Forestry: An International Journal of Forest Research, 2026, **99**, 1–17
https://doi.org/10.1093/forestry/cpaf062
Advance access publication date 14 October 2025
Dataset paper

SegmentedForests: a labelled dataset of terrestrial LiDAR point clouds for semantic segmentation of forests

Diego Laino[1], Carlos Cabo[2,*], Celestino Ordóñez[2], Rodolfo Bolanos[1], Romain Janvier[3], Federico Giulioni[4], Miriam Herrmann[5], Andrew Hudak[6], Russell Parsons[7], Cristina Santin[1,8]

[1]Biodiversity Research Institute (IMIB), Spanish National Research Council (CSIC)—University of Oviedo-Principality of Asturias, Gonzalo Gutierrez Quiros Street, Mieres, Asturias 33600, Spain
[2]Department of Mining Exploitation and Prospecting, University of Oviedo, Gonzalo Gutierrez Quiros Street, Mieres, Asturias 33600, Spain
[3]Independent Consultant, Nancy 54600, France
[4]Department of Agricultural, Food and Environmental Sciences, Università Politecnica delle Marche, 10 Brecce Bianche Street, Ancona 60131, Italy
[5]Department of Remote Sensing and Geoinformation, Institute of Geographic Sciences, Freie Universität Berlin, Malteserstr. 74 -100, Berlin 12249, Germany
[6]United States Department of Agriculture Forest Service, Rocky Mountain Research Station, Forestry Sciences Laboratory, 1221 South Main Street, Moscow, ID 83843, United States
[7]United States Department of Agriculture Forest Service, Rocky Mountain Research Station, Fire Sciences Laboratory, 5775 W Broadway Street, Missoula, MT 59808, United States
[8]Centre for Wildfire Research, Swansea University, Wallace Building, Singleton Campus, Swansea SA2 8PP, Wales, United Kingdom

[*]Corresponding author. Department of Mining Exploitation and Prospecting, University of Oviedo, Mieres, 33600, Asturias, Spain. E-mail: carloscabo@uniovi.es

Abstract

Semantic segmentation of point clouds using deep learning (DL) has been the subject of research in forestry in recent years due to its potential applications. Several scientific and management disciplines, such as biodiversity monitoring, ecosystem carbon assessments, or forest management could benefit from this technique. However, it requires manual segmentation of point clouds to be used as training data. This process is highly labour-intensive and time-consuming, and there is a notable lack of publicly available datasets to support the development of accurate DL semantic segmentation models for forestry and forest ecology applications. Here, we present SegmentedForests, a curated dataset of manually segmented ground-based point clouds from forest plots, specifically designed to facilitate the training and validation of semantic segmentation models. This publicly available dataset contains >920 million labelled points from 14 forest plots, acquired using both terrestrial laser scanning (TLS) and mobile laser scanning (MLS) technologies. It covers two hectares of broadleaf, conifer, and mixed stands from different bioclimatic regions and features >1600 trees across 16 tree species. Each point cloud is labelled into multiple vegetation classes (up to 16), such as tree stems, branches, grass, shrubs, and down wood, as well as non-vegetation elements commonly present in forest scenes, including rocks, people, and stakes. Data splits to facilitate DL model development using our dataset are provided as well. The dataset is available at https://zenodo.org/records/17396681. By releasing this annotated dataset, we seek to address the critical need for publicly available, high-quality training data for DL models that perform semantic segmentation of ground-based point clouds in forest ecosystems.

Keywords: forestry; deep learning; artificial intelligence; TLS; supervised segmentation; 3D computer vision

Introduction

Characterizing forest structure is fundamental to a wide range of forestry and ecology applications, including forest monitoring and management planning (McElhinny et al. 2005), biodiversity conservation (McElhinny et al. 2005, Bütler et al. 2013), carbon cycle modelling (Luyssaert et al. 2007, Disney et al. 2018) and wildlife habitat assessment (Ehbrecht et al. 2017, Rehush et al. 2018). Forest structure encompasses diverse attributes such as tree size, shape, and density, canopy cover or understorey composition.

