DATA NOTE

Diameter, height and species of 42 million trees in three European landscapes generated from field data and airborne laser scanning data [version 1; peer review: 2 approved with reservations]

Raphaël Aussenac[1-3], Jean-Matthieu Monnet[1], Matija 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]CIRAD, UPR Forêts et Sociétés, Yamoussoukro, Cote d'Ivoire
[3]Forêts et Sociétés, Université de Montpellier, CIRAD, Montpellier, France
[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

v1 **First published:** 14 Feb 2023, **3**:32
https://doi.org/10.12688/openreseurope.15373.1
Latest published: 14 Feb 2023, **3**:32
https://doi.org/10.12688/openreseurope.15373.1

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. 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 available data: inventory and ALS data.
We evaluated the overall reliability of our workflow by comparing the stands dominant heights measured by ALS to those calculated from

Open Peer Review

Approval Status ? ?

	1	2
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.

the trees we generated. Overall, the landscapes we generated are in good agreement with the landscapes they aim to reproduce.

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; **Matija 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, Matija 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 1; peer review: 2 approved with reservations]** Open Research Europe 2023, **3**:32 https://doi.org/10.12688/openreseurope.15373.1

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

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 better understand at which spatial scale ecological processes have an effect and thus help address the scaling issues in ecology [With, 2019]. Such data could also prove useful for testing forest management strategies and accurately predicting the evolution of ecosystem services.

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: less than 5 hours on a six-core laptop are needed to generate the 35 million trees making up the 51,500 ha of forest in the Bauges Geopark.

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

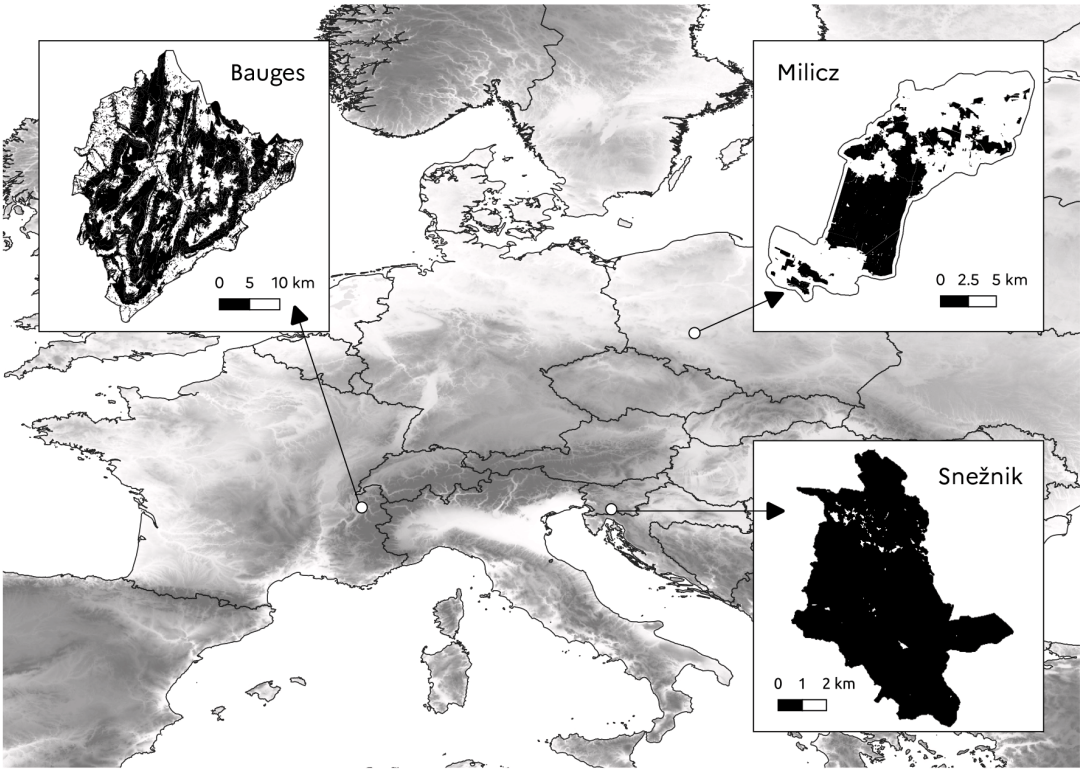

Figure 1. Location of study areas. The black areas show the forested areas.

