Diameter, height and species of 42 million trees in three European landscapes generated from field data and airborne laser scanning data

Raphaël Aussenac, Jean-Matthieu Monnet, Matija Klopčič, Pawel Hawrylo, Jaroslaw Socha, Mats Mahnken, Martin Gutsch, Thomas Cordonnier, Patrick Vallet

DATA NOTE

REVISED Diameter, height and species of 42 million trees in three European landscapes generated from field data and airborne laser scanning data [version 2; peer review: 2 approved]

Raphaël Aussenac[1-3], Jean-Matthieu Monnet[1], Matija Klopčič[4], Paweł Hawryło[5], Jarosław Socha[5], Mats Mahnken[6], Martin Gutsch[6], Thomas Cordonnier[1,7], Patrick Vallet[1]

[1]Université Grenoble Alpes, INRAE, LESSEM, 2 rue de la Papeterie-BP 76, F-38402 St-Martin-d'Hères, France
[2]Forêts et Sociétés, Université de Montpellier, CIRAD, Montpellier, France
[3]CIRAD, UPR Forêts et Sociétés, Yamoussoukro, Cote d'Ivoire
[4]University of Ljubljana, Biotechnical Faculty, Department of Forestry and Renewable Forest Resources, Jamnikarjeva 101, 1000 Ljubljana, Slovenia
[5]Department of Forest Resources Management, Faculty of Forestry, University of Agriculture in Krakow, Al. 29 Listopada 46, 31-425 Krakow, Poland
[6]Potsdam Institute for Climate Impact Research (PIK), Member of the Leibniz Association, Telegrafenberg, 14473 Potsdam, Germany
[7]Office National des Forêts, Département Recherche Développement Innovation, Direction Territoriale Bourgogne-Franche-Comté, 21 rue du Muguet, 39100 Dole, France

v2 **First published:** 14 Feb 2023, **3**:32
https://doi.org/10.12688/openreseurope.15373.1
Latest published: 05 Dec 2023, **3**:32
https://doi.org/10.12688/openreseurope.15373.2

Abstract

Ecology and forestry sciences are using an increasing amount of data to address a wide variety of technical and research questions at the local, continental and global scales. However, one type of data remains rare: fine-grain descriptions of large landscapes. Yet, this type of data could help address the scaling issues in ecology and could prove useful for testing forest management strategies and accurately predicting the dynamics of ecosystem services.

Here we present three datasets describing three large European landscapes in France, Poland and Slovenia down to the tree level. Tree diameter, height and species data were generated combining field data, vegetation maps and airborne laser scanning (ALS) data following an area-based approach. Together, these landscapes cover more than 100 000 ha and consist of more than 42 million trees of 51 different species.

Alongside the data, we provide here a simple method to produce high-resolution descriptions of large landscapes using increasingly

Open Peer Review

Approval Status ✓ ✓

	1	2
version 2 (revision) 05 Dec 2023	✓ view	✓ view
version 1 14 Feb 2023	? view	? view

1. **Fabian Fischer**, University of Bristol, Bristol, UK

2. **Nikolai Knapp**, Thünen Institute of Forest Ecosystems, Eberswalde, Germany

Any reports and responses or comments on the article can be found at the end of the article.

available data: inventory and ALS data.

We carried out an in-depth evaluation of our workflow including, among other analyses, a leave-one-out cross validation. Overall, the landscapes we generated are in good agreement with the landscapes they aim to reproduce. In the most favourable conditions, the root mean square error (RMSE) of stand basal area (BA) and mean quadratic diameter (Dg) predictions were respectively 5.4 m2.ha-1 and 3.9 cm, and the generated main species corresponded to the observed main species in 76.2% of cases.

Keywords
forest, inventory, landscape, tree-level, airborne laser scanning, downscaling

This article is included in the Horizon 2020 gateway.

This article is included in the Forest and Forestry Sciences gateway.

Corresponding author: Raphaël Aussenac (raphael.aussenac@proton.me)

Author roles: Aussenac R: Conceptualization, Data Curation, Formal Analysis, Methodology, Software, Supervision, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Monnet JM**: Conceptualization, Data Curation, Formal Analysis, Methodology, Resources, Software, Validation, Writing – Original Draft Preparation, Writing – Review & Editing; **Klopčič M**: Data Curation, Investigation, Resources, Writing – Original Draft Preparation, Writing – Review & Editing; **Hawryło P**: Data Curation, Investigation, Resources; **Socha J**: Conceptualization, Funding Acquisition, Project Administration, Resources, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing; **Mahnken M**: Conceptualization, Methodology, Writing – Original Draft Preparation, Writing – Review & Editing; **Gutsch M**: Conceptualization, Methodology; **Cordonnier T**: Conceptualization, Funding Acquisition, Methodology, Project Administration, Supervision, Validation, Writing – Original Draft Preparation, Writing – Review & Editing; **Vallet P**: Conceptualization, Formal Analysis, Funding Acquisition, Methodology, Project Administration, Supervision, Validation, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This research was financially supported by the European Union's Horizon 2020 research and innovation programme under the grant agreement No 773324 (ForestValue - Innovating forest-based bioeconomy [ForestValue]). This work was carried out within the framework of the I-Maestro project, supported under the umbrella of ERA-NET Cofund ForestValue by ADEME (FR), FNR (DE), MIZS (SI), NCN (PL). This work was also supported by the GRAINE program of ADEME (FR) in the framework of the PROTEST project (convention n°1703C0069).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Aussenac R, Monnet JM, Klopčič M *et al.* **Diameter, height and species of 42 million trees in three European landscapes generated from field data and airborne laser scanning data [version 2; peer review: 2 approved]** Open Research Europe 2023, **3**:32 https://doi.org/10.12688/openreseurope.15373.2

First published: 14 Feb 2023, **3**:32 https://doi.org/10.12688/openreseurope.15373.1

REVISED **Amendments from Version 1**

In this new version we provide an in-depth evaluation of the generated landscapes. The Dataset validation section has been completely revised. This new evaluation is presented synthetically in the General approach section and some results are mentioned in the abstract. In the Algorithm section, we have gone into more detail to clarify the functioning of our downscaling algorithm. In the introduction, we further explain the interest of our fine-grain large-scale datasets. In the ALS metrics section, additional information were provided on how the sensitivity of point cloud metrics to scanner acquisitions was handled.

Any further responses from the reviewers can be found at the end of the article

Introduction

In recent years, a considerable effort has been made to make forest inventory data available, and to aggregate them at the continent [Mauri *et al.*, 2017] or at the global scale [Cazzolla Gatti *et al.*, 2022; Liang *et al.*, 2016]. These data make it possible to study ecological processes at fine scales (at the inventory plot scale) as well as at coarse scales (by aggregating inventory plots). At the forest or landscape scale however, they are of limited use as they hardly capture forest- or landscape-level ecological processes. Denser networks of inventory plots or large-scale inventories are needed. However, beyond a certain area, large-scale inventories become too costly and plot networks are preferred. Yet, fine-grain descriptions of large forest areas could help address the pervasive scaling issues in forest ecology, modelling and management. In practice, such data could help better understand at which spatial scale ecological processes emerge in forest ecosystems [Craven *et al.*, 2020; With, 2019]. They could also be extremely valuable to compare forest dynamics models operating at different scales (organ, tree, stand, landscapes) and evaluate their validity across scales [Papaik *et al.*, 2010]. They could ultimately help develop and test management strategies at different spatial scales [Seidl *et al.*, 2013].

Airborne Laser Scanning (ALS) surveys are a promising way forward to address this challenge, as they can provide high-resolution data over wide areas. However, retrieving individual tree attributes from ALS point clouds remains a challenge in particular in closed-canopy forests. At present, one solution is to combine ALS data with tree-level field data [Lamb *et al.*, 2018; Silva *et al.*, 2016].

Here we present three datasets describing three large European landscapes in France (Bauges Geopark ≈ 89,000 ha), Poland (Milicz forest district ≈ 21,000 ha) and Slovenia (Snežnik forest ≈ 4700 ha) down to the tree level. Individual trees were generated combining inventory plot data, vegetation maps and ALS data. Together, these landscapes (hereafter virtual landscapes) cover more than 100,000 ha including about 64,000 ha of forest and consist of more than 42 million trees of 51 different species.

In addition to the datasets, we provide here a simple method to predict the diameter, height and species of all trees in a landscape using increasingly available data: inventory and ALS data. This method also has the advantage of being fast: about 1 hour on an height-core laptop is needed to generate the 42 million trees making up the 64,000 ha of forest of our three landscapes.

Methods

Three study areas

Three European study areas were used as bases for our virtual landscapes: the Bauges Geopark, the Milicz forest district and the Snežnik forest (Figure 1).

The Bauges Geopark is a mountainous area located in the French Alps between 255 and 2672 m above sea level (a.s.l.). It is a karst mountain range characterised by a steep and irregular topography. The annual rainfall is about 1100 mm, and the average annual temperature is 8°C at Bellecombe-en-Bauges (850 m a.s.l.). Monthly temperatures range from -1.3 to 17.1°C. The Bauges Geopark covers a total area of 89,324 ha including 51,564 ha of forest (21,073 ha of public forest and 30,491 ha of private forest). The main tree species are beech (*Fagus sylvatica*), fir (*Abies alba*) and spruce (*Picea abies*) which are mostly found in uneven-aged mixed stands, but the area is characterised by a great diversity of tree species. In particular, mixed stands of broadleaf species are found at low elevation.

The Milicz forest district is located in the province of Lower Silesia in south west Poland at a mean elevation of 126 m a.s.l. (elevation ranging from 96 to 227 m a.s.l.). Much of the area is almost flat or slightly undulating with gentle slopes. This part of the landscape is covered by developed terraces and aeolian formations. The remaining part of the landscape is a slightly undulating moraine plateau above which irregularly shaped moraine hills are found. The average annual rainfall is 565 mm and the mean annual temperature is 8.2°C. Monthly temperatures range from -1.3 to 17.8°C. The Milicz forest district covers a total area of 21,086 ha including 7713 ha of public forest. Small patches of private forest are also found in the landscape but they were not considered here as no field data were collected there. The public forest is largely dominated by pure stands of Scots pine (*Pinus sylvestris*). Pure and mixed stands of oak (*Quercus robur*) and beech are also found, but in a much smaller proportion.

The Snežnik forest is located in the Dinaric Mountains in southern Slovenia between 572 and 1792 m a.s.l. The Dinaric Mountains are a karst mountain range composed mainly of limestone and dolomite and characterised by an irregular and diverse topography and rockiness. The area has abundant precipitation (over 2000 mm annually on average), which is evenly distributed throughout the year. The average annual temperature is 6.5°C, with a mean monthly maximum temperature of around 16°C in July and a mean minimum of -3.4°C in January.

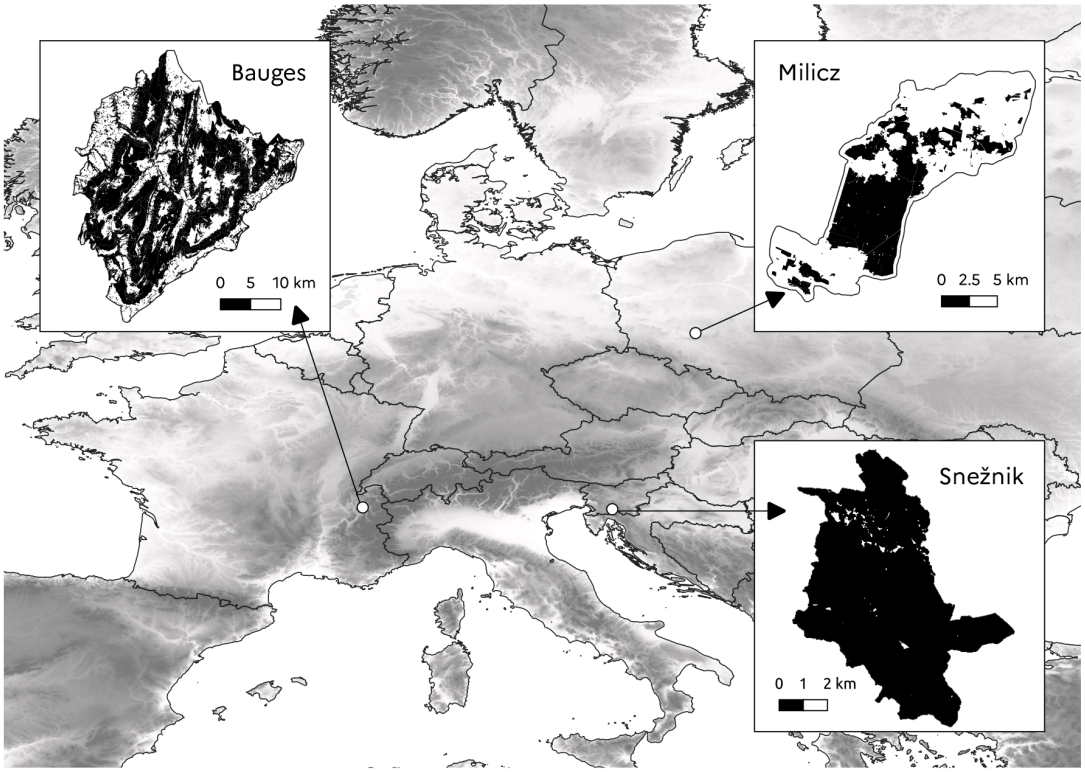

Figure 1. Location of study areas. The black areas show the forested areas.

The study area spans over 4725 ha and is almost completely covered by public forest (4660 ha). The main tree species are fir and beech, which are mostly found in uneven-aged mixed stands. Interestingly, in this study area, the upper forest limit is formed by beech stands and not conifer stands.

General approach

Here we outline the approach we adopted to produce the virtual landscapes corresponding to our three study areas (Figure 2).

First, we produced raster maps of stand total basal area (BA), mean quadratic diameter (Dg) and proportion of broadleaf trees BA (BA_b) at a 25 m resolution (see *ALS mapping*). For that, we used ALS point clouds along with field data (tree diameter and species identity). Thereafter, we generated trees in each 25x25 m^2 cell, specifying their diameter at breast height (dbh), number (n) and species (sp; see *Downscaling algorithm*). For that, we first assigned to each cell a stand from the field data based on the similarity of their BA, Dg and BA_b values (calculated as the Euclidean distance between each cell and each field plot in the three-dimensional space made up by the scaled values of BA, Dg and BA_b). We then transformed the structure of the stand chosen from the field data (by changing the trees dbh, basal area and weight) to reach the BA and BA_b values of the cell. Finally, we used diameter-height models to assign heights (h) to all trees (see *Heights models*).

