus santalinus	Red sanders	II	#7
Quercus mongolica	Mongolian oak	III	#5
Swietenia spp.§	True Mahogany	II	#4, #5, #6
Taxus spp. §	Yew	II	#2

LEGEND:

§ While the genus description is included in the *CITESwoodID* database, not all species of these genera are CITES listed.

***** For species listed under CITES Annex I, there is absolute protection without exceptions. For this reason, no footnote(s) is(are) included.

Footnote #1: All parts and derivatives, except: seeds, cut flowers and fruits

Footnote #2: Designates all parts and derivatives except: seeds and pollen; and finished products packaged and ready for retail trade

Footnote #4: Designates all parts and derivatives, except: seeds and fruits

Footnote #5: Designates logs, sawn wood, and veneer sheets.

Footnote #6: Designates logs, sawn wood, veneer sheets and plywood.

Footnote #7: Logs, wood chips, powder, and extracts

Footnote #10: Logs, sawn wood, veneer sheets, including unfinished wood articles used for the fabrication of bows for stringed musical instruments

Footnote #11: Logs, sawn wood, veneer sheets, plywood, powder, and extracts. Finished products containing such extracts as ingredients, including fragrances, are not considered to be covered by this annotation

Footnote #12: Logs, sawn wood, veneer sheets, plywood, and extracts. Finished products containing such extracts as ingredients, including fragrances, are not considered to be covered by this annotation

Footnote #14: All parts and derivatives except: seeds fruits and leaves

Footnote #15: All parts and derivatives are included, except: Leaves, flowers, pollen, fruits, and seeds; Non-commercial exports of a maximum total weight of 10 kg per shipment

Footnote #17: Logs, sawn wood, veneer sheets, plywood and transformed wood.

7.2 APPENDIX 2: EXTENDED INFO ON EXPLORATION, CHECKING, AND ANALYSES OF GENETIC DATA

7.2.1 DECISION TREES FOR PROVENANCE & SPECIES IDENTIFICATION

Diagram showing the general steps required for the process of data analysis for provenance identification.