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. 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. 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, by comparing the stands dominant heights measured by ALS ($Hdom_{ALS}$) to those calculated from the trees we generated ($Hdom_T$; see *Dataset validation*).

ALS mapping

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

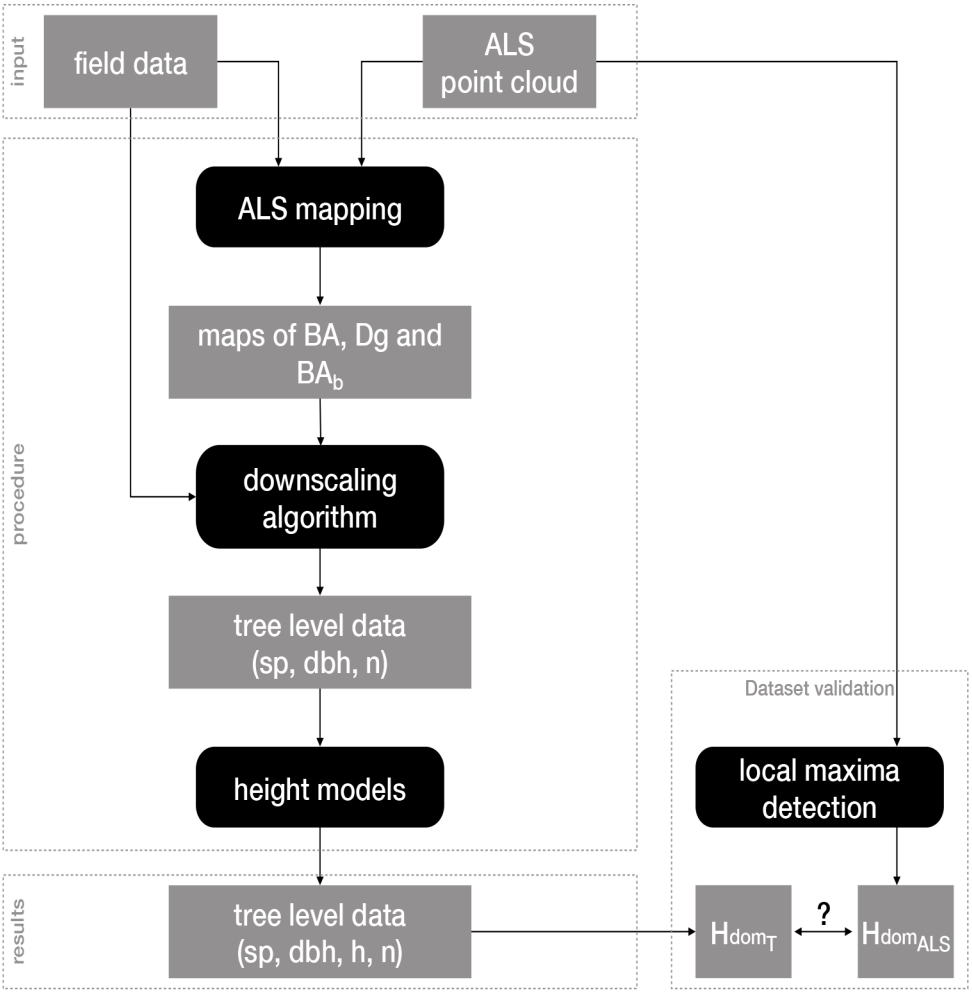

Figure 2. Workflow overview. Black boxes correspond to data generation steps feeding each other with datasets represented by grey boxes. BA: basal area; Dg: mean quadratic diameter; BA_b: BA proportion of broadleaf trees; sp: species; dbh: diameter at breast height; h: height; n: number of trees; $Hdom_{ALS}$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

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. Mean point densities were respectively 5.5 and 24.4 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 average point density was 16.8 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 average point density was 18.1 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 or from the derived CHM, 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 tree_segmentation function) from the CHM. 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 25x25 m^2 cell of the raster layout covering each acquisition, in order to build the prediction dataset (mapping step).

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).

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.