In recent years, ground-based laser scanning technologies, such as terrestrial laser scanning (TLS) and mobile laser scanning (MLS), have emerged as transformative tools for generating 3D models of physical objects. These technologies generate highly detailed 3D point clouds that capture the spatial distribution of elements in space with accuracy (Calders et al. 2020). In the context of forestry, the increasing level of geometric detail from these ground-based point clouds has facilitated the development of forest 'digital twins': virtual representations of forests that enable in-depth analysis of their spatial structure (Qiu et al. 2023). Ground-based 3D point clouds have therefore been extensively examined for their suitability to measure key forest attributes such as tree position, diameter at breast height (DBH) and tree height (TH) (Cabo et al. 2018; Fassnacht et al. 2023; Laino et al. 2024). This could be seen as a digitalization of traditional forest inventory. However, in addition to basic forest inventory parameters, ground-based point clouds enable derivation of more quantitative descriptors of forest structure (Disney et al. 2018, Fassnacht et al. 2023). For instance, algorithms utilizing these datasets can estimate timber volume (Puletti et al. 2019, Hyyppä et al. 2020, Alvites et al. 2021, Prendes Pérez et al. 2021), analyze canopy characteristics (Chianucci et al. 2020, Schraik et al. 2021), or assess understorey composition (Alonso-Rego et al. 2020, Tian et al. 2023). Thus, ground-based point clouds

Handling editor: Dr. Rubăn Manso
Received 25 March 2025; **revised** 18 August 2025; **accepted** 16 September 2025