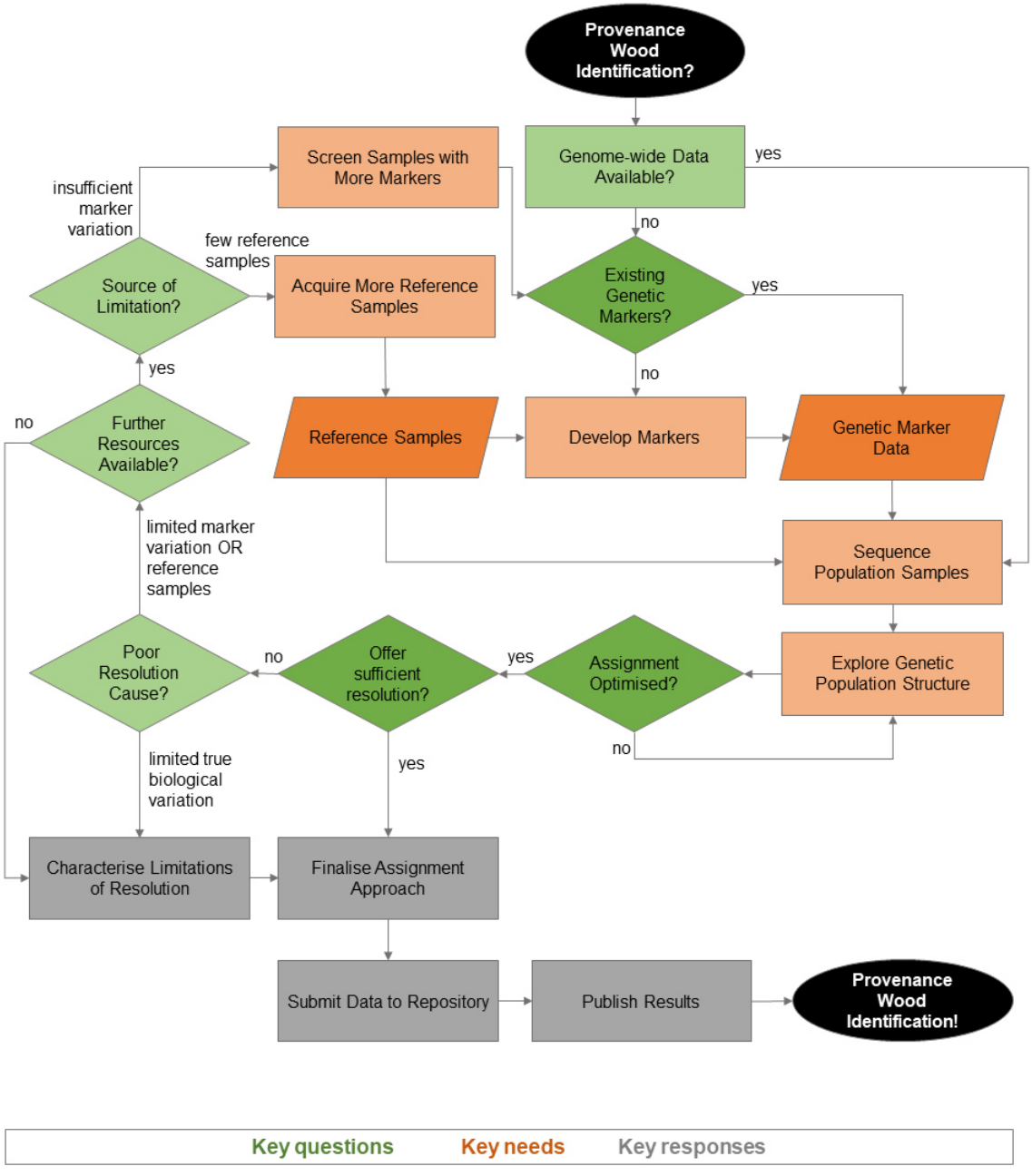
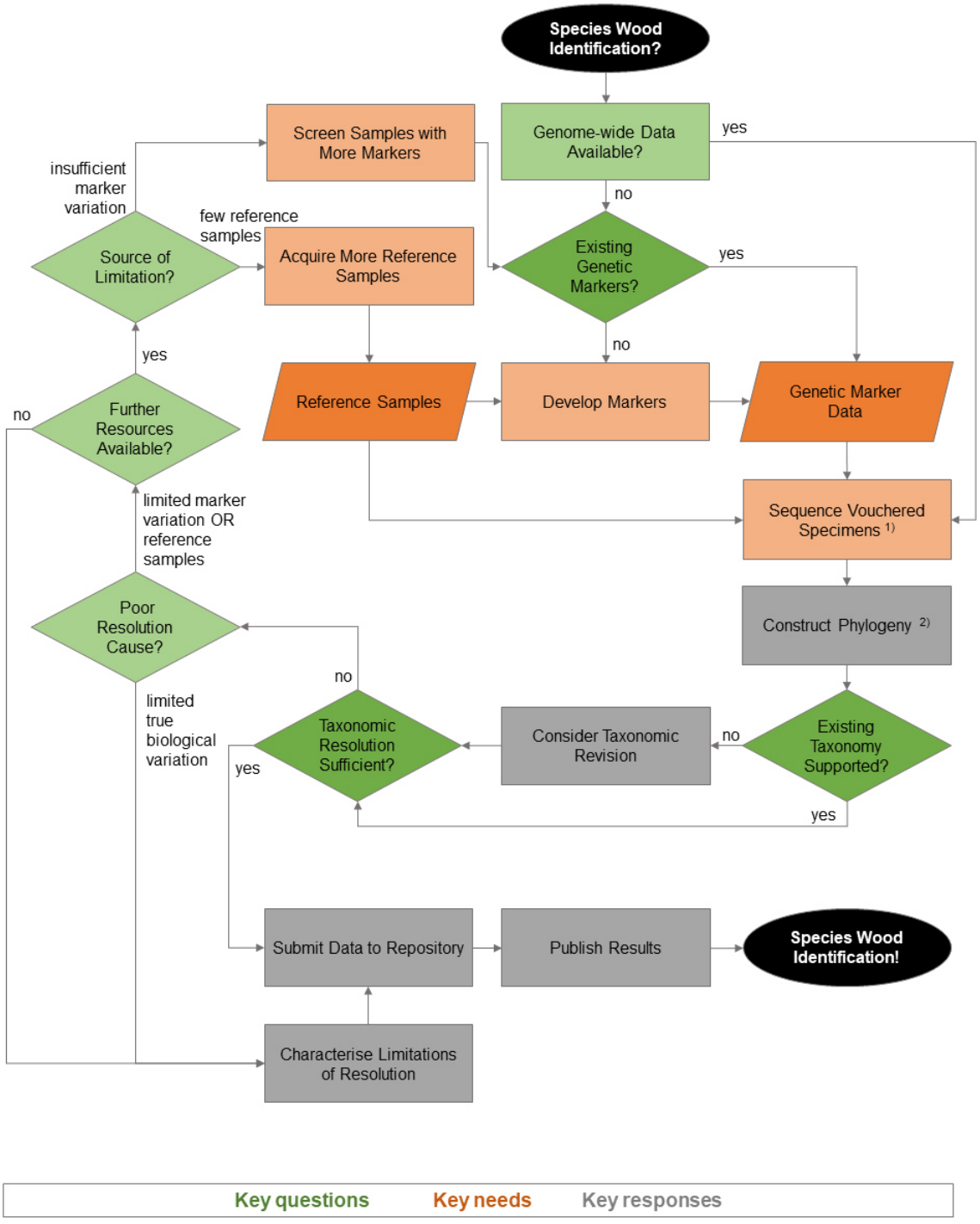


Diagram showing the general steps required for the process of data analysis for species identification.



[1] Complemented with additional, unvouchered specimens to overcome potential sample size limitations.

[2] Phylogeny is not necessary/possible if you make use of known molecular markers. A grouping/cluster analysis can be sufficient.

7.2.2 SOFTWARE USED IN CLUSTERING ANALYSES

Table 11: Evaluation of some of the most used software for genetic data analysis.

Software	Advantages[32]	Disadvantages
Clustering		
Geneland (+ *CLUMPP*) (Guillot *et al.* 2008)	Geographically explicit clustering method that can also include phenotypic data (*e.g.* wood morphology, chemistry). Takes a wide range of genetic input data (diploid, haploid, co-dominant, and dominant).	Computationally intensive for large datasets. Requires basic *R* knowledge.
STRUCTURE (+ *CLUMPP*) (Pritchard *et al.* 2000; Jakobsson & Rosenberg 2007; Earl & vonHoldt 2012)	*STRUCTURE* takes haploid, diploid and polyploid genotypes (through the recessive alleles parameter), as well as information such as population, sampling location and phenotype. Missing data is not a big issue, unless very little data is available or if there are a lot of null alleles.	Model assumptions are HWE and LE. Computationally intensive for large datasets. Not spatially explicit[33].
Clustering testing		
assignPop (Chen *et al.* 2018)	Identification of uninformative and most informative loci. Integrated data visualisation.	Requires basic *R* knowledge.
GeneClass (Piry *et al.* 2004)	Possible to estimate exclusion probabilities during self-assignment procedure.	Most statistics require HWE and LE.
Oncor	Bayesian Clustering method, extensive self-assignment tests options (leave-one-out procedure)	

[32]All *R* packages (see Table 2) have the advantage that *R* scripts are reproducible and publishable.