We evaluated the overall reliability of our workflow, i.e. its ability to produce virtual landscapes as close as possible to the real ones (see *Dataset validation*). In particular we carried out a leave-one-out cross validation (LOOCV) on our entire workflow. This analysis consisted in:

- comparing the observed and predicted values of BA, Dg, Ba_b and the quantiles of tree height and diameter;
- comparing the observed and predicted values of species abundance at the landscape level;
- calculating the frequency at which the most abundant species was correctly predicted at the cell level;

As a complement, we also compared the stands dominant heights measured by ALS ($Hdom_{ALS}$) to those calculated from the trees we generated ($Hdom_T$). Finally, we compared the spatial distribution of species to current expert knowledge.

ALS mapping

The so-called "area-based" approach is a workflow commonly implemented for mapping stands variables in operational

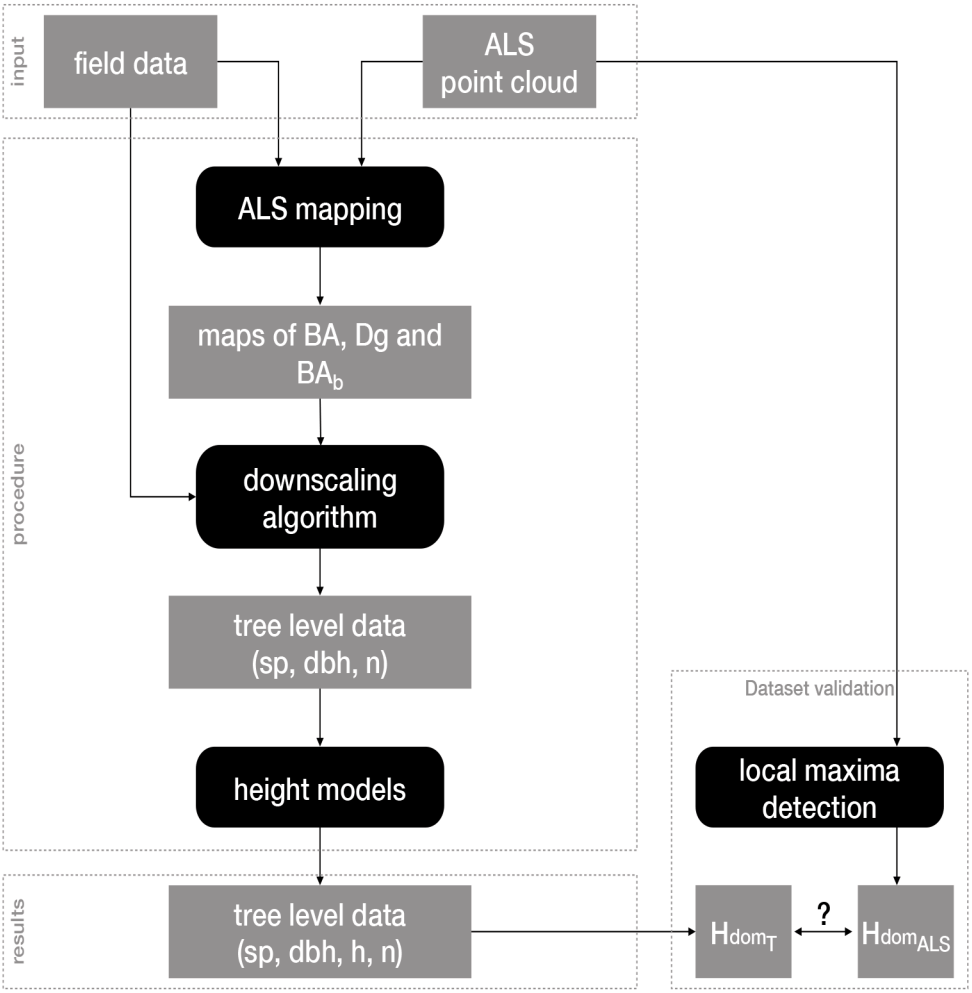

Figure 2. Workflow overview. Black boxes correspond to data generation steps feeding each other with datasets represented by grey boxes. BA: basal area; Dg: mean quadratic diameter; BA_b: BA proportion of broadleaf trees; sp: species; dbh: diameter at breast height; h: height; n: number of trees; $Hdom_{ALS}$ and $Hdom_T$: stands dominant heights measured by ALS or calculated from the generated trees, respectively.

conditions [White *et al.*, 2013]. It is based on the synergistic use of field plots and ALS point clouds. Estimation models for target forest variables are fitted with point clouds statistics, also called metrics, as predictor variables. Field plots are used for training the models. For the mapping step the predictor variables are computed in each cell of a raster layout over the whole acquisition area, and then the models are applied to obtain wall-to wall-estimates. This workflow was implemented in each study area.

Forest areas. Reference areas for forest mapping were defined as the intersection of two layers for each site, one defining the administrative boundary and one defining the forest mask. Those extents are respectively:

- Bauges: the Geopark administrative extent with the forest mask defined by the BD Forêt v2 from the National Institute of Geographic and Forest Information [IGN, 2019], excluding the "herbaceous", "moors" and "*Populus* plantations" categories;
- Milicz: the public forests of Milicz with the forest mask defined by the Forest Data Bank [Bureau for Forest Management and Geodesy, 2020];
- Snežnik: the forest management units of Leskova Dolina and Snežnik with the forest mask defined by Snežnik-forest cover [Service, 2020].

Field data. In the Bauges, a local forest inventory with 320 plots was implemented in 2018. On each plot, all living trees with a dbh larger than 17.5 cm and within a 15 m horizontal distance from the plot centre had their dbh, position and species recorded. Trees with a dbh between 7.5 and 17.5 cm

were counted according to simplified categories of diameter and species (coniferous / broadleaf). Plot centres were geolocated with survey-grade GNSS (Global Navigation Satellite System) receivers. Plots co-registration with the ALS data was improved when possible by comparing the positions of trees with the Canopy Height Model (CHM) derived from the point cloud.

At Milicz, a local forest inventory with 901 plots of 12.62 m radius was carried out in 2015. Species and diameter of all living trees with dbh above 7 cm were recorded. Plot centres were geolocated with survey-grade GNSS receivers.

At Snežnik, a total of 515 plots were inventoried, in 2013 for plots located in the Leskova Dolina management unit and in 2014 for plots located in the Snežnik management unit. Trees with a dbh above 30 cm within a 12.61 m distance from the plot centre had their diameter and species recorded. Trees with a dbh between 10 and 30 cm were recorded within a 7.98 m distance from the plot centre. Plot centres were geolocated with commercial-grade GNSS receivers.

The following stand-level variables were computed for each plot: total basal area (BA) in $m^2.ha^{-1}$, mean quadratic diameter (Dg) in cm and the proportion of broadleaf species in basal area (BA_b). Weights were applied to correct for sampling intensity in the case of nested plots (Bauges and Snežnik).

ALS data. The Bauges was covered by two ALS acquisitions with different settings and equipment. The southern part was covered between June and September 2016, the northern part in September 2018. Point densities computed at 25 m resolution in forest areas were respectively 5.9 ± 3.1 and 27.6 ± 13.3 m^{-2}. Intensity values were normalised by dataset, by subtracting the mean and dividing by the standard deviation of intensity values of points located inside the extent of field plots covered by each acquisition.

Milicz was covered by an ALS acquisition in August 2015. The point density was 16.5 ± 7.1 m^{-2}. The point cloud contains colour values extracted from aerial pictures with near infra-red, red and green bands.

Snežnik was covered by an ALS acquisition between February 14th and November 21st 2014. Forests might have been both in leaf-on and leaf-off conditions. The point density was 18.4 ± 10.1 m^{-2}. An ice storm occurred in Leskova Dolina management unit between January 30th and February 10th 2014. This event damaged the forest stands, and happened between the field inventory and the ALS acquisition. It affected the quality of the derived maps (see *Mapping*) and the realism of our virtual landscape (see *Dataset validation*).

ALS metrics. All computations were performed with R software. Terrain metrics (aspect, elevation and slope) were computed by fitting a plane surface to points classified as ground. Before the computation of vegetation metrics, ALS point clouds were normalised, *i.e.* height above ground was computed for each point. Two types of metrics were then computed from the points classified as vegetation with a height above 2 meters (this limit was set to remove points of shrubs and low vegetation from the analysis):

- Point cloud metrics were directly computed from the point cloud using the aba_metrics function from the lidaRtRee R package. Those metrics summarise the geometry of the point cloud in a given area.

- Tree metrics were computed with the std_tree_metrics function from the characteristics of local maxima extracted from the CHM with the tree_segmentation function. CHM resolution was set to 0.5 m at Milicz, and 1 m at Snežnik and the Bauges due to higher variability of point density. Local maxima with a height lower than 5 m were discarded. Those metrics summarise the characteristics of trees detected in a given area of the point cloud. One of the tree metrics is the ALS dominant height ($Hdom_{ALS}$), which is the mean height of the six highest local maxima. In case less than six maxima were present, the mean height of all maxima was used.

The metrics were computed for each field plot based on the point cloud located inside their extent, in order to build the dataset for model calibration (training step). The metrics were also computed in each 25×25 m^2 cell of the raster layout covering each acquisition, in order to build the prediction dataset (mapping step). Each metric map was visually checked for spatial patterns potentially linked to acquisition patterns, which eventually led to:

- discard some intensity-related metrics in Sneznik study area;

- remove ALS points acquired with a scan angle larger than 21 degrees in Milicz study area, in order to achieve a trade-off between metrics robustness, point density and comprehensive coverage of the study area.

Models. For BA and Dg, we searched for the linear regression model that yielded the highest adjusted-R^2 with at most $n = 6$ independent variables among the above-mentioned ALS metrics. The model was given by:

$$\hat{y} = a_0 + \sum_{i=1}^{n} a_i x_i \quad (1)$$

with $\hat{y}$ the estimated value, $(a_i)i\in\{0,...,n\}$ the model parameters and $(x_i)i\in\{1,...,n\}$ the selected metrics. Two data transformations were also tested: a logarithm transformation of all variables and a Box-Cox transformation of the dependent variable. The logarithm transformation of all variables turns the model at Equation 1 into:

$$\hat{y} = e^{(a_0)} \times \prod_{i=1}^{n} x_i^{a_i} \quad (2)$$

A bias correction factor had to be applied to the fitted values to obtain the predictions (P):

$$P = \hat{y} \times e^{\left(\frac{v}{2}\right)} \quad (3)$$

with v the variance of the model residuals.

The Box-Cox transformation consists in determining the λ parameter that best normalises the distribution of the dependent variable (Y). It is determined using the maximum likelihood-like approach of Box & Cox [1964] (*powerTransform* function of car R package). Y is given by:

$$Y = \frac{(y^{\lambda} - 1)}{\lambda} \quad (4)$$

Equation 1 is then fitted with Y instead of y. The predictions P are obtained by applying the inverse Box-Cox transformation to the fitted values $\hat{Y}$ and a bias correction factor:

$$P = (\lambda \hat{Y} + 1)^{\frac{1}{\lambda}} \times \left(1 + \frac{v}{2} \times \frac{1-\lambda}{(\lambda \hat{Y} + 1)^2}\right) \quad (5)$$

For broadleaf proportion (BA_b), values are bounded to [0, 1]. A binomial generalised linear model with logit link was therefore fitted with the *glm R* function. The model was given by:

$$log\left(\frac{\widehat{BA_b}}{1 - \widehat{BA_b}}\right) = a_0 + \sum a_i x_i \quad (6)$$

All metrics were at first included in the model and then a stepwise selection was used to reduce their number (*stepAIC* function of the *MASS R* package).

Stratification. When calibrating a statistical relationship between forest stand variables, which are usually derived from diameter measurements and ALS metrics, one relies on the hypothesis that the interaction of laser pulses with the leaves and branches structure is constant on the whole area. However, differences can be expected either due to variations in acquisition settings (flight parameters, scanner model), in forests (stand structure and composition) or in topography (slope). Better models might be obtained when calibrating stratum-specific relationships, provided each stratum is more homogeneous regarding the laser interaction with the vegetation. A trade-off has to be achieved between the within-strata homogeneity and the number of available plots for calibration in each stratum.

Depending on the study areas, different ancillary data are available for stratification. At the Bauges, two layers were used: species composition (mixed, broadleaf, coniferous) derived from the BD Forêt v2 and ALS survey. At Milicz, the following information was available for a total of 2175 stands: dominant species (coniferous, *Quercus*, other broadleaf) and stand age. At Snežnik, the following information was available for a total of 1536 stands: forest management unit (FMU: Snežnik or Leskova Dolina) and broadleaf proportion in volume, which is converted into a two (broadleaf or coniferous) or three-levels factor (adding the mixed category). The metrics selected in the 32 models for BA and Dg (which include at most six independent variables) are presented in Table S1 of the *Extended data*.

Field plots and raster cells were assigned to the category of the polygon which contains their centres.

Mapping. Stratifications were compared based on expert knowledge taking into account the following criteria: minimum number of observations in strata, prediction error and number of variables in the model. The retained stratifications for the prediction models and the root mean square error (RMSE) of prediction estimated in leave-one-out cross validation are presented in Table 1.

Table 1. Stratification and root mean square error (RMSE) of predictions for the three study areas and three forest variables. BA: basal area ($m^2.ha^{-1}$); Dg: mean quadratic diameter (cm); BA_b: broadleaf BA proportion (%).

study area	Variable	RMSE	Stratification: number and combinations
Bauges	BA	8.3	6: composition x ALS survey
	Dg	4.2	6: composition x ALS survey
	BA_b	20.3	3: composition
Milicz	BA	5.4	7: (coniferous x 5 age classes), *Quercus sp.*, other broadleaf
	Dg	3.7	3: coniferous, *Quercus sp.*, other broadleaf
	BA_b	12.9	2: coniferous, broadleaf
Snežnik	BA	9.6	4: FMU x composition (2 classes)
	Dg	7.6	6: FMU x composition (3 classes)
	BA_b	19.3	2: FMU

Prediction accuracy is better for mean diameter and lower for BA, which is common when estimated with ALS. Precision is quite low for broadleaf proportion, which could be expected as spectral data are usually better than ALS at classifying species. Prediction accuracy was higher at Milicz, intermediate at the Bauges and lower at Snežnik. Milicz was well suited for making predictions with its dense ALS data, homogeneous stands and precise co-registration. The Bauges has precise co-registration, but heterogeneous forest stands and two different ALS datasets. At Snežnik the data were much noisier, especially because of the ice storm event. The maps we created are presented in Figure 3.