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 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 m^2 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 m^2 cell a field plot based on the similarity of their BA, Dg and BA_b.
   (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.

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

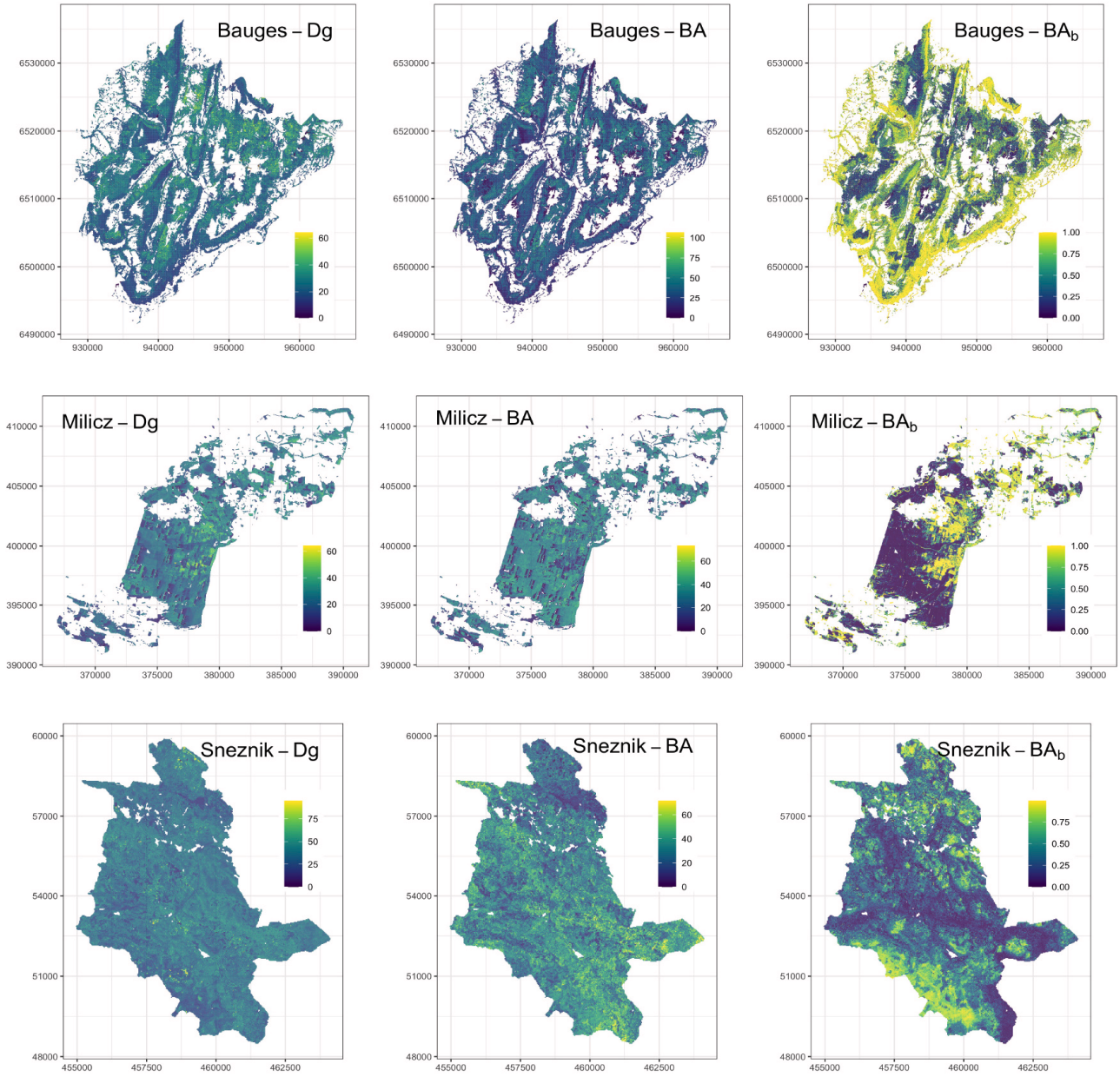

Figure 3. Airborne laser scanning (ALS) maps of forest variables for our three study areas at a 25 m resolution. Dg: mean quadratic diameter (cm), BA: basal area ($m^2.ha^{-1}$) and (BA_b): proportion of broadleaf BA.