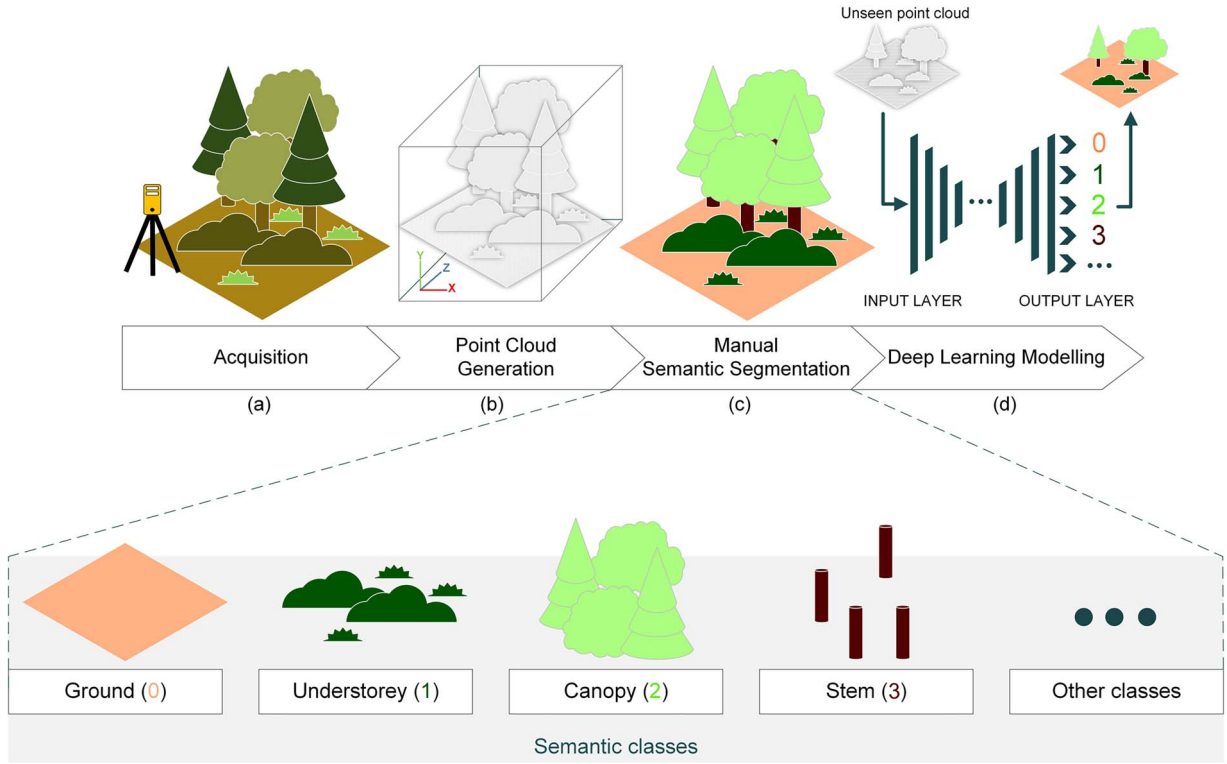

Figure 1. Schematic overview of a generic workflow to build DL semantic segmentation models for ground-based point clouds of forest plots. (a) The process starts by scanning the forest plots in the field. (b) Then, the scans are processed and point clouds generated. This may requireregistering processing (i.e. cropping, de-noising, subsampling) and sometimes post-processing (i.e. voxelization). (c) Once the point clouds have been generated, they are manually segmented. An example of semantic classes is illustrated. (d) Finally, the segmented point clouds can be used to build DL models that segment forest scenes. A simplified DL model is shown (top right), which assigns semantic labels to new, unseen point clouds after being trained with the manually segmented point clouds.

enable characterizing forest stand structure with unprecedented detail and accuracy, surpassing the limitations of traditional field-based methods and airborne laser scanning (ALS), used in several nation-wide inventories (Fassnacht et al. 2023). Whist ALS offers broader spatial coverage, its sampling density and top-down perspective limit their ability to resolve stem form or fine-scale structural features near the ground, such as understorey vegetation (White et al. 2016). In contrast, TLS and MLS provide detailed, close-range data ideal for developing high-resolution forest digital twins and supporting tasks that require precise spatial differentiation between structural components.

In this context, semantic segmentation has emerged as a key tool for unlocking the full potential of ground-based point clouds (Lines et al. 2022a). Semantic segmentation assigns a class to each point. In the context of forestry, this means labelling each and every point as ground, stem, branch, or understorey vegetation, for instance (illustrated in Fig. 1). This process enables the decomposition of complex forest scenes into specific and interpretable structural components, facilitating downstream analyses (Lines et al. 2022a). However, traditional segmentation methods often rely on data (Breiman 2001, Belton et al. 2013, Béland et al. 2014) or rule-based (Tao et al. 2015, Ma et al. 2016) modelling approaches, which can struggle with the compositional complexity and structural variability of natural forests (Xi et al. 2020).

To address the challenges of segmenting complex structures, artificial intelligence (AI) techniques, particularly those in the realm of machine learning (ML), have gained prominence in 3D point cloud analysis. Supervised deep learning (DL), a subfield of ML, has revolutionized segmentation tasks with the application of modern neural network architectures (Guo et al. 2021; Halperin and Eisl 2025). Modern neural networks can learn hierarchical representations directly from raw point cloud data (Qi et al. 2017, Guo et al. 2021, Engel et al. 2021, Zhao et al. 2021, Robert et al. 2023), enabling models to identify intricate patterns and relationships within these datasets that traditional methods might overlook (Kulicki et al. 2024). Although point cloud segmentation has a long-standing history in Computer Vision and Robotics (Nguyen and Le 2013), its application to forest-related research is relatively recent. Thus, semantic segmentation of forest point clouds via DL has the potential to automate and dramatically enhance forestry and forest ecology applications (Kaijaluoto et al. 2022, Lines et al. 2022b). However, the effectiveness of DL methods depends on the availability of large, high-quality, annotated datasets. Despite the potential usefulness of semantic segmentation of point clouds in forest-related applications is immense, there are however very few publicly available datasets designed for this purpose (Lines et al. 2022a, Kulicki et al. 2024).