[33] This might be of advantage with datasets containing 'cryptic' species.

Table 11 continued

Software	Advantages	Disadvantages
Assignment		
assignPop	Broad capabilities of testing assignment power and accuracy varying the number of individuals and markers to be included based on machine learning procedures. Can also include phenotypic data (*e.g.* wood morphology, chemistry).	Requires basic *R* knowledge.
GDA-NT	Conducts assignment tests and self-assignment. Uses different criteria for the likelihood estimation	Most statistics require HWE and LE.
GeneClass	Possible to estimate exclusion probabilities during self-assignment procedure.	Most statistics require HWE and LE.
Geneland	Allows estimation of assignment probability of individuals to populations.	Computationally intensive for large datasets. Requires basic *R* knowledge.
GeoAssign	Grouping of reference samples can contain admixed individuals.	
Oncor	Allows estimation of assignment probability of individuals to populations.	
STRUCTURE	Interesting option POPINFO to assign unknown samples when reference clusters are available.	No self-assignment possible.

7.2.3 DATA EXPLORATION WITH MVA

Multivariate analysis (MVA) of genetic data is a good way to first approach a genetic dataset and getting a quick glance at the structure and strength of signal in the dataset, which will indicate which other types of analysis we should aim for (*e.g. STRUCTURE* with/without admixture and allele correlation, *Geneland* with/without allele correlation). *adegenet* (see Table 2-3) offers at least three different types of MVA:

- Principal component analysis at the individual level. Each individual is represented by a dot; the closer two dots are, the more genetically similar they are (shared alleles); dots can be colour coded based on any *a priori* variable we want (*e.g.* species, population samples, inferred populations, collection dates, phenotypes, *etc.*). PCA gives us an idea of the strength of the structure (dots clustered or spread), and the overlap among the different groups.
- Correspondence analysis (CA) at the population sample level. Each population sample is represented by a dot; the closer two dots are, the more genetically similar they are (allele frequencies); dots can be colour coded based on any *a priori* variable we want (*e.g.* species, population sample, inferred population, collection date, *etc.*). CA can be especially useful to start hinting populations, and which sampling locations may belong to the same population.
- Discriminant analysis of principal components at the individual level. This method allows the estimation of the number of genetic clusters (*i.e.* species or populations) and the membership probability of each individual to each genetic cluster. The results can then be visualized through dot-plots, table-plots, or structure-like-plots.

A typical *adegenet* analysis would start with an exploratory PCA where *a priori* known variables (*e.g.* species, biogeographical regions, suspected populations, or population sample codes) are used for colour coding. The different axes of the PCA should be explored. *adegenet* allows separating the data across loci (for example with microsatellites), and then do single-loci PCAs to compare the structure signal across loci. Subsequently, a CA will show us which population samples are more similar to each other and whether there is any obvious clustering among population samples, suggesting true populations. Finally, the DAPC can be explored thoroughly to evaluate the best estimate of *K*, the number of clusters (true biological populations). DAPC of different *K* values can then be compared with each other to look at the stability of individual assignments across different *K* values. The loadings of the different axes of DAPC can be used to identify which loci contribute most to the

signal contained in each axis. This way we can identify loci which are most informative for particular questions.

7.2.4 DATA CHECKING

A possibility for checking the data is to evaluate the distribution of missing data across the different species, sampling locations, inferred populations, and individuals, and make an informed decision based on the (non-)randomness of the presence of null alleles. *poppr* allows the visualisation of the percentage missing data per loci across population samples, individuals, and loci (Fig. 22). Whole samples, single individuals, and/or loci can then be removed through a cut-off threshold (*e.g.* allowing 50%, 25%, 5% missing data).

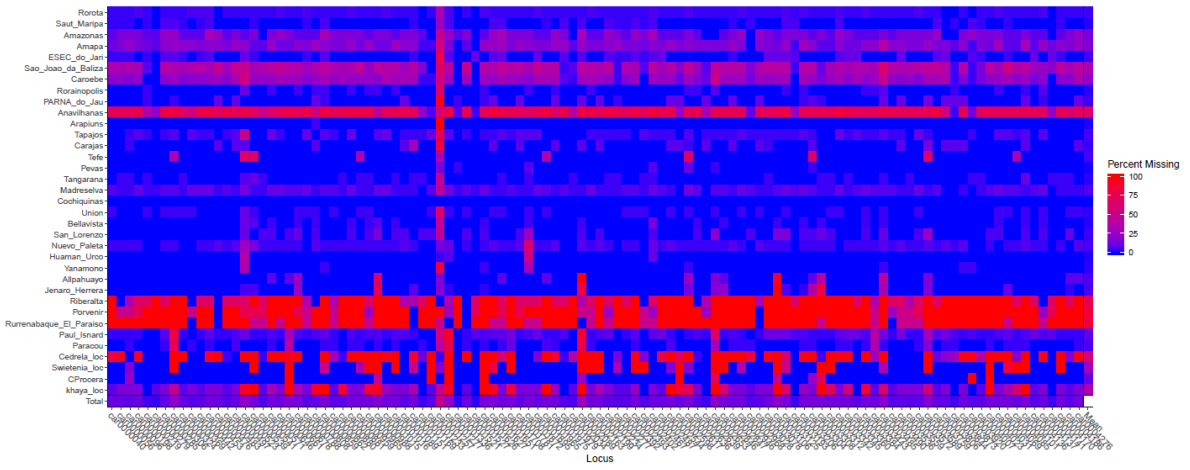


Fig. 22: Graph showing how missing genetic data are not randomly distributed. Percentage of missing data for SNPs (vertical lines) across 35 population samples (horizontal lines) containing over 800 individuals of *Carapa* and other Meliaceae species. Certain loci have high rates of non-amplification regardless of sample (vertical red lines), and certain samples have high rates of non-amplification (up to 100%) regardless of loci (horizontal red lines). The samples towards the bottom of the graph were species miss-identified in the field (*Cedrela*), while the samples towards the top with 25-50% missing data were samples with poor quality DNA.