(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 m^2 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 apply to all trees dbh of the field plots. 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 per ha of each tree as follows:

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

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, $Prop_{S_{pF}}$ is the BA proportion of species S_p in broadleaf (reps. coniferous) species in the field plot, and $Prop_{tree_F}$ is the BA proportion of this tree in species S_p in the field plot.

(c) Finally, we divided ω by 16 and performed a Bernouilli draw on the decimal part of the obtained values (as an example, a weight of 5.63 has a 63% chance of becoming 6, and a 37% chance of becoming 5) to get integer values corresponding to the weight of the trees in the 25×25 m² cells. As this rounding of the weights slightly modifies the stand total BA, we transformed again the trees dbh to reach the total BA provided by the ALS mapping using the trees BA and their integer weight (ω_{int}) as follows:

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

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.

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.

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*.

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

To assess the realism of the virtual landscapes we generated, 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

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

[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. As shown in Figure 2 $Hdom_{ALS}$ is totally independent from the procedure that generates the trees. Thus, comparing $Hdom_{ALS}$ and $Hdom_T$ makes it possible to evaluate the overall reliability of our workflow.

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.

Results

Overall, with R^2 values ranging from 0.61 to 0.83 (Figure 4), $Hdom_{ALS}$ and $Hdom_T$ were consistent with each other. This indicates that the virtual landscapes are in good agreement with the landscapes they aim to reproduce. However, $Hdom_{ALS}$ and $Hdom_T$ showed some divergence at Snežnik: $Hdom_T$ tends to be overestimated as $Hdom_{ALS}$ decreases. This 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.

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.

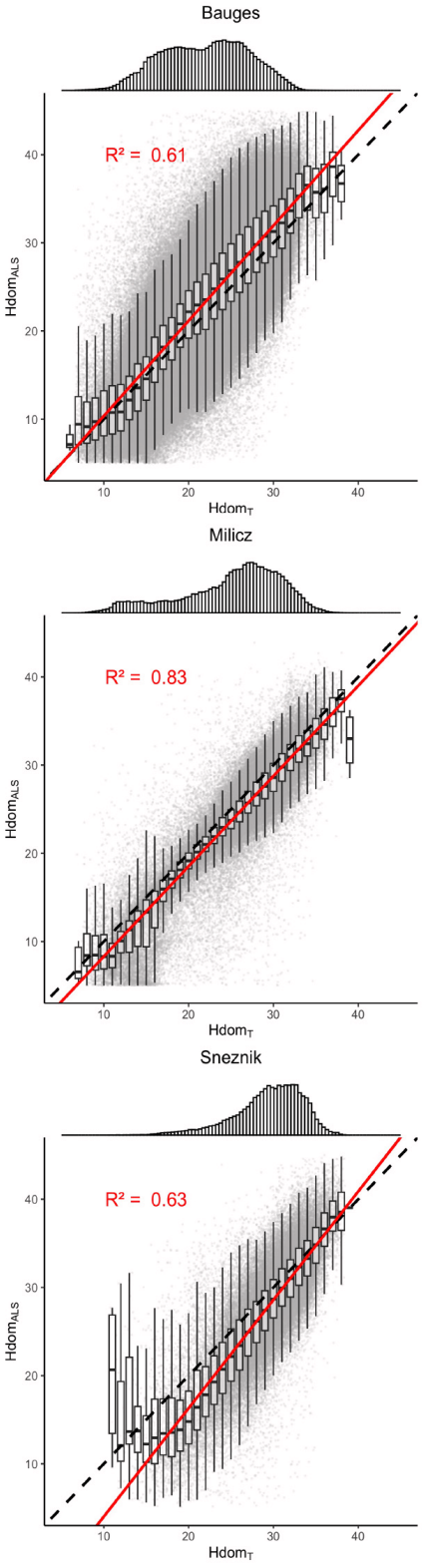

Figure 4. Comparison of the stands dominant heights measured by ALS ($Hdom_{ALS}$) to those calculated from the generated trees ($Hdom_T$). The top panels show the distribution of $Hdom_T$. The dashed lines indicate the $y = x$ line. The red lines correspond to the regression lines. The regression R-Squared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