Downscaling algorithm

Field data. At Milicz and Snežnik, we used the same dbh measurements as those used to calibrate the ALS models (from 901 plots at Milicz and from 515 plots at Snežnik, see *ALS mapping - Field data*). At the Bauges, we could not use the dbh measurements used to calibrate the ALS models because trees with a dbh smaller than 17.5 cm were not measured but

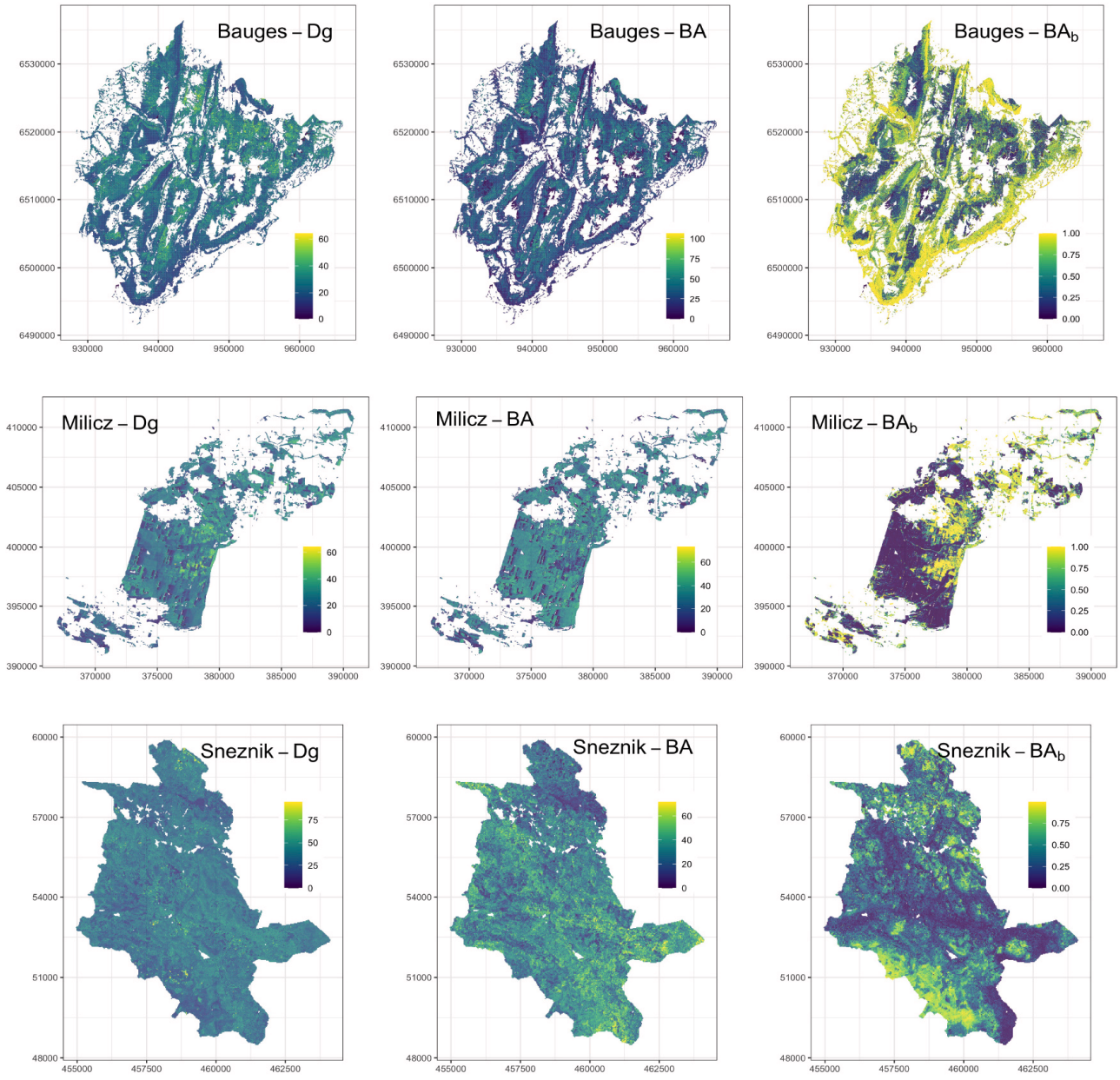

Figure 3. Airborne laser scanning (ALS) maps of forest variables for our three study areas at a 25 m resolution. Dg: mean quadratic diameter (cm), BA: basal area ($m^2.ha^{-1}$) and (BA_b): proportion of broadleaf BA.

counted by diameter classes. Instead, we used the tree diameter measurements from the 258 forest plots of the French National Forest Inventory (NFI) located in the study area. Those plots were inventoried between 2005 and 2018. They consist of three concentric plots of 6 m, 9 m and 15 m radius, where small (7.5 < dbh < 22.5 cm), medium (dbh < 37.5 cm) and big trees (dbh > 37.5 cm) were measured, respectively. At the Bauges, we used an additional information on forest vegetation: the map of forest types [IGN, 2019], which we also used to delineate the forest areas (see *Forest areas*).

Algorithm. Our algorithm consisted in associating to each 25×25 m² cell a field plot based on the similarity of their dendrometrical variables, and then in modifying the trees dbh, basal area and weight of this field plot in order to reach the total BA and the proportion of broadleaf BA (BAb) of the cell (*i.e.* the values provided by the ALS maps). The algorithm breaks down as follows:

1. First, we calculated the total basal area (BA), mean quadratic diameter (Dg) and proportion of broadleaf BA (BA_b) of all field plots.

2. Second, we associated to each 25×25 m² cell a field plot based on the similarity of their BA, Dg and BA_b. These 3 variables were chosen for matching because together they provide a synthetic yet fairly accurate picture of the stands.

   (a) For this, we scaled the values of BA, Dg and BA_b between 0 and 1. We scaled the ALS and field data together to account for the possible differences in their range.

   (b) We then calculated the Euclidean distance between each cell and each field plot in the three-dimensional space made up by the scaled values of BA, Dg and BA_b.

   (c) Finally, we associated to each cell the closest field plot in this three-dimensional space. For the Bauges study area, we assigned to each 25×25 m² cell a forest type (*e.g.* pure beech, mixed deciduous forest, among others) from the map of forest types. We then associated the closest field plot sharing the same forest type to each cell.

3. Third, we transformed the field plots stand structure so that it matched the BA and BA_b values of the cells they were associated with.

   (a) For this, we first calculated α, a multiplier correction coefficient to be applied to all tree diameters of a field plot. The idea is to increase or decrease tree diameters so that their Dg reaches the Dg value of the cell to which they are associated. α is given by:

$$\alpha = \frac{Dg_{ALS}}{Dg_F} \quad (7)$$

   with Dg_{ALS} the Dg value of the cell given by the ALS mapping, and Dg_F the Dg value calculated with the dbh of the trees from the field plot.

   (b) Thereafter, we calculated the weight (ω in $n.ha^{-1}$) of these trees with corrected diameters, so that the generated stand matches the BA and BAb values of the cell it is associated with. ω is given by:

$$\omega = \frac{40000}{\pi} \times \frac{ba_{tree_{ALS,F}}}{(\alpha.dbh_F)^2} \quad (8)$$

   where dbh_F is the tree dbh in the field plot, and $ba_{tree_{ALS,F}}$ is the tree individual basal area derived from the ALS mapping and the field plot data using the following equation:

$$ba_{tree_{ALS,F}} = BA_{ALS} \times Prop_{BC_{ALS}} \times Prop_{Sp_F} \times Prop_{tree_F} \quad (9)$$

   where BA_{ALS} is the total BA of the cell given by the ALS mapping, $Prop_{BC_{ALS}}$ is the BA proportion of broadleaf (resp. coniferous) trees given by the ALS mapping, $Prop_{Sp_F}$ is the BA proportion of species Sp in broadleaf (reps. coniferous) species in the field plot, and $Prop_{tree_F}$ is the BA proportion of this tree in species Sp in the field plot.

   (c) Finally, we divided ω by 16 to get the weight of the trees in the 25×25 m² cells (ω being a weight per ha and 16 being the surface area ratio between 1 ha and a 25×25 m² cell). In doing so, the obtained tree weights can be either integer or decimal. However, the objective of our algorithm is to generate for each cell a list of individual trees with their associated diameter, height and species. From this perspective, decimal weights are not useful. We cannot simply round the tree weights to the nearest integer as this can lead to a significant over- or underestimation of the total number of trees in the cells. This is because the decimal part of the tree weights in the 25×25 m² cells is not the result of a random draw but directly depends on the surface area ratio between the field plot and the cell. As an example: 1 tree inventoried on a 400 m² field plot will always obtain a weight of 1.56 in a 25 × 25 m² cell, and a weight of 2 after rounding to the nearest integer. In order to obtain integer tree weights in the 25 × 25 m² cells while avoiding this bias, we performed a Bernoulli draw on the decimal part of the tree weights. As an example, a weight of 1.56 has a 56% chance of becoming 2, and a 44% chance of becoming 1. As this rounding of the weights slightly modifies the total BA of the generated stand, we transformed

[1] The scale factor 40000 is the product of two scale factors: 4 × 10000. The scale factor 4 comes from the formula linking a surface area *S* to a diameter d ($S = \pi \frac{d^2}{4}$); while the scale factor 10000 accounts for the difference in units between the diameters (in cm) and the basal areas (in m²).

again the trees dbh to reach the total BA provided by the ALS mapping using the trees BA and their integer weights (ω_{int}) as follows:

$$dbh_{final} = \sqrt{\frac{40000^{1}}{\pi} \times \frac{ba_{tree_{ALS,F}}}{\frac{16}{\omega_{int}}}} \quad (10)$$

As this last transformation only compensates for the rounding, the changes in dbh are minor.

This procedure has multiple benefits (see proofs in *Extended data*): it makes it possible to reach the BA and BA_b values given by the ALS mapping. It also maintains the Dg ratios observed on the field plots between the different species. The Bernouilli draw used to get integer tree weights only adds a minor variability. We created the three virtual landscapes by applying this algorithm to each study area separately.

Heights models

We developed individual diameter-height models for the three study areas to assign heights to all generated trees.

Field data. At Snežnik and Milicz, the diameter and height measurements come from the same field plots used for the ALS models calibration (see *ALS mapping - Field data*). At the Bauges, no height measurements were collected in the field plots used to calibrate the ALS models. We therefore used the tree diameter and height measurements of the 240 French NFI plots located in the study area (inventoried between 2005 and 2016). At Milicz and the Bauges, the heights were measured for all species in all diameter classes. At Snežnik, tree heights were measured only on two to four trees from the upper layer. The number of trees with both diameter and height measurements in each study area is summarised per species in Table 2.

Table 2. Number of trees for the diameter-height models calibration in each study area and for each species. For each study area, all the species with less than 100 observations are grouped into the "other species" category.

Species	Number of trees for		
	Bauges	Milicz	Snežnik
Abies alba	468		638
Acer pseudoplatanus	181	228	
Alnus glutinosa		823	
Betula pendula		1 519	
Carpinus betulus		808	
Fagus sylvatica	705	2 199	435
Fraxinus excelsior	209		
Larix decidua		709	
Picea abies	551	2 183	325
Pinus sylvestris		24 995	
Prunus serotina		191	
Quercus petraea	130		
Quercus rubra		308	
Quercus undefined*		1 916	
Tilia cordata		311	
Other species	642	522	29
TOTAL	2 886	36 712	1 427

*At Milicz, the *Quercus* undefined is mainly *Quercus robur*.

Models. We used a mixed effect model to predict individual tree height from the ratio between the tree dbh and the stand Dg (to account for the tree social status) and from the stand Dg (to account for the stand development stage). We considered the site effect as a random effect. Finally, as the variance of height increases with height due both to increasing measurement errors and to individual cumulative variations, we accounted for heteroscedasticity by modelling the error term with a power of the fitted values. The model is given by:

$$h_{tot} = 1.3 + (1 + \alpha_{site}) \times \alpha_{sp} \times \left(1 - e^{(-\alpha_1 \times Dg^{\alpha_2})}\right) \times \left(1 - e^{(-\beta_{sp} \times \frac{dbh}{Dg})}\right)^{\gamma} + \epsilon \quad (11)$$

where α_{sp}, α_1, α_2, β_{sp} and γ are parameters to be estimated; and α_{site}, a random effect accounting for the site effect. This model has an asymptotic form: α_{sp} corresponds to the species-specific asymptotic value, and β_{sp} is the species-specific speed for reaching the asymptotic value.

At Snežnik, most of the trees selected for height measurement were dominant or co-dominant trees. Moreover, more than half of the plots only had two observations. This precludes to fit the part of the curve with small diameters within the stand. We solved this issue by assuming that the within-stand relationship at the Bauges was similar at Snežnik, as these landscapes are quite similar in terms of species, stand structure (mostly uneven-aged), or elevation (mountains). Therefore, for Snežnik height predictions, we used the β_{sp} and γ fitted values of the Bauges model.

We fitted one mixed effect model for each study area using the *nlme* function from the *nlme R* package. We modelled the residual errors using a *varPower* function of the fitted values. The parameters are presented in Table 3, Table 4, and Table 5 for the three study areas.

Dataset validation

Method

We carried out a leave-one-out cross validation (LOOCV) to evaluate the realism of the virtual landscapes we generated. This consisted in excluding a field plot from our entire

Table 3. Parameters of the Bauges diameter-height model.

Parameter	Value	Standard error	p-value
$\alpha_{Fa.sy.}$	41.05595	4.3	<10^{-3}
$\alpha_{Pi.ab.}$	55.11821	5.8	<10^{-3}
$\alpha_{Ab.al.}$	48.46640	5.1	<10^{-3}
$\alpha_{Fr.ex.}$	40.94293	4.3	<10^{-3}
$\alpha_{Ac.ps.}$	37.95001	4.0	<10^{-3}
$\alpha_{Qu.pe.}$	36.64676	4.2	<10^{-3}
$\alpha_{OtherSp.}$	36.87834	3.8	<10^{-3}
α_{1}	0.01594	0.0030	<10^{-3}
α_{2}	1.26326	0.10	<10^{-3}
$\beta_{Fa.sy.}$	1.71474	0.08	<10^{-3}
$\beta_{Pi.ab.}$	0.99226	0.05	<10^{-3}
$\beta_{Ab.al.}$	1.17894	0.06	<10^{-3}
$\beta_{Fr.ex.}$	2.01951	0.12	<10^{-3}
$\beta_{Ac.ps.}$	2.08068	0.12	<10^{-3}
$\beta_{Qu.pe.}$	1.56216	0.16	<10^{-3}
$\beta_{OtherSp.}$	1.84067	0.08	<10^{-3}
γ	1.42595	0.05	<10^{-3}
Power of the variance model			0.51
Standard deviation of the plot level random effect			0.14
Standard deviation of residual error			0.59

Table 4. Parameters of the Milicz diameter-height model.