Several population genetics analysis programmes (*poppr*, *assignPop*) allow the identification and removal of **uninformative loci** at this stage. The cut-off value can be either set manually or set as a default (*i.e.* fewer than x variant observations in the whole dataset, or MAF < 0.01). It is worth running this analysis to check the proportion of uninformative loci in the dataset, even if the uninformative loci are not subsequently removed, to verify there are no problems with the genotyping.

Identical genotypes can occur in the dataset and could be indicative of three different phenomena that need to be dealt with differently: (1) contamination (technical error), (2) lack of discriminatory power with the number of loci (statistical weakness), or (3) clones occurring in the study species (true biological signal). *poppr* allows identifying identical genotypes at several levels (strata: *e.g.* population sample, inferred population, species) or at the whole dataset level, and can remove identical genotypes from the dataset. Whether identical genotypes should be removed or not will depend on the nature of the identical genotypes. If **contamination** is suspected (*e.g.* single individual sampled repeatedly, labelling errors, or identical genotypes in nearby wells within a genotyping plate), then identical genotypes should be removed as they do not represent true biological signals. If we do **not have enough discriminatory power** to distinguish individuals, we should carefully consider the implications in terms of an individual assignment exercise and acknowledge such lack of statistical power. If **clones occur in the target species**, we should consider the frequency of clonal individuals in each population sample, and the potential implications in the individual assignment exercises. Variance in the frequency of clones in each inferred population will strongly impact assignment probabilities and accuracy.

Once we have removed all dubious and missing data, we can produce graphs of sample sizes (number of individuals per population sample), number of loci, and allelic richness per population sample. We can also state the percentage loss of data through the quality checking stage.

7.2.5 BAYESIAN CLUSTERING ALGORITHMS

7.2.5.1 *STRUCTURE*

STRUCTURE includes Bayesian clustering algorithms, which group individuals into genetic clusters, trying to maximise HWE. **Model assumptions** are based on HWE and no LD, but the method seems to be quite robust to model deviations. The user can choose between different models: no mixture/admixture, correlation of allele frequencies or not, prior population information or not. *STRUCTURE* takes haploid, diploid, and polyploid genotypes (through the recessive alleles parameter), as well as information such as population, sampling location, and phenotype. **Missing data** rate is not a big issue, unless very little data is available or if there are many null alleles.

It is recommended to use **the admixture model with correlated allele frequencies** as it usually better reflects real populations. It usually better detects subtle structure (closely related populations), for example if two clusters originated through isolation-by-distance. However, it might overestimate the number of clusters. **The**

independent allele frequencies model can be used if the allele frequencies are expected to be quite different.

Some authors recommend as well to have a rather balanced sampling design, as *STRUCTURE* might not detect all underrepresented true clusters and tend to merge them. The question is however how to select the individuals without losing potential cryptic groups and how to deal with uneven sampling design, which is very frequent in tropical species (due to species abundance and population accessibility).

The input file contains an ID number for each individual, an ID for the population and the genotype. The user can specify how the missing data should be coded (0, -1, -9, *etc.*). The next step is to **choose the model, the number of burn-in and MCMC iterations**, **the number of K clusters to test and the number of runs for each K value** (for guidance 100,000 burn-in, 100,000 MCMC and 30 repeats for each K are recommended). The software runs on *Windows* through a user interface, but running the software on command-line through *R* is more effective and runs can even be done in parallel on different cores on a *linux* system (function 'parallel.structure').

Outputs for each single run (one repeat for one given *K* value) include the individual probabilities of membership of each cluster and should only include those. All the outputs can be zipped together and run with *CLUMPACK* to visualize the bar plots of cluster membership for each *K* value. Put together with results provided by *StructureHarvester*, the optimal *K* value can then be chosen. Because differentiation levels among clusters can strongly differ, **one single analysis is rarely enough to catch all subtle structure within the dataset**. Using the optimal *K* value from the first analysis and the *CLUMPACK* output for this *K* value, each individual will be assigned to one of the clusters or considered as admixed (*e.g.* probability of belonging to a cluster <0.8). Then the analysis will be repeated for each cluster separately until no further structure is observed (too many admixed individuals).

The option POPINFO is especially useful to assign unknown samples when reference clusters are available (from a previous analysis). This might be useful to enlarge existing reference data. However, how this method performs in comparison to classical genetic assignment algorithms for the analysis of test samples is unknown (no self-assignment possible).

Pros & Cons

Pros: *STRUCTURE* is one of the most popular software for Bayesian Clustering analysis and downstream analysis with CLUMPACK is straightforward. It is also easy to use different methods to look at optimal *K* value. The options to use prior information on the sample can be explored and is widely used in publications dealing

with hybridization among species within a genus (taxonomically identified 'pure' individuals can be used to force *STRUCTURE* to assign samples with unclear species ID to the reference individuals, and thereby also indicating if hybridization occurred). The possibility to call *STRUCTURE* from *R* enables scripting of many runs, thus making easier to use computing clusters.

Cons: This method requires good computing facilities and time to prepare input files. The model is not spatially explicit, but this might be of advantage with dataset containing 'cryptic' species.

7.2.5.2 *GENELAND*

Geneland is a geographically explicit **genetic-assumptions-based** Bayesian clustering algorithm that allows the estimation of geographically contiguous and genetically consistent populations from georeferenced and genotyped individuals. It works best when individuals are sampled randomly across the area of interest (*e.g.* species range) but can also cope with individuals collected as population samples (*i.e.* several individuals from the same sampling location). *Geneland* uses co-dominant or dominant genetic markers as well as sequence data, it can cope with null alleles, and can also include phenotypic data (*e.g.* morphology, microchemistry).