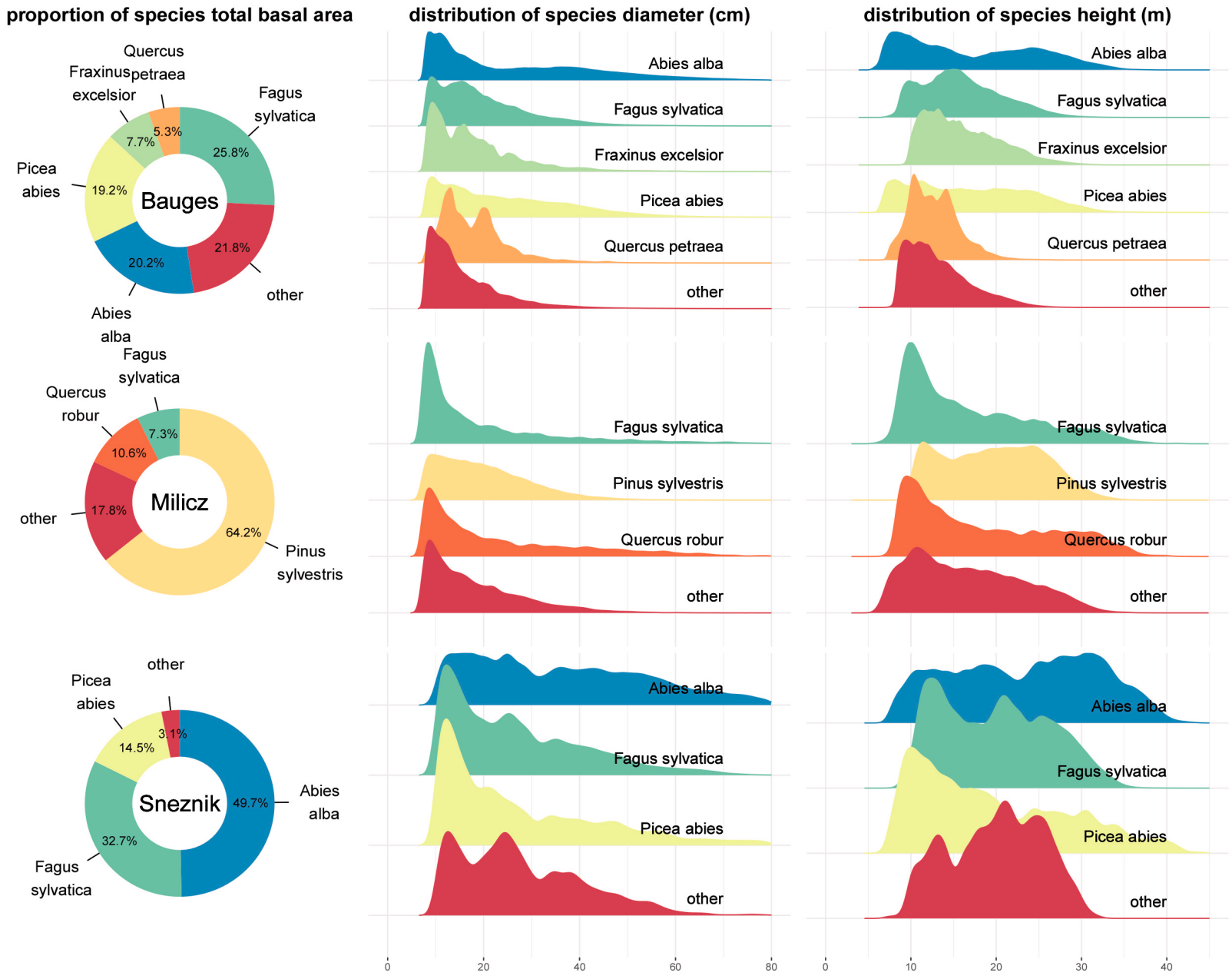

Figure 5. Main species basal area proportion, diameter distribution and height distribution in the three virtual landscapes. Species accounting for less than 5% of the virtual landscapes total basal area were grouped in the 'other' category.

Open Research Europe 2023, 3:32 Last updated: 12 MAY 2023

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
- 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.

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**

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**

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 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.

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.