Existing datasets mostly focus on instance segmentation (i.e. individual tree segmentation). In opposition to semantic segmentation, instance segmentation not only classifies points but also differentiates between individual objects within the same class (for example, separating each tree from its neighbours) (Guo et al. 2021). However, these datasets primarily targeted tree-level metrics, overlooking the complex spatial context of full forest scenes. For example, Calders (2014) provides a dataset of 31 individual trees from a TLS point cloud of a native eucalypt forest in Victoria, Australia. Similarly, Weiser et al. (2022) includes 249 individual trees from 12 forest plots scanned using TLS, uncrewed aerial

vehicle laser scanning (ULS), and ALS technologies, whist Tockner et al. (2022) offers a dataset of 515 individual tree point clouds acquired using MLS from Austrian forests. However, none of these datasets include semantic labels for the points, limiting their utility for semantic segmentation. An exception is the recently published FOR-Instance dataset (Puliti et al. 2023), which represents a significant step forward in providing semantically labelled point cloud data. This dataset includes ULS point clouds comprising 2.8 ha of forest across 29 plots in five countries: Norway (20 plots), the Czech Republic (3 plots), Austria (1 plot), Australia (1 plot), and New Zealand (5 plots), where individual points have been segmented into six semantic classes: stem, woody branches, live branches, terrain, low vegetation, and bordering points. Whilst this dataset expands the scope of semantic segmentation possibilities in forestry, it is important to note that it is not ground-based data but rather acquired through ULS technology. This distinction may affect its direct applicability to certain ground-level forestry applications, such as those requiring high-density TLS point clouds for detailed structural analysis of below-canopy components.

In contrast, to the extent of the authors' knowledge, only one publicly available dataset specifically addresses semantic segmentation of ground-based point clouds from forest environments: Cheng et al. (2024). This dataset includes MLS and ULS data from diverse environments, including a 3-hectare (ha) forest plot in New Jersey, USA, an intensively managed 0.2 ha forest plot in Virginia, USA, and 26 ha of pistachio orchards and 20 ha of almond orchards in California, USA. However, the semantic labels in this dataset are minimal, with stems being the only vegetation component given a dedicated label. All other points are classified as 'ground' or 'miscellaneous', which limits the applicability of the dataset for developing and evaluating models that require a more detailed understanding of forest structure.

A more recent study by Liang et al. (2024) describes a dataset comprising six TLS point clouds manually segmented into semantic classes such as 'ground', 'trunk', 'first-order branch', 'higher-order branch', 'foliage', and 'miscellany'. Whilst this dataset could potentially advance semantic segmentation research, it is not publicly available, further underscoring the scarcity of accessible, high-quality annotated datasets for this purpose.

The lack of public benchmarking datasets is a major obstacle for the progress of semantic segmentation in forestry. Many studies rely on custom datasets to demonstrate specific applications of existing or adapted DL methods for semantic segmentation (Xi et al. 2020; Krisanski et al. 2021; Shen et al. 2022; Xi et al. 2023; Oviedo de la Fuente et al. 2024). Whilst this approach can yield useful insights, it involves intensive and time-consuming data collection and processing, and the results are frequently restricted to specific ecosystems and sensor types. Moreover, the variability in the types and structures of data used in these studies makes it difficult to compare methods or generalize findings across different forest types or environmental conditions. (Lines et al. 2022a, Ma et al. 2023, Van den Broeck et al. 2023). As a result, the field urgently requires datasets that are publicly accessible, cover diverse forest ecosystems, and include detailed semantic labels for a wide range of forest components.