Input files consist of text files containing individual genotypes and geographical coordinates. The analyst will have to choose which type of allele frequency model, uncorrelated or correlated allele frequencies, is most appropriate for the system under study. **Knowledge on the biology of the organism is key here**. Populations frequently exchanging migrants or recently separated will have strongly correlated allele frequencies. Conversely, different congeneric species will probably have uncorrelated alleles frequencies. Also, you need to evaluate whether you want to maximise your chances of detecting subtle population structure, or whether you would rather make sure you only detect structure that is strong and probably biologically relevant. **It is recommended to start with the uncorrelated model and then assess the effects of changing to a correlated model**. *Geneland* can cope with and estimate the frequency of null alleles operating at each locus. It can also include spatial coordinate uncertainty, which can be useful to detect genetic outliers (*e.g.* migrants).

When running the analysis, you have to **choose the number of iterations and the frequency at which you sample iterations**, which will depend on the number of individuals and the strength of the structure in the system. To know the number of iterations needed, you will need to do several runs (*e.g.* 10) and evaluate the stability of the results. For exploratory analysis of an average dataset 100,000 iterations is a

good starting point, however, for publication, 1,000,000 iterations are probably required.

Pros & Cons

Pros: *Geneland* is a powerful tool for estimating the number of panmictic populations and the individual membership probability to each population. The inclusion of geographic coordinates makes it particularly pertinent for the construction of reference databases for forest tree species. The basic analysis is relatively simple to use, but *Geneland* has many add-ons (*e.g.* inclusion of phenotypes, GIS, simulation of genetic data) that can prove useful in the context of timber traceability.

Cons: *Geneland* is computationally intensive (processing power and time) and may require large amounts of disk space.

7.2.5.3 GDA-NT

GDA-NT is a population genetics analysis software. It calculates allele frequencies, effective number of alleles, fixation index within each group or population as well as F_{ST}, Hedrick's F'_{ST}, Rst, Gregorius' delta, Nei's genetic distance, Gregorius' genetic distance among populations. *GDA-NT* includes a **locus quality option**. After automatic removal of loci, which have missing data at all individuals at least in one group, average allele call rate, diversity, F-index, differentiation, and correlation among genetic and geographic distance can be estimated for all loci. This is useful to select loci for cost-effective genotyping. *GDA-NT* also estimates **Gregorius' genetic distance** among pairs of populations (which is a metric distance) and differentiation (delta). Based on genetic distance among groups, an output for PAST can be created to visualize groups with an UPGMA dendrogram. The reference data can be directly checked for their performance in genetic assignment with a leave-one-out procedure and calculation of several well-known likelihood criteria.

The different data analysis steps

GDA-NT takes a .csv file or a .txt. The first lines contain the number of populations, the number of loci, the highest value expected for an allele, the loci names, followed by the genotypes arranged per population. In the population labels, coordinates can be given.

Assignment of test samples

For the assignment of test samples, *GDA-NT* takes the input file from the reference population and another input file with the test samples. Test samples are coded like populations in the input, or it is possible to make a group assignment by coding batches of genotypes as populations. The genetic assignment module contains several algorithms to estimate likelihoods for genetic assignment and allows exclusion probabilities calculation with simulations.

Test samples always need to be analysed twice (two independent DNA extractions), and it is often useful to conduct PCR with different amounts of DNA template. A consensus genotype can be added in the analysis. It is however important to code each repeat of the same test sample as 'population' (no group analysis).

The use of the software, as for all other methods, is not the trickiest step for the analysis. Care should be taken at the reference data and the loci used for identification.

Pros & Cons

Pros: *GDA-NT* includes options (loci quality check, self-assignment) and estimates which are not available in most software. No strong computer requirements.

Cons: The software crashes very often and input/outputs are not in an easy format. The results of the self-assignment do not include exclusion probability calculation (gives only the percentage of samples which are more likely to stem from a given population). This option is only available for test samples.

7.2.5.4 *ASSIGNPOP*

assignPop offers an analytical framework for **assessing the reliability and accuracy of reference databases for both genetic and phenotypic data** (*e.g.* morphology, microchemistry).

assignPop works by first, splitting the data into training and test datasets, evaluates the power of each feature (*i.e.* SNPs, microelement) to detect the defined populations, and then estimates the assignment accuracy based on the success of assignment of the test dataset to the training dataset. *assignPop* allows specifying how many individuals per population or the % of each population to include in the training dataset, as well as testing a range of values to evaluate the impact of the number of individuals included in each population in the reference database. It also allows specifying how many times to do the resampling to estimate confidence

intervals. *assignPop* offers five different classification models and has integrated graphic tools to visualise the effects of varying the number of individuals and loci included in the analysis. You can subset the number of loci included in the analysis by either random selection or high-grading, which allows you to compare the assignment accuracy of your baseline with all loci or with only the best features of your dataset (*e.g.* best 25 SNPs).

Pros & Cons

Pros: *assignPop* improves on already available reference database testing analyses, by making it possible to run many tests sequentially across a range of values (% of individuals, % of loci, information content, classification method, *etc.*). Furthermore, the possibility of including the analyses in *R* scripts allows for replication and publication of the analytical procedure.

It is applicable for both genetic and phenotypic data, which allows direct comparison of the assignment power of different identification tools for timber tracking.

Cons: It requires some basic *R* training.

7.2.5.5 *GeoAssign*

GeoAssign uses a nearest-neighbour approach to assign individuals to a user-defined group. A reference **input** file and a test input file need to be prepared. In the reference file, geographical coordinates for all individuals, together with the group/population/genetic cluster are necessary. In the test file, the *a priori* claim (*i.e.* declared species or provenance) should be indicated.