Parameter	Value	Standard error	p-value
$\alpha_{Pi.sy.}$	48.55802	2.3	<10^{-3}
$\alpha_{Fa.sy.}$	48.01692	2.3	<10^{-3}
$\alpha_{Pi.ab.}$	60.35196	3.1	<10^{-3}
$\alpha_{Qu.un.}$	52.24210	2.5	<10^{-3}
$\alpha_{Be.pe.}$	51.60844	2.5	<10^{-3}
$\alpha_{Al.gl.}$	49.34039	2.4	<10^{-3}
$\alpha_{Ca.be.}$	36.73985	1.8	<10^{-3}
$\alpha_{La.de.}$	52.06992	2.5	<10^{-3}
$\alpha_{Ti.co.}$	45.25535	2.4	<10^{-3}
$\alpha_{Qu.ru.}$	45.74754	2.4	<10^{-3}
$\alpha_{Ac.ps.}$	41.50894	2.2	<10^{-3}
$\alpha_{Pr.se.}$	36.18532	2.9	<10^{-3}
$\alpha_{OtherSp.}$	54.94652	2.8	<10^{-3}
α_{1}	0.01958	0.001	<10^{-3}
α_{2}	1.13831	0.035	<10^{-3}
$\beta_{Pi.sy.}$	2.73192	0.024	<10^{-3}
$\beta_{Fa.sy.}$	1.98085	0.032	<10^{-3}
$\beta_{Pi.ab.}$	1.20700	0.035	<10^{-3}
$\beta_{Qu.un.}$	1.62943	0.027	<10^{-3}
$\beta_{Be.pe.}$	2.11097	0.037	<10^{-3}
$\beta_{Al.gl.}$	2.04760	0.045	<10^{-3}
$\beta_{Ca.be.}$	2.86677	0.063	<10^{-3}
$\beta_{La.de.}$	2.33369	0.050	<10^{-3}
$\beta_{Ti.co.}$	1.89682	0.064	<10^{-3}
$\beta_{Qu.ru.}$	2.38748	0.095	<10^{-3}
$\beta_{Ac.ps.}$	2.56340	0.102	<10^{-3}
$\beta_{Pr.se.}$	2.04373	0.150	<10^{-3}
$\beta_{OtherSp.}$	1.50792	0.019	<10^{-3}
γ	1.55264	0.040	<10^{-3}
Power of the variance model			0.16
Standard deviation of the plot level random effect			0.09
Standard deviation of residual error			1.09

Table 5. Parameters of the Snežnik diameter-height model.

Parameter	Value	Standard error	p-value
$\alpha_{Ab.al.}$	66.17413	5.4	<10^{-3}
$\alpha_{Fa.sy.}$	53.81402	4.4	<10^{-3}
$\alpha_{Pi.ab.}$	76.82544	6.3	<10^{-3}
α_{1}	0.0251	0.0036	<10^{-3}
α_{2}	1.00672	0.075	<10^{-3}
$\beta_{Ab.al.}$*	1.17894	* taken from the Bauges model	
$\beta_{Fa.sy.}$*	1.71474		
$\beta_{Pi.ab.}$*	0.99226		
γ*	1.42595		
Power of the variance model			-0.56
Standard deviation of the plot level random effect			0.077
Standard deviation of residual error			15.8

workflow and comparing the predicted values obtained to the observed values. This operation was repeated within each landscape for all field plots. We calculated the root mean square error (RMSE) of the predictions of BA, Dg, Ba_b and the quantiles of tree height and diameter. As part of the LOOCV, we also compared the observed and predicted values of species abundance at the landscape level (in BA) and calculated the frequency at which the most abundant species was correctly predicted at the cell level.

As a general validation of our approach, we compared the stand dominant heights estimated by ALS ($Hdom_{ALS}$) to those calculated from the trees we generated ($Hdom_T$). We expect $Hdom_{ALS}$ to be as close to reality as possible, as tree height is among the most reliable ALS measurement [Van Leeuwen & Nieuwenhuis, 2010] and can be derived from ALS data with little processing and no field data. $Hdom_{ALS}$ therefore serves here as a reference to which $Hdom_T$ is compared.

In practice, $Hdom_T$ is calculated as the mean height of the six highest trees, while $Hdom_{ALS}$ is calculated as the mean height of the six highest local maxima (see *ALS metrics*). In case less than six trees/maxima were found, the mean height of all trees/maxima was used. These dominant heights are calculated at the 25×25 m^2 cell level. There is some circularity in comparing HdomALS and HdomT as models predicting BA, Dg and BA_b from ALS point clouds may include ALS derived height metrics or more generally metrics which are correlated with the dominant height estimated from ALS point clouds. The results of this comparison must therefore be interpreted with caution.

Finally, we examined the spatial distribution of species at each site and compared it to current expert knowledge.

Results

Overall, the virtual landscapes are in good agreement with the landscapes they aim to reproduce. The generated stand structures and compositions are consistent with the observations and make it possible to distinguish stands at different stages of development and with different compositions.

At Milicz, predictions are the most accurate. The RMSE of all evaluated variables are the lowest in comparison with the other landscapes (Table T1). Species abundance at the landscape level is also better reproduced (Figure F1). Finally, in 76.2% of cases, the generated main species corresponds to the observed main species. This higher quality of predictions can be explained by the fact that Milicz has the highest density of inventory plots and the least complex landscape, with a predominance of even-aged monospecific stands and the lowest species diversity among our three landscapes.

At the Bauges and Sneznik, the RMSE of the evaluated variables are comparable (Table T1.). In contrast, predictions of species abundance at the landscape level are more accurate at Sneznik (Figure F1). The same applies to the compositions predicted at the plot level: the predicted main species corresponds to the observed main species in 63.1% of cases at Sneznik and in 37.2% of cases at the Bauges. However, two datasets were used in the Bauges. In the local forest inventory

Table T1. Root mean square error (RMSE) of predictions for the three study areas obtained from the leave-one-out cross validation (LOOCV) carried out on our entire workflow. BA: basal area ($m^2.ha^{-1}$); Dg: mean quadratic diameter (cm); BA_b: broadleaf BA proportion (%); dbh: diameter at breast height (cm); h: tree height (m); $Q_{0.5}$ and $Q_{0.95}$: fiftieth and ninety-fifth percentiles, respectively, of the distribution of dbh and h. $dbhQ_{0.5}$ is not considered as it is almost similar to Dg. The RMSE values from the LOOCV of ALS models presented in Table 1, are added here in brackets to facilitate comparisons. At Sneznik, RMSE of $hQ_{0.5}$ could not be calculated as only dominant trees were measured on the field. At the Bauges, RMSE of $hQ_{0.5}$ and $hQ_{0.95}$ could not be calculated as no tree heights were measured in the field plots used to calibrate the ALS models.

Study area	Variable	RMSE
Bauges	BA	9.5 (8.3)
	Dg	5.4 (4.2)
	BA_b	21.6 (20.3)
	$dbhQ_{0.95}$	16.1
	$hQ_{0.5}$	-
	$hQ_{0.95}$	-
Milicz	BA	5.4 (5.4)
	Dg	3.9 (3.7)
	BA_b	13.1 (12.9)
	$dbhQ_{0.95}$	8.8
	$hQ_{0.5}$	5.0
	$hQ_{0.95}$	2.6
Sneznik	BA	9.6 (9.6)
	Dg	7.9 (7.6)
	BA_b	20.0 (19.3)
	$dbhQ_{0.95}$	13.1
	$hQ_{0.5}$	-
	$hQ_{0.95}$	4.7

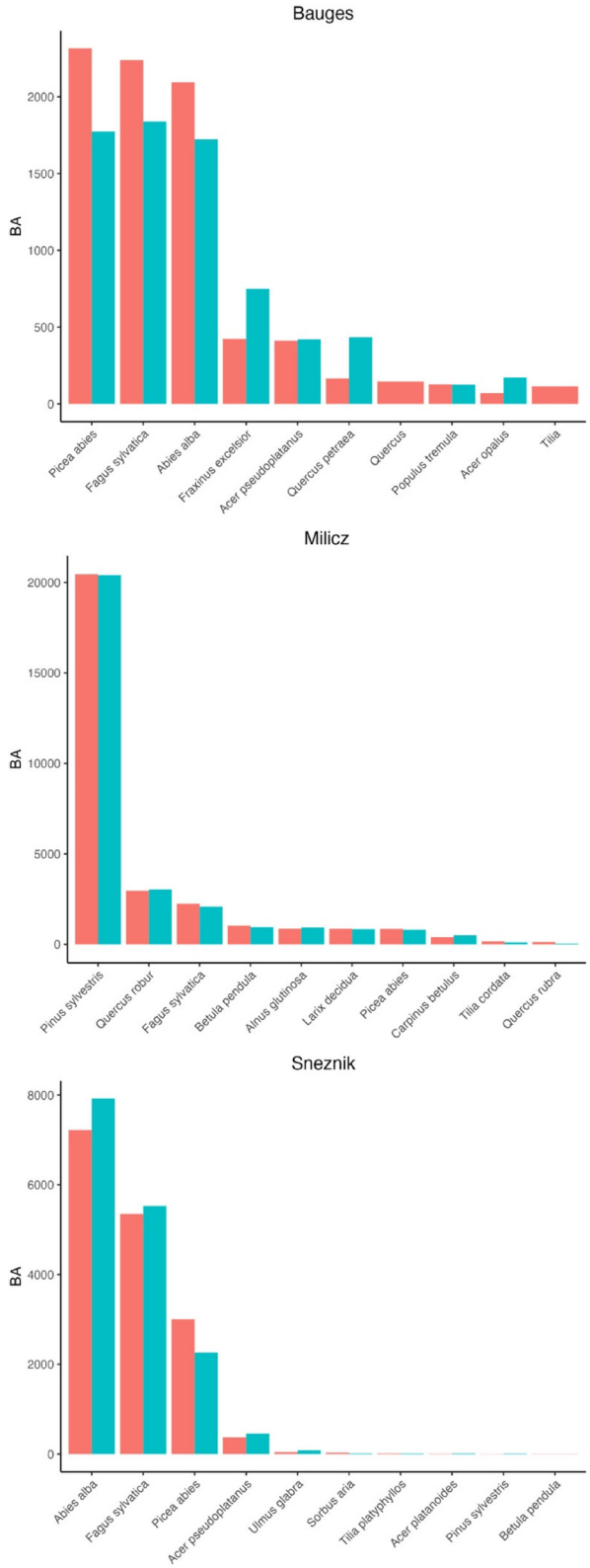

Figure F1. Predicted (blue) and observed (red) species abundance in BA (m^2) at the landscape level. In the Bauges, we only considered trees with a dbh greater than 17.5, as smaller trees were not identified in the local forest inventory (LFI) but only grouped in two categories (coniferous and broadleaf). Also, some predictions are missing in the Bauges because some trees in the LFI were not identified at the species level and therefore can't find a match in the generated trees which all receive a species name.

(LFI) not all trees were identified at the species level and trees with a dbh between 7.5 and 17.5 cm were not measured but counted by diameter classes and grouped in two categories (coniferous and broadleaf). This led us to use a local subset of the NFI from which composition is derived in our downscaling algorithm. The poorer composition predictions in the Bauges might therefore partly be an artefact arising from the evaluation itself, as the LFI may not be suitable to serve as a field reference.

The fact that the RMSE values obtained from the LOOCV carried out on our entire workflow are almost similar to the RMSE values obtained from the LOOCV of ALS models shows that the downscaling algorithm hardly adds any error (Table 1, Table T1). The main way of increasing the realism of our virtual landscapes would therefore be to improve the ALS models.

With R^2 values ranging from 0.61 to 0.83 (Figure 4) and RMSE values below 5 m, HdomALS and HdomT are consistent with one another. This provides a general validation of our workflow. As discussed above, the better predictions obtained at Milicz might steam from the higher density of inventory plots and the lower complexity of the landscape. At Sneznik, HdomT tends to be overestimated as HdomALS decreases. This divergence could be due to the ice storm that occurred between the field inventory and the ALS acquisition and that might have biased the ALS models.

Overall, species spatial distribution in the virtual landscapes is consistent with field observations. In the Bauges, pure and mixed stands of fir and spruce are more abundant at higher elevation while mixed stands of broadleaf species are found at lower elevation. At Milicz, pure stands of Scots pine are found at lower elevation while broadleaf species and mixed stands appear at higher elevation. Finally, at Sneznik, pure beech stands are found at higher elevation while fir is found at lower elevation in pure or mixed stands (a specific feature of the site).

Our procedure is not free of flaws and some outliers are present in the generated data (i.e. stands with extreme values of BA, Dg, tree height or density). These outliers are a direct consequence of the uncertainties associated with the models we used. The realism of the stands associated with these extreme values is open to question. However, separating realistic from unrealistic stands seems difficult as extreme values can be locally observed. It is therefore up to the users of the dataset to decide whether or not to consider these stands depending on their objectives.

Virtual landscapes overview

Overall, 42,394,479 trees belonging to 51 different species were generated: 35,134,985 trees of 40 different species were generated at the Bauges, 5,726,420 trees of 32 different species at Milicz and 1,533,074 trees of 16 different species at Snežnik. The main species BA proportion as well as their h and dbh distributions are shown in Figure 5 for each virtual landscape.