To mitigate the scarcity of datasets described above, our main objective is to provide a public dataset for development of semantic segmentation models based on DL. For this, we introduce here the SegmentedForests dataset and highlight its potential as a benchmarking resource for developing DL-models for semantic segmentation of ground-based point clouds of forest scenes.

Materials and methods

The SegmentedForests dataset includes 14 ground-based point clouds from 14 forest plots (Plot 1–Plot 14) distributed across four countries: Spain (five plots), Austria (five plots), United Kingdom (one plot), and the United States (three plots) (Fig. 2). The plots vary in size, covering areas from 650 m^2 to 3600 m^2. The plots were selected to represent a wide range of forest types, structures, and ecological conditions across different geographic regions, rather than to be representative of any single forest type or management regime. The dataset aims to capture variability relevant to ecological monitoring and structural characterization, rather than being tailored specifically for commercial forestry applications. Site selection was constrained by availability of high-quality ground-based scans, manual annotations, and permission for data sharing. In total, the dataset spans over 2.3 ha of forests that feature >1600 trees from 16 different species.

Plot characteristics

The dataset includes plots from Mediterranean, Oroborealoid, Oceanic Temperate, and Continental Temperate bioclimatic regions and incorporates both broadleaf (i.e. trees with broad, flat leaves) and coniferous forests (i.e. composed of trees with needle-like leaves and cones). These include both natural forests and plantations, offering a broad spectrum of forest ecosystems for analysis. The dataset spans a range of forest ages, from old-growth forests with complex structures to young stands, as well as different acquisition periods through the year, which affects the phenological state of the vegetation. Additionally, some plots feature dense understorey vegetation, including shrubs and small trees, whist others have a more open understorey, providing a range of vertical structural complexity. The dataset also includes forests with different slopes and varying levels of disturbances, such as areas with fallen trees, down wood, gaps, and other structural heterogeneities. Table 1 and Table 2 contain qualitative and quantitative descriptions of the plots, respectively, and Fig. 3 shows a Principal Component Analysis (PCA) biplot of the quantitative descriptors.

Whilst thoughtful care was taken to ensure structural and ecological diversity, it is unfeasible in practice to produce a dataset that is completely balanced across all possible forest types or structural configurations. As such, users using SegmentedForests for training DL models should be aware of potential biases. In addition to the qualitative and quantitative descriptions above, class distributions per semantic category are reported later, in Table 3.

Plots 1–5 (Spain)

The five plots located in Spain (Fig. 2, Table 1) represent diverse forest types and species, including both natural forests and plantations across distinct climatic regions within the country. Plot 1 is situated in the Cantabrian Mountains (Cordillera Cantábrica) in Cantabria, northern Spain. The Cantabrian Mountains fall within the Oroborealoid Bioclimatic Region as defined by Walter's classification (Walter 1979). The oroborealoid climate is representative high-altitude zones in temperate latitudes, where the influence of altitude creates cool, humid conditions that resemble boreal climates. These regions are typified by cold winters, mild summers, and high annual precipitation, often >1200 mm, with a marked influence of Atlantic weather systems. Plots 2, 3, and 4 are situated north of the Cantabrian Mountains, in the region between this mountain range and the Cantabrian Sea, which falls under the Nemoral Bioclimatic Region (Walter