It is possible to test the reference data against itself to proceed to self-assignment (reference data quality). The **output** will contain results for each sample separately but also summary statistics (percentage of samples correctly assigned). *GeoAssign* returns two outputs: one indicating the most likely population/group of each test sample/genotype as well as the exclusion probability for the most likely group. *GeoAssign* compares the genetic distance of the test samples with all reference samples and checks within each group how many 'neighbour' samples they contain in comparison to what you would expect with a random distribution. Likely populations are the ones containing more neighbours than expected (significant positive index). This first output provides all index values for each reference group and help the user in case there is more than one likely group. The second output provides exclusion probabilities for each group. Samples are assigned to the most-likely group with a non-significant exclusion probability. If the complete output is requested, a file for

each test sample giving the genetic and geographic distances to all reference samples sorted from the most-related to less-related will be provided. It is then easy to see whether the genetic neighbours are also the geographic neighbours.

Pros & Cons

Pros: Web-based application. It is possible to use the software for metric data (isotopes) or for genetic data. Detects identical genotypes. Less sensitive to substructure within the groups of reference individuals and does not require HWE nor LE (based on Gregorius' genetic distance).

Cons: Some genetic clusters might have a large distribution, in this case the interest of the method over other genetic assignment algorithms is decreased.

7.3 APPENDIX 3: BACKGROUND INFO ON ISOTOPES

7.3.1 THE ORIGIN OF CHEMICAL COMPOUNDS IN WOOD

Trees are complex organisms that take water and carbon dioxide and fix them into sugars by **photosynthesis**. Those sugars are then subsequently re-arranged into gradually more complex compounds that go on to form structural and functional elements of the tree. Trees are also able to take other compounds from the soil and the air by varying means such as **symbiotic bacteria or fungi** through their roots. As the origin of most compounds within a tree is from photosynthesis, it is helpful to think in context of that metabolic system.

$6H_2O + 6CO_2 \rightarrow C_6H_{12}O_6 + 6O_2$ (general chemical equation for photosynthesis)

7.3.2 THE ORIGIN OF STABLE ISOTOPE RATIOS IN WOOD

Stable isotope analysis is a means by which the origin of compounds can be tracked. Stable isotope patterns in tree wood document the ambient environmental conditions such as topography, geography, climate, soil, atmosphere, available water, emissions.

Almost all elements exist in multiple forms, called isotopes. Isotopes contain the same number of electrons and protons but differ in the number of neutrons making their atomic mass (expressed in Atomic Mass Units, AMU) different. In a periodic table, the atomic mass for any given element is an average representing the abundance and mass of the various isotopes for these elements (Table 10). Some elements have stable isotopes, these are forms of the element that do not undergo radioactive decay or have such long half-lives that it might as well be concluded that they do not undergo significant radioactive decay. Most elements that have stable isotopes are light elements (*cfr.* the periodic table).

Table 10: Stable isotopes typically measured in wood and their origin[34].

Element	Symbol	Natural abundance (%)	Mass (AMU)	Origin/cause of the isotope ratio's in trees
Hydrogen	^{1}H	99,99	1,008	Source of water, environmental conditions, drought stress (influencing plant metabolism and transpiration).
	^{2}H	0,01	2,014	
Carbon	^{12}C	98,89	12,000	Source of carbon dioxide, drought stress, plant metabolism (C3, CAM, C4 plant).
	^{13}C	1,11	13,003	
Nitrogen	^{14}N	99,63	14,003	Source of bio-available nitrogen, plant metabolism (non/leguminous plant).
	^{15}N	0,37	15,000	
Oxygen	^{16}O	99,76	15,995	Source of water and carbon dioxide, environmental conditions, drought stress (influencing plant metabolism and transpiration).
	^{18}O	0,20	17,999	
Sulphur	^{32}S	95,00	31,972	Plant-available sulphur in soil/bedrock, atmosphere (*e.g.* sea-spray).
	^{34}S	4,22	35,967	
Strontium[35]	^{86}Sr	9,86	85,909	Plant-available strontium in soil/bedrock, water, atmosphere (*e.g.* sea-spray).
	^{87}Sr	7,04	86,908	

As the chemical properties of an element are dictated by their charge rather than their mass, isotopes of elements cannot be chemically separated, they must be physically separated. In the natural world there can be many opportunities for elements to physically separated/sorted in some manner. These events are often referred to as '**fractionations**.' For example, distillation processes can influence the abundance of heavy oxygen and hydrogen isotopes in the distillate. This is possible because it takes a greater amount of free energy to 'boil' a heavy water molecule than it would an ordinary water molecule. The net effect of distillation is that the distillate is lower in abundance of heavy isotopes whereas the source liquid becomes greater in abundance of heavy isotopes post-distillation. **In nature, the greatest form of distillation is the water cycle** where ocean water forms clouds that precipitate water.

[34] References: Thode *et al.* 1961, Seal 2006, Horacek *et al.* 2009, 2018, Gori *et al.* 2013.

[35] Strontium is not a major element in trees but as it has the same charge as calcium, it often occupies places that would otherwise contain calcium.