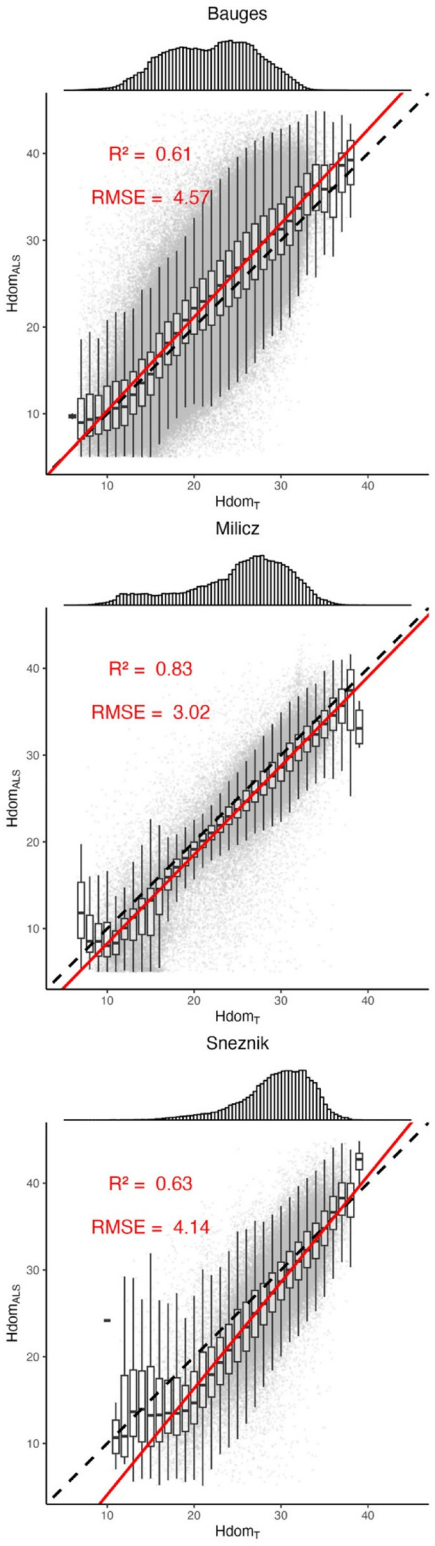

Figure 4. Comparison of the stands dominant heights measured by ALS ($Hdom_{ALS}$, in m) to those calculated from the generated trees ($Hdom_T$, in m). The top panels show the distribution of $Hdom_T$. The dashed lines indicate the y = x line. The red lines correspond to the regression lines. The root mean square error (RMSE) values between $Hdom_{ALS}$ and $Hdom_T$, as well as the regression R-Squared values are shown in red.

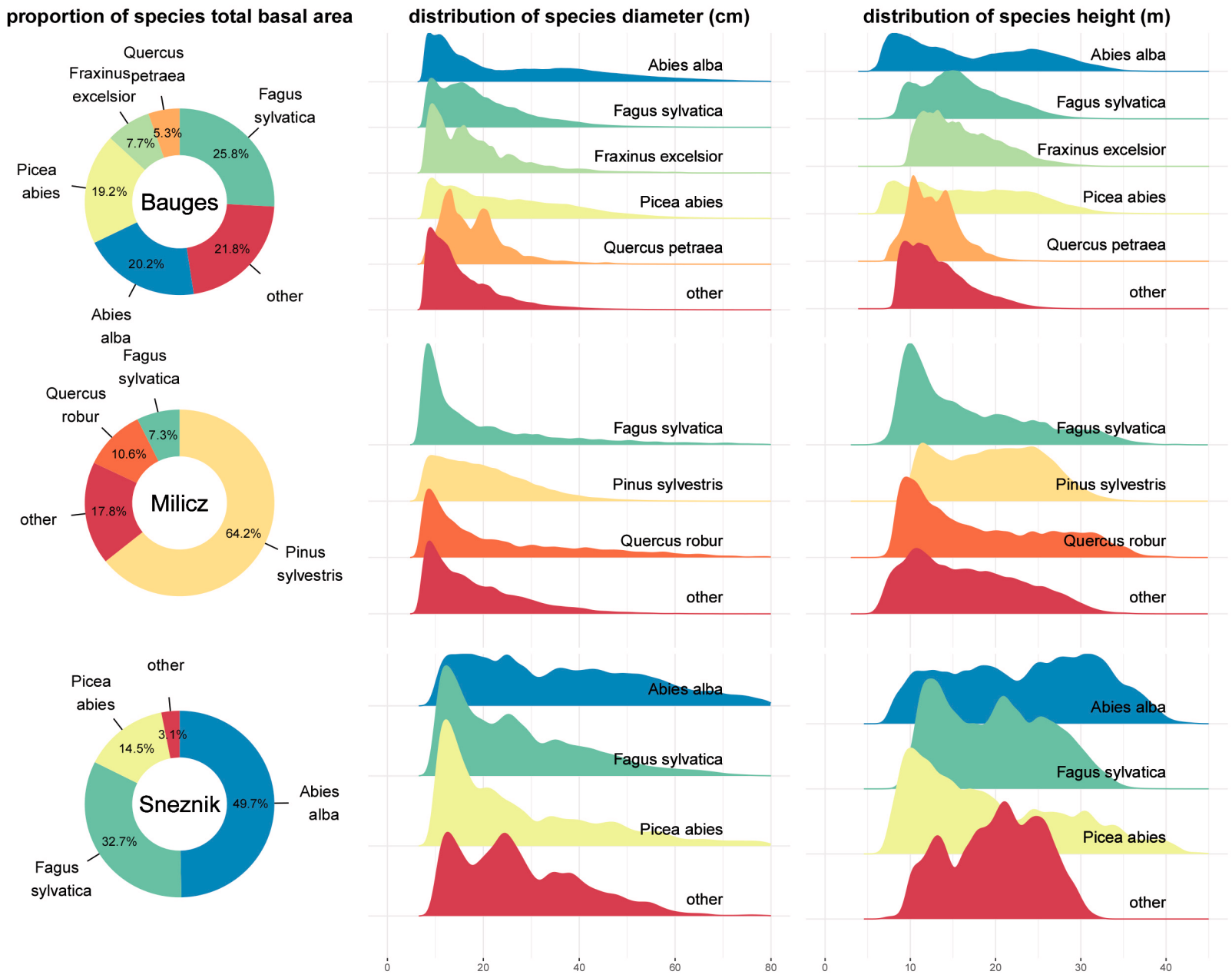

Figure 5. Main species basal area proportion, diameter distribution and height distribution in the three virtual landscapes. Species accounting for less than 5% of the virtual landscapes total basal area were grouped in the 'other' category.

Open Research Europe 2023, 3:32 Last updated: 24 JAN 2024

Data availability

Underlying data

Bauges

- The maps of forest types (BD Forêt®V2) are available to download from the National Institute for Geographic and Forestry Information website at https://geoservices.ign.fr/bdforet, under the Etalab open license 2.0.
- The French National Forest Inventory data are available to download from the National Institute for Geographic and Forestry Information website at https://inventaire-forestier.ign.fr/dataifn/, under the Etalab open license 2.0.
- The local forest inventory dataset is available for non-commercial use upon request to Jean-Matthieu Monnet (jean-matthieu.monnet@inrae.fr). A data sharing agreement will have to be established, with the following restrictions:
  - data are available for internal use only and cannot be distributed;
  - results obtained from the data can be displayed or distributed provided they do not allow the estimation of growing stock in individual private properties;
  - data funding (Ademe grant 1703C0069) should be cited.
- ALS data in the northern part (Haute-Savoie) are available to download from the Recherche Data Gouv dataverse at https://doi.org/10.57745/ZUT1MJ, under the Etalab open license 2.
- ALS data in the southern part (Savoie) can be purchased upon request to (Régie de Gestion des Données Savoie Mont Blanc) at https://www.rgd.fr/.

Milicz

- The stand data in the ESRI Shapefile format are available to download from the Polish Forest Data Bank at https://www.bdl.lasy.gov.pl/portal/wniosek-en.
- The local forest inventory dataset and ALS data are available for non-commercial use upon request to Jarosław Socha (jaroslaw.socha@urk.edu.pl). A data sharing agreement will have to be established, with the following restrictions:
  - data are available for internal use only and cannot be distributed;
  - data funding (REMBIOFOR - BIOSTRATEG1/267755/4/NCBR/2015) should be cited.

Sneznik

- The forest inventory data (in *.xlsx and *.shp formats) and maps of forest types and species mixture (in *.shp format) are available upon request to Slovenia Forest Service (zgs.tajnistvo@zgs.si; rok.pisek@zgs.si). A data sharing agreement will have to be established, with the following restrictions:
  - data are only available for the study that is the subject of the agreement;
  - Slovenia Forest Service should be acknowledged for providing the data in all publications.
- ALS data are available to download from the Slovenian Environment Agency website at http://gis.arso.gov.si/evode, under the terms of the international Creative Commons 4.0 license (http://www.evode.gov.si/fileadmin/user_upload/Lidar_pogoji_uporabe.pdf):
  - the data user must indicate the data source at each publication of data or products, specifying "Slovenian Environmental Agency, type of data and period to which the data refer or the date of the database".

Extended data

Zenodo: I-MAESTRO data: 42 million trees from three large European landscapes in France, Poland and Slovenia. https://doi.org/10.5281/zenodo.7462440 [Aussenac *et al.*, 2022].

For each virtual landscape we provide a table (in .csv format) with the following columns:

- cellID25: the unique ID of each 25x25 m^2 cell
- sp: species latin names
- n: number of trees. n is an integer >= 1, meaning that a specific set of species "sp", diameter "dbh" and height "h" can be present multiple times in a cell.
- dbh: tree diameter at breast height (cm)
- h: tree height (m)

We also provide, for each virtual landscape, a raster (in .asc format) with the cell IDs (cellID25) which makes data spatialisation possible. The coordinate reference systems are EPSG: 2154 for the Bauges, EPSG: 2180 for Milicz, and EPSG: 3912 for Sneznik.

We provide Table S1 presenting the metrics used in the 32 stratum-specific prediction models of BA and Dg.

Finally, we provide a proof of how, in the downscaling algorithm, multiplying the trees dbh by the α correction coefficient makes it possible to reach the cells BA value derived from the ALS mapping.

Acknowledgments

The authors would like to thank the ONF and PNR du Massif des Bauges for their contribution to the field and ALS data collection in the French study area, as well as the IGN for providing freely the French National Forest Inventory data. The authors also wish to thank the Slovenia Forest Service for providing the forest inventory data from the Slovenian study area, and the Ministry of Education, Science and Sport of the Republic of Slovenia for funding the project. Finally, the authors would like to thank the Polish Forest Management and Geodesy Bureau for providing data from the Polish study area.

References

Aussenac R, Monnet JM, Klopcic M, *et al.*: **I-maestro data: 42 million trees from three large european landscapes in france, poland and slovenia.** 2022.
http://www.doi.org/10.5281/zenodo.7462440

Box GEP, Cox DR: **An analysis of transformations.** *Journal of the Royal Statistical Society: Series B (Methodological).* 1964; **26**(2): 211–243.
Publisher Full Text

Bureau for Forest Management and Geodesy: **Forest data bank.** 2020.
Reference Source

Cazzolla Gatti R, Reich PB, Gamarra JGP, *et al.:* **The number of tree species on earth.** *Proc Natl Acad Sci U S A.* 2022; **119**(6): e2115329119.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Craven D, van der Sande MT, Meyer C, *et al.:* **A cross-scale assessment of productivity–diversity relationships.** *Glob Ecol Biogeogr.* 2020; **29**(11): 1940–1955.
Publisher Full Text

IGN: **La BD Forêt ® v2 - Une cartographie forestiere nationale pour les territoires.** 2019.
Reference Source

Lamb SM, MacLean DA, Hennigar CR, *et al.:* **Forecasting forest inventory using imputed tree lists for lidar grid cells and a tree-list growth model.** *Forests.* 2018; **9**(4): 167.
Publisher Full Text

Liang J, Crowther TW, Picard N, *et al.:* **Positive biodiversity-productivity relationship predominant in global forests.** *Science.* 2016; **354**(6309): aaf8957.
PubMed Abstract | **Publisher Full Text**

Mauri A, Strona G, San-Miguel-Ayanz J: **Eu-forest, a high-resolution tree occurrence dataset for europe.** *Sci Data.* 2017; **4**: 160123.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Papaik MJ, Fall A, Sturtevant B: **Forest processes from stands to landscapes: exploring model forecast uncertainties using cross-scale model comparison.** *Can J For Res.* 2010; **40**(12): 2345–2359.
Publisher Full Text

Seidl R, Eastaugh CS, KramerK: **Scaling issues in forest ecosystem management and how to address them with models.** *Eur J For Res.* 2013; **132**: 653–666.
Publisher Full Text

Silva CA, Hudak AT, Vierling LA, *et al.:* **Imputation of individual longleaf pine (*pinus palustris* mill.) tree attributes from field and lidar data.** *Can J Remote Sens.* 2016; **42**(5): 554–573.
Publisher Full Text

Slovenia Forest Service: **Gis database on forest stands.** Slovenia Forest Service, Ljubljana, Slovenia. 2020.

van Leeuwen M, Nieuwenhuis M: **Retrieval of forest structural parameters using lidar remote sensing.** *Eur J Forest Res.* 2010; **129**(4): 749–770.
Publisher Full Text

White JC, Wulder MA, Varhola A, *et al.:* **A best practices guide for generating forest inventory attributes from airborne laser scanning data using an area-based approach.** Technical report, Natural Resources Canada, Canadian Forest Service, Canadian. Wood Fibre Centre, Victoria, BC. 2013; **89**(6): 722–723.
Publisher Full Text

With KA: **14Scaling Issues in Landscape Ecology.** In: *Essentials of Landscape Ecology.* Oxford University Press, 2019; 14–41.
Publisher Full Text

Open Peer Review

Current Peer Review Status: ✔ ✔

Version 2

Reviewer Report 24 January 2024

https://doi.org/10.21956/openreseurope.18268.r36612

✔ **Fabian Fischer**
School of Biological Sciences, University of Bristol, Bristol, England, UK

The authors have done an impressive job addressing the reviewers' comments and the paper now presents an intriguing study of how one could produce large-scale inventory maps. The validation procedure, in particular, is now comprehensive and, by relying on leave-one-out cross validation (LOOCV) across the whole workflow, provides readers with a clear idea of the quality of the product.