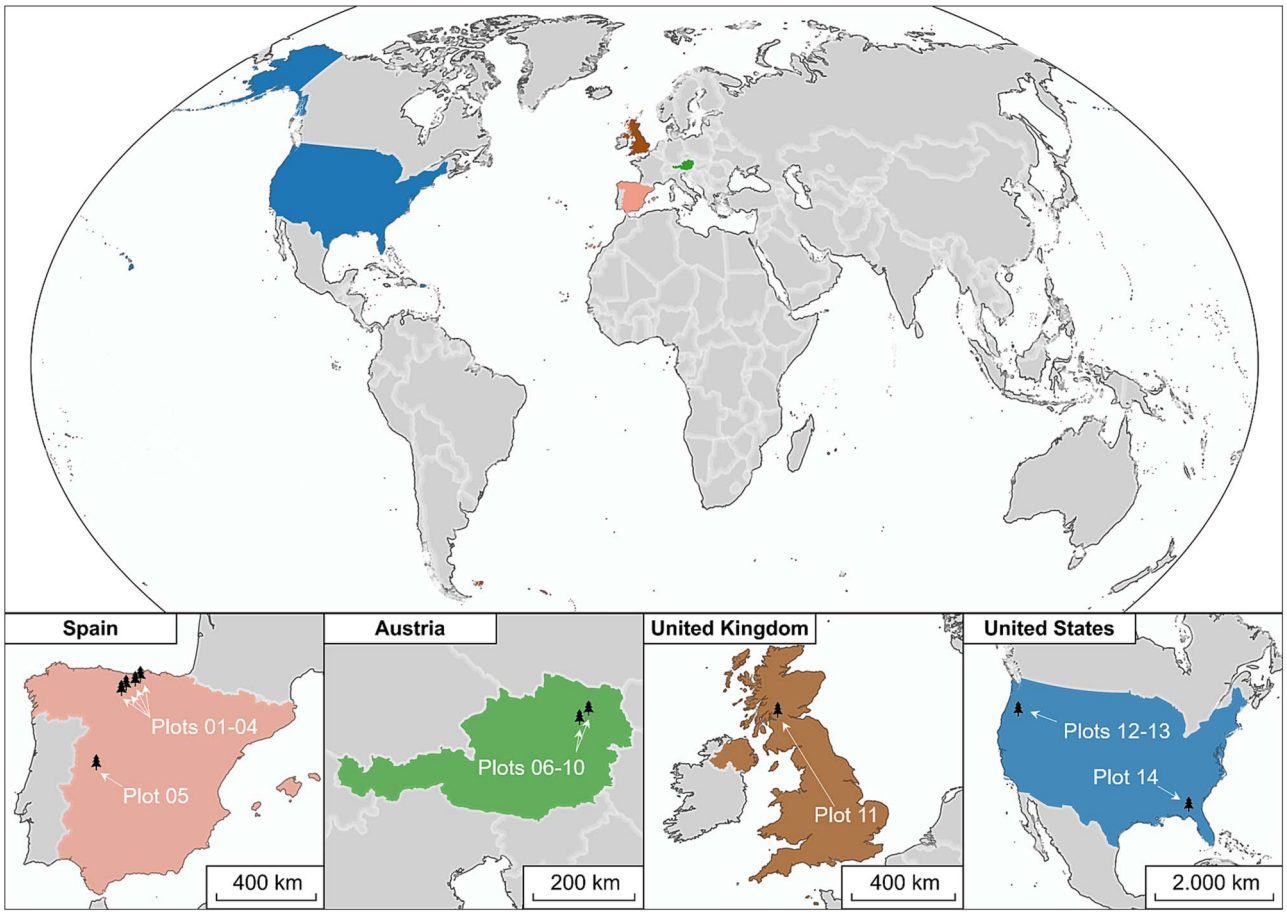

Figure 2. Locations of the 14 forest plots included in the SegmentedForests dataset. Plots 1–5 are located in different regions in Spain, plots 6–10 are located in the Vienna woods in Austria, plot 11 is located in Loch Lomond in United Kingdom and plots 12–14 are located in the southeastern and northwestern regions in the United States.

1979). This region is characterized by mild temperatures and high annual precipitation, promoting dense and productive forest growth. Plot 5 is located in Cáceres, within the Sistema Central mountain range, and falls under the Mediterranean Bioclimatic Region (Walter 1979). This region is characterized by hotter summers, colder winters, and lower annual precipitation compared to coastal or Northern Spain, reflecting the typical conditions of Mediterranean climate.