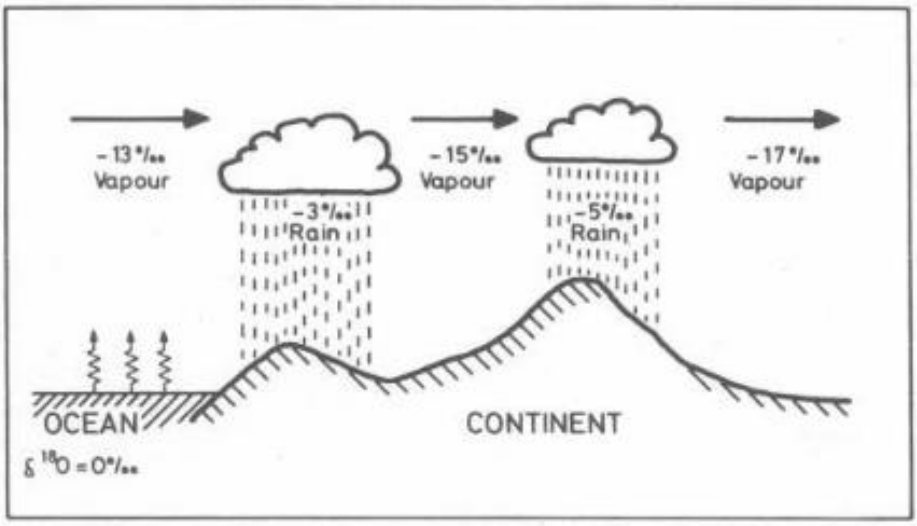


Fig. 23: Schematic of oxygen isotope ratio fractionation of ocean water to precipitation (Siegenthaler 1979).

As clouds lose water in the form of precipitation, the process can be considered a fractional distillation as the liquid water phase (rain) is continually removed, causing further changes in the isotopic composition of the water vapour in the cloud as the cloud progresses.

7.3.3 THE GEOGRAPHIC ORIGIN OF WOOD

Since the photosynthesis and fractionation processes vary with geographical variables (such as temperature, rainfall, presence of symbiotic bacteria and fungi, topography, geology, …) stable isotope ratios can be used to determine/control the geographic origin of wood.

REFERENCES:

Gori, Y., R. Wehrens, M. Greule, F. Keppler, L. Ziller, N. La Porta and F. Camin (2013). Carbon, hydrogen and oxygen stable isotope ratios of whole wood, cellulose and lignin methoxyl groups of Picea abies as climate proxies. Rapid Communications in Mass Spectrometry 27(1): 265-275.

Horacek, M., M. Jakusch and H. Krehan (2009). Control of origin of larch wood: discrimination between European (Austrian) and Siberian origin by stable isotope analysis. Rapid Commun. Mass Spectrom. 23: 3688–3692

Horacek, M., G. Rees, M. Boner and J. Zahnen (2018). Comment on: Developing forensic tools for an African timber: […], by Vlam et al., 2018. Biological Conservation 226: 333-334.

Seal, R.R. II (2006). Sulfur Isotope Geochemistry of Sulfide Minerals. USGS Staff - Published Research. 345. https://digitalcommons.unl.edu/usgsstaffpub/345

Siegenthaler, U. (1979). Stable Hydrogen and Oxygen Isotopes in the Water Cycle. Lectures in Isotope Geology pp 264-273 Springer-Verlag Berlin. DOI: 10.1007/978-3-642-67161-6_22

Thode, H.G., J. Monster and H.B. Dunford (1961). Sulphur isotope geochemistry. Geochimica et Cosmochimica Acta 25:159-174. Pergamon Press Ltd. Printed in Northern Ireland.

7.4 APPENDIX 4: GUIDELINES FOR THE BUILDING OF IDEAL REFERENCE DATA COLLECTIONS

PRIOR NOTE: **Don't let the perfect be the enemy of the good**. The requirements below are meant to provide the ideal conditions for a reference database. However, suboptimal reference data are still better than no reference data at all.

- ☑ Use only **samples of known identity** (species, variety, geographic origin).
  - ➢ Samples collected in the field with GPS coordinates of the exact location and identified from the standing tree by a botanist.
  - ➢ Samples from a curated wood collection linked to herbarium vouchers
  - ➢ Samples identified by an experienced wood anatomist.
- ☑ Collect **enough samples[36] and material per sample**
  - ➢ to cover internal variation
  - ➢ as a back-up (for lost or destroyed material)
  - ➢ for use with additional methods (see Table 1 in the GTTN sampling guide for advice on sample quantity).
- ☑ Use only **samples without cracks or any signs of decay** (by *e.g.* bacteria, fungi), and avoid the use of chemicals or pencil on the wood samples (see the GTTN sampling guide for advice on storing samples).
- ☑ Include samples representing the **variability of identification requests** (different regions, commercial timbers, different sizes, sap- and heart-wood, …).
- ☑ Include samples representing the **environmental variability within the taxon's distribution range**.
- ☑ In case of **reference images**, use the same sample for the images and the reference descriptions.

[36] The IAWA Hardwood List (IAWA 1989) recommends using at least five vouchered wood samples per species description.

7.5 APPENDIX 5: METHOD INDEPENDENT ADVICE FOR DATA STORAGE & MANAGEMENT

Labelling

- **Avoid long labels** for files or excel sheets.
- **Set up a naming convention** for samples to allow classification and automatization (see below). Keep the same order of terms and do not add supplemental information.
- **Be consistent** in word use (language, names for countries, groups).
- **Use unique labels** for individual and population samples.
- **Keep a record of name changes** of samples, to be able to go back to the original naming. After analysis, some samples might need to get a different name if they were misidentified in the field. Later on, also this naming might turn out to be wrong, or you might find a physical or electronic location where the old name is still used and you want to check.
- **Flag undocumented samples**.