I have only a few minor comments:

- R packages should generally be cited in the references, but this seems not the case (I could not find references for, e.g, *MASS* package, *lidaRtRee* package)
- From my understanding, the whole workflow was applied separately at each site, but this is not immediately clear when reading the paper. It would be worth mentioning this early on somewhere in the methods section.
- For the validation tables (Tables 1 and T1), it would have been nice to also see a relative RMSE (rRMSE), i.e., RMSE divided by the standard deviation or mean of the corresponding variable. This would help readers understand whether these prediction errors are large or not, make BA/D_g/BA_b errors comparable between each other and also make estimates better comparable across sites
- Generally speaking, I found that there was not much information on the ALS data processing. How was ground classification done? Did you remove noise? Is it all done with the *lidaRtRee* R package? This is not hugely important for your study, and if it was all done already by the data providers and with a variety of methods, you could also just state that.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: My areas of expertise are in lidar processing, individual-based modelling, as well as the creation of simulated forest stands (cf. my 2020 paper on this topic, mentioned in the review), which is very close to what the authors have been working on.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 29 December 2023

https://doi.org/10.21956/openreseurope.18268.r36613

Nikolai Knapp
Thünen Institute of Forest Ecosystems, Eberswalde, Germany

Thank you for revising the manuscript, for adding further validation and for the detailed answers to the comments. I only have one further remark: It would be good to mention the used lidar metrics explicitly, either by summarizing them in the text or listing them in a table. If they are only referenced as functions in the lidaRtRee package, readers have to search in the package documentation and, more importantly, the approach may become unreproducible, in case the package changes in the future and the used metrics were not documented. Otherwise, I congratulate the authors on this well written manuscript and I am looking forward to future applications of the method.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Forest monitoring, forest modeling, lidar remote sensing

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 12 May 2023

https://doi.org/10.21956/openreseurope.16618.r31138

Nikolai Knapp
Thünen Institute of Forest Ecosystems, Eberswalde, Germany

The paper presents an innovative approach to generate forest stand structure information at

landscape extent and single tree resolution based on airborne lidar data and inventory plots. The approach is not based on individual tree detection (ITD) from lidar, but operates in an area-based (ABA) fashion at 25 m x 25 m cell scale. The inventory plots serve as lookup tables. Structure metrics are being estimated for every cell in the landscape based on lidar metrics. Then, each cell is being assigned to the most similar stand from the inventory lookup table based on a minimum distance of a set of structure metrics. The dbh values of the trees are then adjusted according to a proposed algorithm, such that the final structure metrics match the ones predicted for the cell. Finally, the generated forest landscapes are being validated by calculating dominant height for each cell based on the generated stands and comparing them to dominant heights directly obtained from lidar. The approach has been applied to three different regions in France, Poland and Slovenia.

The presented approach is very interesting and useful as an efficient solution to generate maps at single tree resolution and landscape extent, which are highly relevant, e.g., for spatial and temporal interpolation of forest inventories and for modelling tasks. The method is well documented and the case studies along with the provided datasets make it an innovative publication. However, I have listed some comments below, which the authors should consider during revision.

Detailed comments:

- In the Abstract, I suggest to remove the tilde signs from 100~000~ha.
- On page 4 “For that, we first assigned to each cell a stand from the field data based on the similarity of their BA, Dg and BAb values.” it should already be briefly mentioned how “similarity” is defined, i.e. minimum distance of normalized values.
- I suggest to mention earlier (in the Abstract or Introduction), that the study follows an ABA approach, because readers might expect an ITD approach, if the final product are landscapes at tree level.
- Why were BA and Dg chosen as the structure metrics for matching? Would it not be important to also consider metrics that capture stem size heterogeneity / stem size distribution?
- On page 6, what is meant by “Point cloud metrics were directly computed from the point cloud or(?) from the derived CHM”? I suggest to list all lidar metrics which were used in a table.
- In Table 1, why are RMSE values for BAb > 1? In case they are given in percent, please add “(%)” to the caption.
- On page 9, the multiplication by 40000/pi and the division by 16 need to be explained. I suspect they convert values to the 1 ha and then back to the 25-m scale, however these scale factors should be explained explicitly. Also, the purpose of the rounding under “c)” should be better explained.
- Figure 4: What is the explanation for the seemingly better fit (higher R^2) in Milicz compared to Bauges?

General comments (for a possible Outlook):

- Unlike an ITC approach, the presented method does not provide precise tree positions within the 25-m cells. Are there ways to expand the approach to additionally generate tree positions?

- Would it be possible/useful to add a height correction algorithm based on ALS heights (local maxima), similar to the dbh adjustment algorithm?

Comments about the data:

- The information about the coordinate reference system is missing. I was not able to georeference the asc files in a GIS.

It would be better to use unique file names, e.g. “milicz_cellID25.asc” etc. to be able to load all rasters in one GIS session.

Is the rationale for creating the dataset(s) clearly described?

Yes

Are the protocols appropriate and is the work technically sound?

Yes

Are sufficient details of methods and materials provided to allow replication by others?

Yes

Are the datasets clearly presented in a useable and accessible format?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Forest monitoring, forest modeling, lidar remote sensing

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 24 Nov 2023

Raphaël Aussenac

We would like to thank the reviewers for their useful comments which have helped significantly improve the manuscript. A special effort has been made to provide an in-depth evaluation of the generated landscapes, to clarify the functioning of our downscaling algorithm and to consolidate the rational of our work.

This revision has led us to optimise our code (better memory management, more effective parallelisation, etc.). It is now much faster, taking less than 1 hour to generate the 42 million trees of our three landscapes, while it used to take “less than 5 hours” to generate the 35 million trees of the Bauges landscape. We updated the introduction accordingly.

Please find below our replies to your comments.

The paper presents an innovative approach to generate forest stand structure information at landscape extent and single tree resolution based on airborne lidar data and inventory plots. The approach is not based on individual tree detection (ITD) from lidar, but operates in an area-based (ABA) fashion at 25 m x 25 m cell scale. The inventory plots serve as lookup tables. Structure metrics are being estimated for every cell in the landscape based on lidar metrics. Then, each cell is being assigned to the most similar stand from the inventory lookup table based on a minimum distance of a set of structure metrics. The dbh values of the trees are then adjusted according to a proposed algorithm, such that the final structure metrics match the ones predicted for the cell. Finally, the generated forest landscapes are being validated by calculating dominant height for each cell based on the generated stands and comparing them to dominant heights directly obtained from lidar. The approach has been applied to three different regions in France, Poland and Slovenia.

The presented approach is very interesting and useful as an efficient solution to generate maps at single tree resolution and landscape extent, which are highly relevant, e.g., for spatial and temporal interpolation of forest inventories and for modelling tasks. The method is well documented and the case studies along with the provided datasets make it an innovative publication. However, I have listed some comments below, which the authors should consider during revision.

Detailed comments:

In the Abstract, I suggest to remove the tilde signs from 100~000~ha. **Done as suggested**

On page 4 “For that, we first assigned to each cell a stand from the field data based on the similarity of their BA, Dg and BAb values.” it should already be briefly mentioned how “similarity” is defined, i.e. minimum distance of normalized values. **We added a definition of similarity at the end of the sentence as follows: “For that, we first assigned to each cell a stand from the field data based on the similarity of their BA, Dg and BA_b values (calculated as the Euclidean distance between each cell and each field plot in the three-dimensional space made up by the scaled values of BA, Dg and BA_b).”**

I suggest to mention earlier (in the Abstract or Introduction), that the study follows an ABA approach, because readers might expect an ITD approach, if the final product are landscapes at tree level. **Thank you for pointing that out. To make that point clear from the beginning, we now specify in the abstract that we use an ABA: “Tree diameter, height and species data were generated combining field data, vegetation maps and airborne laser scanning (ALS) data following an area-based approach.” To be consistent throughout the article, we also replaced, in the ALS mapping section, “area-based method” for “area-based approach.”**

Why were BA and Dg chosen as the structure metrics for matching? Would it not be important to also consider metrics that capture stem size heterogeneity / stem size distribution? **BA and Dg were chosen because together they capture the development**

stage of stands. Along with BA_b, they provide a synthetic yet fairly accurate picture of the stands and seem therefore appropriate for our matching procedure. We added this point in the Algorithm section as follows: "2. Second, we associated to each 25×25 m^2 cell a field plot based on the similarity of their BA, Dg and BA_b. These 3 variables were chosen for matching because together they provide a synthetic yet fairly accurate picture of the stands." Stem size heterogeneity or distribution would certainly bring another information layer. However, stem-related measurements (number, size, etc.) obtained from ALS are not very accurate. Including such variables in the procedure could in fact lead to less relevant matches.

On page 6, what is meant by "Point cloud metrics were directly computed from the point cloud or(?) from the derived CHM"? I suggest to list all lidar metrics which were used in a table. **We removed the second part of the sentence "or from the derived CHM" for better clarity. As for the lidar metrics, see response to Reviewer 1's comment on ALS metrics.**

In Table 1, why are RMSE values for BAb > 1? In case they are given in percent, please add "(%)" to the caption. **RMSE values are indeed given in percent. We modified the table caption as suggested.**

On page 9, the multiplication by 40000/pi and the division by 16 need to be explained. I suspect they convert values to the 1 ha and then back to the 25-m scale, however these scale factors should be explained explicitly. Also, the purpose of the rounding under "c)" should be better explained. **We added an explanation of the 40000 scale factor in a foot note as it appears in two different equations: "The scale factor 40000 is the product of two scale factors: 4 x 10000. The scale factor 4 comes from the formula linking a surface area *S* to a diameter *d* ($S=\pi d42$); while the scale factor 10000 accounts for the difference in units between the diameters (in cm) and the basal areas (in m^2)."**

We also clarified the rounding procedure and the use of the scale factor 16 by modifying the "c)" paragraph as follows: "c) Finally, we divided ω by 16 to get the weight of the trees in the 25×25 m^2 cells (ω being a weight per ha and 16 being the surface area ratio between 1 ha and a 25×25 m^2 cell). In doing so, the obtained tree weights can be either integer or decimal. However, the objective of our algorithm is to generate for each cell a list of individual trees with their associated diameter, height and species. From this perspective, decimal weights are not useful. We cannot simply round the tree weights to the nearest integer as this can lead to a significant over- or underestimation of the total number of trees in the cells. This is because the decimal part of the tree weights in the 25×25 m^2 cells is not the result of a random draw but directly depends on the surface area ratio between the field plot and the cell. As an example: 1 tree inventoried on a 400 m^2 field plot will always obtain a weight of 1.56 in a 25 x 25 m^2 cell, and a weight of 2 after rounding to the nearest integer. In order to obtain integer tree weights in the 25 x 25 m^2 cells while avoiding this bias, we performed a Bernoulli draw on the decimal part of the tree weights. As an example, a weight of 1.56 has a 56% chance of becoming 2, and a 44% chance of becoming 1. As this rounding of the weights slightly modifies the total BA of the generated stand, we transformed again the trees dbh to reach the total BA provided by the ALS mapping

using the trees BA and their integer weights (ω_{int}) as follows: [...]"

Figure 4: What is the explanation for the seemingly better fit (higher R^2) in Milicz compared to Bauges? **The better predictions obtained at Milicz are discussed in the revised version of the validation section: "This higher quality of predictions can be explained by the fact that Milicz has the highest density of inventory plots and the least complex landscape, with a predominance of even-aged monospecific stands and the lowest species diversity among our three landscapes."**

General comments (for a possible Outlook): **We provide below some answers to the following two general comments. However, we would rather not include in the article the discussed points because our responses can only be speculative at this stage and further analyses would be required to provide solid answers. Our article is already rather dense for a data note, as pointed out by reviewer1, and we would prefer not to lengthen it further with speculative considerations.**

Unlike an ITC approach, the presented method does not provide precise tree positions within the 25-m cells. Are there ways to expand the approach to additionally generate tree positions? **With the ABA and the LFI (local forest inventory) field plots as look up table, the tree lists in each pixel is very close to existing stands from the landscape. With an ITC method, the dominant trees and their position and heights can be retrieved, and the diameter and species estimated. In case the detection parametrization is chosen so as to avoid comission errors, stands need to be populated with additional trees to compensate for the omission errors. Our proposed workflow could be adapted to assign detected positions to trees in the list according to their sorted heights and then to randomly (or based on a model) assign positions to trees with no detected position.**
Another possibility would be to use a semi-ITC approach as proposed by [https://doi.org/10.1016/j.rse.2009.12.004] to assign tree groups from detected clusters in the LFI reference plots to similar detected clusters in the landscape. This approach is interesting as it:

- **provides the coordinates of the main detected trees**
- **directly provides a tree list at the lidar « detection cluster » scale, which can then be aggregated in larger areas (pixels, polygons)**

The main drawback with this approach is that it requires a point density compatible with ITC analysis. In the southern part of the Bauges study area, the point density is too low to implement this approach.

Would it be possible/useful to add a height correction algorithm based on ALS heights (local maxima), similar to the dbh adjustment algorithm? **Our approach is based on the fact that stands basal area (BA), quadratic diameter (Dg) and density (N) are deterministically linked. As it stands, it seems difficult to add a height correction based on ALS heights to our approach. There would be one too many unknowns in the equation system, making it intractable. A different approach could be developed, where instead of trying to reach the stands BA by modifying tree diameters, the aim would be to reach the stands total volume by modifying individual tree volumes. For this purpose, stand volume models from ALS point cloud should be created as well as**

models predicting individual tree volume as a function of tree height and diameter. The latter could be constrained by ALS heights. However, volume allometries come with their own uncertainty and whether the generated stands would be more realistic following this procedure remains to be tested. More generally, there might be a trade-off between adjusting to the local lidar values, and making sure that we create an unbiased landscape with valid stands.

Comments about the data:

The information about the coordinate reference system is missing. I was not able to georeference the asc files in a GIS. T**his is indeed an oversight. Thank you for pointing it out. We specified the coordinate reference systems for each site in the presentation of our dataset on the zenodo website and in the Extended data section: "We also provide, for each virtual landscape, a raster (in .asc format) with the cell IDs (cellID25) which makes data spatialisation possible. The coordinate reference systems are EPSG: 2154 for the Bauges, EPSG: 2180 for Milicz, and EPSG: 3912 for Sneznik." We also added the coordinate reference systems in the R script provided alongside our dataset on the zenodo website. We took this opportunity to replace the raster package, which may no longer be fully maintained in the near future, with the terra package.**

It would be better to use unique file names, e.g. "milicz_cellID25.asc" etc. to be able to load all rasters in one GIS session. **Modified as suggested**

Competing Interests: No competing interests were disclosed.

Reviewer Report 04 April 2023

https://doi.org/10.21956/openreseurope.16618.r30982

?

Fabian Fischer
School of Biological Sciences, University of Bristol, Bristol, England, UK

Overall assessment

The article by Aussenac *et al.* describes a statistical procedure to generate a large data set of individual trees from airborne laser scanning (ALS) and inventory data. The variables include trunk diameter, tree height and species identity, and are provided across three European landscapes. The result is an impressive number of simulated/potential trees, which is a useful data set in forest ecology. As applications, the authors mention studies of scale and (more vaguely) forest management/ecosystem prediction, but one could easily think of a number of other concrete applications, such as input/validation of individual-based models of forest dynamics, or

comparisons with automatically mapped tree crowns from airborne imagery, e.g. as in Weinstein *et al.* 2021[1], Ball *et al*. 2022[2], or spaceborne imagery, as in Tucker, Brandt, Hiernaux, *et al.* 2023[3].