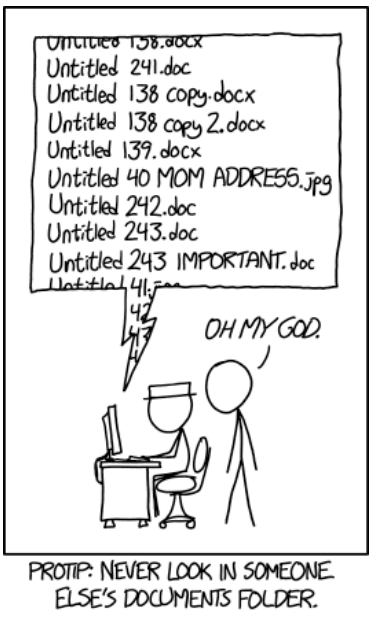


Image source: http://xkcd.com/1459/

Spreadsheet management

- **Keep a backup** of the original data.
- **Add the initial sample number** (given during fieldwork) to the data file.
- **1 dataset, 1 spreadsheet.** Avoid copying different datasets in the same Excel sheet. Mistakes can be made during sorting of data and lead to confusion.
- **Separate formulas and results.** Copy-paste results from formulas somewhere else as text for further analysis, keeping the formula.
- **When sorting, sort all.** Sort data by selecting all columns or all cells instead of by selection of a subset of cells.
- **Be careful with number and date formats.** Managing data from different sources sometimes involves careful transformations (*e.g.* point or comma as decimal separator, different date notations).
- **Easy access from big data analysis software via a relational database using SQL**. To reduce the chance for errors, it is good practice to store large datasets in a relational database that is compatible with most data analysis software (*e.g. R*, *Python*, *PHP*). For this purpose of automatic processing of data files, the abovementioned naming convention for samples is important. The database can include security features to restrict the access to the data, and you can enforce a specific data type input. All data including metadata (all information available for the sample) as well as all data analysis results can be linked by a number or code and stored in the same database.

TIP: For the design of the database there exists user-friendly software (*e.g. MySQL Workbench*, *Sequel Pro*, *phpMyAdmin*, *HeidiSQL*). In addition, there are several database management systems (*e.g. MySQL*, *MariaDB*, *SQL Server*, *Oracle*, *PostgreSQL*) as freeware and commercial versions. Each one of these resources has a large user community that provide comprehensive tutorials to help set everything up.

BOX 11: DATA MANAGEMENT ADVICE FOR *R* USERS

Analyse your data with *R scripts*, where all modifications of the data are recorded. Should you decide to change your mind about your analyses, you can just modify the step on your script, and rerun the analysis. *R scripts* can be shared with others for collaboration, are reproducible, and can be converted into PDFs with *R markdown* for publication as supplementary material. Make sure to annotate your script so that others can follow your line of thought and understand your choices throughout the analysis. Installation of git is recommendable for tracking versions of the script and collaborative work. Scripts can also be uploaded to Github and be shared online.

7.6 APPENDIX 6: METHOD INDEPENDENT ADVICE FOR DATA INTERPRETATION & REPORTING

7.6.1 ADVICE FOR CORRECT INTERPRETATION OF OBSERVATIONS

The precision with which a wood sample can be identified depends on **how much supporting information** is linked to it, how accurate that information is, and the **identity of the sample** (some taxa have a much narrower geographical range than others so assumptions can sometimes be made regarding identity) (Gasson 2011). To arrive at sound ecological conclusions, field experience is more important than statistical significance (Paredes-Villanueva 2018). In addition, when interpreting data, you should always be aware of **potential biases** and interpret accordingly, be critical. Wood is a biological material and hence varies. As Paredes-Villanueva (2018) formulated it: "Finding the variables that explain patterns in your data is like walking through a tropical forest. You always get distracted by random things".

The following variables should be considered during data interpretation:

- ☑ **Taxonomic knowledge**
  - Reference sample with/without voucher
  - Nomenclature/taxonomic changes
- ☑ **Comprehensiveness of the reference database**
  - Sample size (covering species variation)
  - Sample distribution (covering environmental variation)
  - Material types (sapwood, heartwood)
- ☑ **Research environment**
  - Possibility of contamination
  - Equipment used and its calibration
  - Software used and its suitability for the question at hand
  - Statistics/Models used and associated assumptions
- ☑ **Research results**
  - Stability of the identification result
  - Precision of the identification result
  - Statistical correctness of the data processing/mining

In general, always attempt to evaluate the evidence with respect to **alternative propositions** (UNODC 2016). This helps ensure applying an unbiased approach to result interpretation. Ask both questions: "Is the wood from species A?" and "Is this wood from another species?"

7.6.2 QUALITY DATA REPORTING

The task of the wood identification expert is to report on (i) the findings of the analysis, and (ii) all steps undertaken that led to these results. A wood identification/geographic origin report should contain the following information, if applicable to the timber tracking method used:

- ☑ Name of the person who performed the analysis
- ☑ Number of samples used per species/provenance
- ☑ List of where the samples come from (wood collection and location in the field)
- ☑ Number of spectra/markers used per species/provenance
- ☑ A list of the features/markers/elements used for the identification
- ☑ Description of equipment used (name, manufacturer, …)
- ☑ Images at the base of the identification result
- ☑ Number of technical replicates
- ☑ Binning size, signal to noise ratio, and threshold used
- ☑ Description of all statistical analyses
- ☑ The computer-calculated classification
- ☑ Validation results (accuracy and stability)
- ☑ Efficiency rate
- ☑ Rate of positive and negative errors

TIP: Ideally the **wood samples** of the case study should be stored or at least the sub-samples used for the analyses in case there is any future query or comeback.

REFERENCES:

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.

Paredes-Villanueva, K. (2018). Tropical timber forensics: a multi-methods approach to tracing Bolivian *Cedrela* (Doctoral dissertation, Wageningen: Wageningen University).

UNODC (2016). Best Practice Guide for Forensic Timber Identification. New York, United Nations.

Coordinating partners

With support from

by decision of the
German Bundestag

www.globaltimbertrackingnetwork.org

The objective of the Global Timber Tracking Network (GTTN) is to promote the operationalization of innovative tools for wood identification and origin determination, to assist the fight against illegal logging and related trade around the globe. GTTN is an open alliance that cooperates along a joint vision and the network activities are financed through an open multi-donor approach. GTTN phase 2 coordination (2017-2019) is financed by the German Federal Ministry of Food and Agriculture (BMEL).