I also found the paper generally well-written and with a well-thought through methodology for the mapping. The authors carefully tune their models to obtain optimal performance at every step and clearly have spent considerable amounts of time and effort to improve the prediction of stand attributes. In particular, I found the idea of matching predicted basal area to real stands and then filling in/removing trees until the basal area matches intriguing. This bears similarities with model-based estimations of forest attributes/tree attributes from lidar (Hurtt *et al.* 2004[4], Taubert *et al*. 2015[5], Rödig *et al.* 2017[6], Fischer *et al*. 2020[7]) and shares some of these models' advantages (e.g. more fine-scale distribution of biomass, no shrinking to the mean).

However, like these models, the authors' method also involves a lot of complex modelling steps, and it is in the validation step of the procedure that I see deficiencies that need to be addressed. I see two main issues:

a) the robustness of the models to extrapolation issues and spatial autocorrelation is not evaluated, so it is hard to assess how good the models are outside their calibration range and how much we can trust the predictions across the landscape.

b) two of the key attributes of the data set (tree diameter and species identity) are not validated at all, despite featuring prominently in the title and in the results section (Figure 5). This should be a priority in a revised version.

In the following I will provide a few comments on the article following roughly the overall structure, and give suggestions on how to improve the model validation.

Justification for the data set

I see the value of a fine-grain large-scale data set, and having such a data set is indeed rare, but it would be helpful to mention concrete applications. At the moment, the only justification given is the sentence: "Yet, this type of data could help address the scaling issues in ecology and could prove useful for testing forest management strategies and accurately predicting the dynamics of ecosystem services". This is the sentence from the abstract, but the same point is made at the end of the first paragraph. Could the authors rephrase and add literature references in the main text? The vast majority of data sets can be useful for the testing of forest management strategies or predicting dynamics of ecosystem services. What is unique to your data set? Why do we need detailed, tree-based data at large scales?

Model for mapping of tree attributes

ALS metrics: which metrics precisely did you use?

Point cloud properties: Could the authors add information on/discussion of the sensitivity of their point metrics to scanner acquisitions? Lidar scans often exhibit considerable variation in pulse density even within a single acquisition (e.g. scan line centre vs. overlapping scan lines). What is each scan's standard deviation of point/pulse density? Could you include that as a variable in stratification? Could this improve your models (e.g. stratify by pulse densities between 5 and 10,

10 and 15, 15 and 20, etc., or even smaller step sizes)?

Descriptions: I appreciate that the paper is already quite dense, but quite a few steps in the methods section remain unclear to me, particularly in step 3. E.g., in the matching of BA and BAb, why do you need a correction value alpha? Can you explain the weighting better and why it is divided by 16? Maybe this is more exhaustively explained in the Extended Data, but this needs to be clear from the main text already.

Model validation

As pointed out above, this is the point of the paper that needs to be more comprehensive. At the moment, the authors validate their approach by comparing dominant height, as obtained from lidar (mean height of six highest local maxima), to dominant height of the simulated stands, obtained via local allometries (mean height of six highest trees). It is definitely useful to do this comparison and good to see that the results are broadly consistent, so I would keep it in the paper. However, there are issues with circularity, as the authors first use a number of lidar metrics that involve height / basal area-to-height relationships to create the maps and then compare the inferred results (+ independently derived height allometries) again to lidar-derived height metrics. Furthermore, height of the dominant trees may be related to basal area, but it cannot be used to evaluate basal area/tree diameter predictions as such, nor does it validate predicted species composition - both are key features of the data set.

Given that the author's simulation approach seems fast (only ca. 5 hours on a modern laptop, amazing!), another approach suggests itself, namely within-site cross-validation, ideally in the form proposed by Ploton *et al.* 2020[8]. Since a spatially explicit leave-one-out cross-validation, as suggested in Ploton *et al.* 2020[8], may be too computationally intensive, I would recommend the simpler approach proposed in the same paper: for each of the European landscapes, I would recommend the authors to split their field data sets into, e.g., 5 spatially aggregated folds (i.e., spatial clusters), and run their model 5 times, each times using 4 folds to train the model and 1 separate geographic fold of plots to validate the model. In this 1 fold, the authors could directly compare predictions of tree values to actual data according to some simple standard metrics (total basal area, mean quadratic diameter, 95th percentile of diameter, percentage of species xyz, 95th percentile of height, mean height, dominant height). For comparison and to broadly assess whether spatial autocorrelation makes a difference, the authors could do the same validation procedure also with 5 folds containing plots randomly distributed in space (so no spatial clusters). This would only take 25 hours for each validation and give a good impression of how easy it is to accurately map individual trees and species at landscape scale and how realistic the produced inventories are. It would likely also increase interest in the data set, as it would give potential users higher confidence in the results.

Since the paper puts its focus on the value of individual trees, there should, in my opinion, also be one result/validation graph that shows individual trees in some way. It could be, for example, a zoomed-in image of lidar-derived canopy height models + a predicted distribution of trees. If the 5-fold cross-validation is carried out, as above, the authors could simply show sample lidar canopy height models on top of plots, and the diameter distributions for the simulated and the inferred plots.

Overall, it would also be interesting to readers to understand in how far the predicted species

distributions reflect current expert knowledge, but this is not a necessity.

Data set

I had a quick look at the data set. One variable I did not understand was the variable “n” or “number of trees”. Could you explain it a bit better? Does this mean that the specific diameter exists n times in the specific data set? If this is true (and only in this case), I seem to get some cells (very few) of 25m by 25m (e.g. cellID25 = 2439821 in the “Bauges” data set) that contain more than 500 trees with dbh >= 9-10cm per 625m2 and a total basal area >= 6m2 (which would yield roughly 100m^2 per hectare, at densities of 8000 trees). These are outliers, and every model is allowed to have outliers (and nature is full of them too), but it would be interesting to get your take on that in terms of realism/stand type. It could also be part of the validation to assess the edges of the basal area distribution or to give readers a hint what to make of the most extreme values.

References

1. Weinstein BG, Marconi S, Bohlman SA, Zare A, et al.: A remote sensing derived data set of 100 million individual tree crowns for the National Ecological Observatory Network.*Elife*. 2021; **10**. PubMed Abstract | Publisher Full Text
2. Ball J, Hickman S, Jackson T, Koay X, et al.: Accurate delineation of individual tree crowns in tropical forests from aerial RGB imagery using Mask R-CNN. *bioRxiv*. 2022. Publisher Full Text
3. Tucker C, Brandt M, Hiernaux P, Kariryaa A, et al.: Sub-continental-scale carbon stocks of individual trees in African drylands.*Nature*. 2023; **615** (7950): 80-86 PubMed Abstract | Publisher Full Text
4. Hurtt GC, Dubayah R, Drake J, Moorcroft PR, et al.: Beyond Potential Vegetation: Combining Lidar Data and a Height-Structured Model for Carbon Studies. *Ecological Applications*. 2004; **14** (3): 873-883 Publisher Full Text
5. Taubert F, Jahn MW, Dobner HJ, Wiegand T, et al.: The structure of tropical forests and sphere packings.*Proc Natl Acad Sci U S A*. 2015; **112** (49): 15125-9 PubMed Abstract | Publisher Full Text
6. Rödig E, Cuntz M, Heinke J, Rammig A, et al.: Spatial heterogeneity of biomass and forest structure of the Amazon rain forest: Linking remote sensing, forest modelling and field inventory. *Global Ecology and Biogeography*. 2017; **26** (11): 1292-1302 Publisher Full Text
7. Fischer F, Labrière N, Vincent G, Hérault B, et al.: A simulation method to infer tree allometry and forest structure from airborne laser scanning and forest inventories. *Remote Sensing of Environment*. 2020; **251**. Publisher Full Text
8. Ploton P, Mortier F, Réjou-Méchain M, Barbier N, et al.: Spatial validation reveals poor predictive performance of large-scale ecological mapping models.*Nat Commun*. 2020; **11** (1): 4540 PubMed Abstract | Publisher Full Text

Is the rationale for creating the dataset(s) clearly described?

Yes

Are the protocols appropriate and is the work technically sound?

Partly

Are sufficient details of methods and materials provided to allow replication by others?

Yes

Are the datasets clearly presented in a useable and accessible format?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: My areas of expertise are in lidar processing, individual-based modelling, as well as the creation of simulated forest stands (cf. my 2020 paper on this topic, mentioned in the review), which is very close to what the authors have been working on.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 24 Nov 2023

Raphaël Aussenac

We would like to thank the reviewers for their useful comments which have helped significantly improve the manuscript. A special effort has been made to provide an in-depth evaluation of the generated landscapes, to clarify the functioning of our downscaling algorithm and to consolidate the rational of our work.

This revision has led us to optimise our code (better memory management, more effective parallelisation, etc.). It is now much faster, taking less than 1 hour to generate the 42 million trees of our three landscapes, while it used to take "less than 5 hours" to generate the 35 million trees of the Bauges landscape. We updated the introduction accordingly.

Please find below our replies to your comments.

The article by Aussenac *et al.* describes a statistical procedure to generate a large data set of individual trees from airborne laser scanning (ALS) and inventory data. The variables include trunk diameter, tree height and species identity, and are provided across three European landscapes. The result is an impressive number of simulated/potential trees, which is a useful data set in forest ecology. As applications, the authors mention studies of scale and (more vaguely) forest management/ecosystem prediction, but one could easily think of a number of other concrete applications, such as input/validation of individual-based models of forest dynamics, or comparisons with automatically mapped tree crowns from airborne imagery, e.g. as in Weinstein *et al.* 2021[1], Ball *et al*. 2022[2], or spaceborne imagery, as in Tucker, Brandt, Hiernaux, *et al.* 2023[3].

I also found the paper generally well-written and with a well-thought through methodology for the mapping. The authors carefully tune their models to obtain optimal performance at every step and clearly have spent considerable amounts of time and effort to improve the prediction of stand attributes. In particular, I found the idea of matching predicted basal area to real stands and then filling in/removing trees until the basal area matches intriguing. This bears similarities with model-based estimations of forest attributes/tree attributes from lidar (Hurtt *et al.* 2004[4], Taubert *et al*. 2015[5], Rödig *et al.* 2017[6], Fischer *et al*.

2020[7]) and shares some of these models' advantages (e.g. more fine-scale distribution of biomass, no shrinking to the mean).

However, like these models, the authors' method also involves a lot of complex modelling steps, and it is in the validation step of the procedure that I see deficiencies that need to be addressed. I see two main issues:

a) the robustness of the models to extrapolation issues and spatial autocorrelation is not evaluated, so it is hard to assess how good the models are outside their calibration range and how much we can trust the predictions across the landscape.**The leave-one-out cross validation (LOOCV) performed on our entire workflow presented in this revised version of the article gives an indication on the robustness of our approach (see responses to the comments on model validation).**

Regarding lidar modelling, cross-validation (such as used in this study) remains a common method for accuracy assessment of landscape predictions, although some limitations have been highlighted in the literature. Accuracy assessment remains an important research topic, with no consensus on the best practices, as field data are usually scarce and modeling approaches complex (see paragraph 2.5 of the review in https://doi.org/10.1016/j.rse.2021.112477).

We acknowledge that spatial autocorrelation is necessarily present both in the lidar acquisition and in the forest structure. Hence spatial autocorrelation can be expected in the lidar model predictions and errors. Unfortunately, calibration plots were sampled on a regularly spaced grid in each study area which does not make it possible to quantify spatial autocorrelation at distances smaller than the grid step.

b) two of the key attributes of the data set (tree diameter and species identity) are not validated at all, despite featuring prominently in the title and in the results section (Figure 5). This should be a priority in a revised version. **In this revised version, we provide an in-depth evaluation of the generated landscapes (see responses to the specific comments below). We changed the abstract accordingly: "We carried out an in-depth evaluation of our workflow including, among other analyses, a leave-one-out cross validation. Overall, the landscapes we generated are in good agreement with the landscapes they aim to reproduce. In the most favourable conditions, the root mean square error (RMSE) of stand basal area (BA) and mean quadratic diameter (Dg) predictions were respectively 5.4 m^2 and 3.9 cm, and the generated main species corresponded to the observed main species in 76.2% of cases."**

We also modified the paragraph presenting the evaluation in the General Approach section: "We evaluated the overall reliability of our workflow, i.e. its ability to produce virtual landscapes as close as possible to the real ones (see Dataset validation). In particular we carried out a leave-one-out cross validation (LOOCV) on our entire workflow. This analysis consisted in: • comparing the observed and predicted values of BA, Dg, Bab and the quantiles of tree height and diameter; • comparing the observed and predicted values of species abundance at the landscape level; • calculating the frequency at which the most abundant species was correctly predicted at the cell

level; As a complement, we also compared the stands dominant heights measured by ALS (HdomALS) to those calculated from the trees we generated (HdomT). Finally, we compared the spatial distribution of species to current expert knowledge."

In the following I will provide a few comments on the article following roughly the overall structure, and give suggestions on how to improve the model validation.

Justification for the data set

I see the value of a fine-grain large-scale data set, and having such a data set is indeed rare, but it would be helpful to mention concrete applications. At the moment, the only justification given is the sentence: "Yet, this type of data could help address the scaling issues in ecology and could prove useful for testing forest management strategies and accurately predicting the dynamics of ecosystem services". This is the sentence from the abstract, but the same point is made at the end of the first paragraph. Could the authors rephrase and add literature references in the main text? The vast majority of data sets can be useful for the testing of forest management strategies or predicting dynamics of ecosystem services. What is unique to your data set? Why do we need detailed, tree-based data at large scales? **As suggested, we rephrased this part and added some references: "Yet, fine-grain descriptions of large forest areas could help address the pervasive scaling issues in forest ecology, modelling and management. In practice, such data could help better understand at which spatial scale ecological processes emerge in forest ecosystems [Craven et al., 2020; With, 2019]. They could also be extremely valuable to compare forest dynamics models operating at different scales (organ, tree, stand, landscapes) and evaluate their validity across scales [Papaik et al., 2010]. They could ultimately help develop and test management strategies at different spatial scales [Seidl et al., 2013]." Added references: Craven, Dylan, et al. "A cross-scale assessment of productivity–diversity relationships." Global Ecology and Biogeography 29.11 (2020): 1940-1955. Papaik, Michael J., et al. "Forest processes from stands to landscapes: exploring model forecast uncertainties using cross-scale model comparison." Canadian journal of forest research 40.12 (2010): 2345-2359. Seidl, Rupert, et al. "deepl" European Journal of Forest Research 132 (2013): 653-666.**

Model for mapping of tree attributes

ALS metrics: which metrics precisely did you use? **Candidate metrics for selection are those computed with the aba_metrics and std_tree_metrics functions from the lidaRtRee R package, as stated in section ALS mapping. Selected metrics result from an automated procedure applied separately on each study area, forest parameter, stratum (in case a specific model is calibrated for each stratum). The following sentence was added to the Stratification"subsection of the ALS mapping section: "The metrics selected in the 32 models for BA and Dg (which include at most six independent variables) are presented in Table S1 of the *Extended data*."**

The seven stratum-specific logistic regression models obtained for BA_b with stepaAIC variable selection include 20 to 30 metrics and are not presented.

Point cloud properties: Could the authors add information on/discussion of the sensitivity of their point metrics to scanner acquisitions? **Visual inspection of metrics maps was performed to exclude metrics exhibiting spatial patterns which might be linked to acquisition. We hypothesise that the selected metrics and models based on them are robust to variations in acquisition settings, as a systematic sampling of field plots should ensure that the heterogeneities of lidar acquisition and forest structure on the whole landscape are represented in the data. In the case of Milicz, several metrics seemed quite sensitive to points acquired with a large scan angle. Scan angle values above 21 degrees were excluded in order to achieve a trade-off between metrics robustness, point density and comprehensive coverage of the study area. In the case of Sneznik, some intensity-related metrics where also found to be influenced by acquisition pattern and where thus removed from the analysis ("imean", "imax", "isd", "iskew", "ikurt"). The following sentences were added to the manuscript: "Each metric map was visually checked for spatial patterns potentially linked to acquisition patterns, which eventually led to: - discard some intensity-related metrics in Sneznik study area; - remove ALS points acquired with a scan angle larger than 21 degrees in Milicz study area, in order to achieve a trade-off between metrics robustness, point density and comprehensive coverage of the study area."**

Lidar scans often exhibit considerable variation in pulse density even within a single acquisition (e.g. scan line centre vs. overlapping scan lines). What is each scan's standard deviation of point/pulse density? **The mean +- sd point density (/m2) for each study site, computed at 25 m resolution for pixels with at least 1 point above 2 m height is : Bauges (southern part) : 5.9 +- 3.1 Bauges (northern part) : 27.6 +- 13.3 Milicz : 16.5 +- 7.1 Sneznik : 18.4 +- 10.1 Those values are now indicated in the article. Mean values slightly differ from the previous ones which were computed for pixels of the whole landscapes.**

Could you include that as a variable in stratification? Could this improve your models (e.g. stratify by pulse densities between 5 and 10, 10 and 15, 15 and 20, etc., or even smaller step sizes)? **In order to obtain robust models and valid inferences, [https://doi.org/10.5589/m12-052] suggests to use at least 55 observations in calibration, which limits the number of strata that the whole dataset can be partitioned in. We hypothesised that a stratification based on forest structure defined by ancillary data would help in reducing the prediction error. For the Bauges study area where ALS data originate from two different ALS campaigns, campaign was also tested as stratification criterion. Several stratifications where tested, and the one retained for the final map production is the one with the best RMSE improvement in cross validation compared to the single model. Parcimony was also considered when selecting the model. We have not tested a stratification based on pulse densities, as most computed metrics are a priori robust to density and as we have checked that metrics maps did not display any pattern due to acquisition conditions, such as density variations.**

Descriptions: I appreciate that the paper is already quite dense, but quite a few steps in the methods section remain unclear to me, particularly in step 3. E.g., in the matching of BA and BAb, why do you need a correction value alpha? Can you explain the weighting better and

why it is divided by 16? Maybe this is more exhaustively explained in the Extended Data, but this needs to be clear from the main text already. **We have gone into more detail on points 3a and 3b. They now read as follows:**

- **"(a) For this, we first calculated *α*, a multiplier correction coefficient to be applied to all tree diameters of a field plot. The idea is to increase or decrease tree diameters so that their Dg reaches the Dg value of the cell to which they are associated. *α* is given by:"**
- **"(b) Thereafter, we calculated the weight (ω in $n.ha^{-1}$) of these trees with corrected diameters, so that the generated stand matches the BA and BA_b values of the cell to which it is associated. ω is given by:"**

As for the weighting, it is now explained (in a foot note and in point 3c), but see answer to Reviewer 2's comment on scale factors.

Model validation

As pointed out above, this is the point of the paper that needs to be more comprehensive. At the moment, the authors validate their approach by comparing dominant height, as obtained from lidar (mean height of six highest local maxima), to dominant height of the simulated stands, obtained via local allometries (mean height of six highest trees). It is definitely useful to do this comparison and good to see that the results are broadly consistent, so I would keep it in the paper. However, there are issues with circularity, as the authors first use a number of lidar metrics that involve height / basal area-to-height relationships to create the maps and then compare the inferred results (+ independently derived height allometries) again to lidar-derived height metrics. Furthermore, height of the dominant trees may be related to basal area, but it cannot be used to evaluate basal area/tree diameter predictions as such, nor does it validate predicted species composition - both are key features of the data set. **We agree with the reviewer, the comparison of Hdom only partially validates our virtual landscapes. The revised version of the article includes a much more comprehensive evaluation of the virtual landscapes.**

We added this point on circularity in the section presenting the comparation of Hdom as follows: "There is some circularity in comparing Hdom ALS and Hdom T as models predicting BA, Dg and BA b from ALS point clouds may include ALS derived height metrics or more generally metrics which are correlated with the dominant height estimated from ALS point clouds. The results of this comparison must therefore be interpreted with caution."

Given that the author's simulation approach seems fast (only ca. 5 hours on a modern laptop, amazing!), another approach suggests itself, namely within-site cross-validation, ideally in the form proposed by Ploton *et al.* 2020[8]. Since a spatially explicit leave-one-out cross-validation, as suggested in Ploton *et al.* 2020[8], may be too computationally intensive, I would recommend the simpler approach proposed in the same paper: for each of the European landscapes, I would recommend the authors to split their field data sets into, e.g., 5 spatially aggregated folds (i.e., spatial clusters), and run their model 5 times, each times using 4 folds to train the model and 1 separate geographic fold of plots to validate the model. In this 1 fold, the authors could directly compare predictions of tree values to actual data according to some simple standard metrics (total basal area, mean quadratic diameter,

95th percentile of diameter, percentage of species xyz, 95th percentile of height, mean height, dominant height). For comparison and to broadly assess whether spatial autocorrelation makes a difference, the authors could do the same validation procedure also with 5 folds containing plots randomly distributed in space (so no spatial clusters). This would only take 25 hours for each validation and give a good impression of how easy it is to accurately map individual trees and species at landscape scale and how realistic the produced inventories are. It would likely also increase interest in the data set, as it would give potential users higher confidence in the results. **We carried out a leave-one-out cross-validation (LOOCV). This validation is actually not too computationally intensive because it has to run only on a few hundred field plots and not on the whole landscape. We also preferred this LOOCV to a k-fold cross-validation because the latter would require to check that sample size in each stratum would remain relevant in each calibration fold, while performing repetitions of the k-fold samples to assess the reliability of the statistics. Besides we are not certain that spatially correlated folds will help in understanding the robustness of the map. One can expect a model calibrated on a spatially-selected subset to perform poorly on another spatial subset as the predicted area may be different from the calibration area both for the lidar and forest. Spatial k-folds do not allow to assess the local spatial variations of performance of a model calibrated with systematic samples (which is what our models are) in a better way than what a LOOCV does. In both cases, spatial correlation at distances smaller than the grid step, which is interesting e.g. for small area estimations, is impossible to evaluate.**

We added the following paragraph to the method subsection of the validation section: "We carried out a leave-one-out cross validation (LOOCV) to evaluate the realism of the virtual landscapes we generated. This consisted in excluding a field plot from our entire workflow and comparing the predicted values obtained to the observed values. This operation was repeated within each landscape for all field plots. We calculated the root mean square error (RMSE) of the predictions of BA, Dg, Bab and the quantiles of tree height and diameter. As part of the LOOCV, we also compared the observed and predicted values of species abundance at the landscape level (in BA) and calculated the frequency at which the most abundant species was correctly predicted at the cell level."

We presented the results of this new evaluation procedure in the results subsection of the validation section: "Overall, the virtual landscapes are in good agreement with the landscapes they aim to reproduce. The generated stand structures and compositions are consistent with the observations and make it possible to distinguish stands at different stages of development and with different compositions.

At Milicz, predictions are the most accurate. The RMSE of all evaluated variables are the lowest in comparison with the other landscapes (Table T1, Figure 4). Species abundance at the landscape level is also better reproduced (Figure F1). Finally, in 76.2% of cases, the generated main species corresponds to the observed main species. This higher quality of predictions can be explained by the fact that Milicz has the highest density of inventory plots and the least complex landscape, with a predominance of even-aged monospecific stands and the lowest species diversity

among our three landscapes.

At the Bauges and Sneznik, the RMSE of the evaluated variables are comparable (Table T1., Figure 4). In contrast, predictions of species abundance at the landscape level are more accurate at Sneznik (Figure F1). The same applies to the compositions predicted at the plot level: the predicted main species corresponds to the observed main species in 63.1% of cases at Sneznik and in 37.2% of cases at the Bauges. However, two datasets were used in the Bauges. In the local forest inventory (LFI) not all trees were identified at the species level and trees with a dbh between 7.5 and 17.5 cm were not measured but counted by diameter classes and grouped in two categories (coniferous and broadleaf). This led us to use a local subset of the NFI from which composition is derived in our downscaling algorithm. The poorer composition predictions in the Bauges might therefore partly be an artefact arising from the evaluation itself, as the LFI may not be suitable to serve as a field reference.

The fact that the RMSE values obtained from the LOOCV carried out on our entire workflow are almost similar to the RMSE values obtained from the LOOCV of ALS models shows that the downscaling algorithm hardly adds any error (Table 1, Table T1). The main way of increasing the realism of our virtual landscapes would therefore be to improve the ALS models."

Finally, we added the RMSE values associated to the three scatter plots of our Figure 4 for greater consistency in the document and modified the paragraph where this figure is presented: "With R^2 values ranging from 0.61 to 0.83 (Figure 4) and RMSE values below 5 m, $Hdom_{ALS}$ and $Hdom_{T}$ are consistent with one another. This provides a general validation of our workflow. As discussed above, the better predictions obtained at Milicz might steam from the higher density of inventory plots and the lower complexity of the landscape. At Sneznik, $Hdom_{T}$ tends to be overestimated as $Hdom_{ALS}$ decreases. This divergence could be due to the ice storm that occurred between the field inventory and the ALS acquisition and that might have biased the ALS models."

Since the paper puts its focus on the value of individual trees, there should, in my opinion, also be one result/validation graph that shows individual trees in some way. It could be, for example, a zoomed-in image of lidar-derived canopy height models + a predicted distribution of trees. If the 5-fold cross-validation is carried out, as above, the authors could simply show sample lidar canopy height models on top of plots, and the diameter distributions for the simulated and the inferred plots. **Here, we do not exactly agree with the reviewer. We do not generate individual trees, per se, but rather a list of trees for each cell, and those generated trees are not spatialised within the cells. For such a purpose different methods and denser ALS data would be required. Also, it is not clear to us how the comparison of canopy height models and diameter distributions would constitute an evaluation of our algorithm. The comparison of diameter and height quantiles carried out in response to the previous comment seems to us a better way to evaluate the quality of the generated tree lists.**

Overall, it would also be interesting to readers to understand in how far the predicted

species distributions reflect current expert knowledge, but this is not a necessity. **In the method subsection of the validation section, we added: "Finally, we examined the spatial distribution of species at each site and compared it to current expert knowledge."**

And in the results subsection of the validation section, we added: "Overall, species spatial distribution in the virtual landscapes is consistent with field observations. In the Bauges, pure and mixed stands of fir and spruce are more abundant at higher elevation while mixed stands of broadleaf species are found at lower elevation. At Milicz, pure stands of Scots pine are found at lower elevation while broadleaf species and mixed stands appear at higher elevation. Finally, at Sneznik, pure beech stands are found at higher elevation while fir is found at lower elevation in pure or mixed stands (a specific feature of the site)."

Data set

I had a quick look at the data set. One variable I did not understand was the variable "n" or "number of trees". Could you explain it a bit better? Does this mean that the specific diameter exists n times in the specific data set? If this is true (and only in this case), I seem to get some cells (very few) of 25m by 25m (e.g. cellID25 = 2439821 in the "Bauges" data set) that contain more than 500 trees with dbh >= 9-10cm per 625m2 and a total basal area >= 6m2 (which would yield roughly 100m^2 per hectare, at densities of 8000 trees). These are outliers, and every model is allowed to have outliers (and nature is full of them too), but it would be interesting to get your take on that in terms of realism/stand type. It could also be part of the validation to assess the edges of the basal area distribution or to give readers a hint what to make of the most extreme values. **The variable "n" indeed means that a specific diameter exists n times. We clarified that in the presentation of our dataset on the zenodo website and in the Extended data section as follows: "- n: number of trees. n is an integer >= 1, meaning that a specific set of species "sp", diameter "dbh" and height "h" can be present multiple times in a cell."**

The extreme values you identified are indeed outliers. They are a direct consequence of the uncertainties associated with the models. The realism of the stands associated with these extreme values is, of course, open to question. However, it seems difficult to define thresholds separating realistic from unrealistic stands as extreme values can be locally observed. We believe it is up to the users of the dataset to decide whether or not to use all the stands depending on their objectives. We drew the reader's attention to the outliers at the end of the validation section as follows: "Our procedure is not free of flaws and some outliers are present in the generated data (i.e. stands with extreme values of BA, Dg, tree height or density). These outliers are a direct consequence of the uncertainties associated with the models we used. The realism of the stands associated with these extreme values is open to question. However, separating realistic from unrealistic stands seems difficult as extreme values can be locally observed. It is therefore up to the users of the dataset to decide whether or not to consider these stands depending on their objectives."

Finally, it is always tricky to assess the relevance of extreme values as the field data

itself does not allow to estimate them accurately. We believe such an analysis could hardly help in evaluating our approach and we would rather not include it as part of the validation.

Competing Interests: No competing interests were disclosed.