Plot 1 covers an area of 2100 m^2 and represents a typical Northern Spanish montane forest ecosystem (Fig. 4). The plot is dominated by Pyrenean oak (*Quercus pyrenaica*), a deciduous oak species adapted to the cool and humid conditions of mountainous environments in the western and northern Iberian Peninsula. It features medium-sized individuals forming a dense canopy, interspersed with younger trees and sapling and abundant understorey vegetation. It has the highest tree density across the dataset (Table 2).

Plot 2 spans 1600 m^2 and corresponds to a traditionally managed stand of European beech (*Fagus sylvatica*) (Fig. 3). The trees in this plot exhibit signs of historical pollarding, a traditional management practice where branches are periodically cut at a height out of reach of grazing livestock. This has resulted in a characteristic growth pattern with thick trunks and large, sprawling branches that form a dense and continuous canopy that allows limited light penetration to the forest floor. Due to active livestock grazing, the plot features a grassy ground cover and a virtually absent understorey, as the herbivory prevents the establishment and regeneration of shrubs and saplings. Some of these trees have the highest DBH values across the whole dataset (Table 2), reaching diameters of up to 90 cm.

Plot 3 covers 1600 m^2 and represents a typical radiata pine (*Pinus radiata*) plantation (Fig. 4). Globally, there are just over 4 million ha of radiata pine plantations, making this species the most widely planted introduced conifer (Mead 2013). The stand presents a high-density of trees, but the canopy is fragmented. Numerous fallen trees are scattered throughout the plot. The stand is located on a very steep slope with an inclination of 25–30°.

Plot 4 covers 930 m^2 and represents a typical blue gum (*Eucalyptus globulus*) plantation (Fig. 4). Eucalypt plantations cover >20 million ha globally, representing a large portion of the total area of plantation around the globe (131 million ha according to Food and Agriculture Organization of the United Nations 2020). Blue gum is commonly grown for cellulose pulp production in Northern Spain. This particular stand suffers from poor management, resulting in very high tree density, abundant sprouting, and large gaps in the canopy. In addition, the terrain is very steep (24–26°).

Plot 5 covers an area of 800 m^2 and features a plantation of maritime pine (*Pinus pinaster*) (Fig. 4). It has a dense, homogeneous canopy, with tree heights ranging from 9 m to 11 m. The understorey is moderately developed, covering ~30% of the surface area. Notably, no down wood is present. This plot features the second highest tree density in the dataset (Table 1), very close to that of Plot 1. The terrain is almost flat, with negligible slope.

Table 1. Qualitative description of the 14 plots in the SegmentedForests dataset. Coordinates are EPSG:4326.

	Location (°N, °E)	Acquisition date	Type of stand	Type of trees	Main species	Main characteristics
Plot 1	Cantabria, Spain (42.9782, –4.5619)	Dec 2021	Natural	Broadleaf	*Q. pyrenaica*	Very high density, moderate slope
Plot 2	Asturias, Spain (43.1016, –4.4222)	Jul 2022	Natural	Broadleaf	*F. sylvatica*	Mature, sparse trees, thick branches, no understorey
Plot 3	Cantabria, Spain (43.1988, –4.1623)	Nov 2021	Plantation	Coniferous	*P. radiata*	High density, fallen trees, steep slope
Plot 4	Cantabria, Spain (43.3133, –4.0493)	Nov 2021	Plantation	Broadleaf	*E. globulus*	High density, sprouting, dense understorey, steep slope
Plot 5	Caceres, Spain (40.4330, –6.1980)	Jun 2021	Plantation	Coniferous	*P. Pinaster*	Very high density, short trees, even-aged
Plot 6	Vienna Woods, Austria (48.1226, 16.0476)	Sep 2017	Natural	Broadleaf	*F. sylvatica*	Multi-layer, uneven-aged, fallen trees
Plot 7	Vienna Woods, Austria (48.2331, 16.1780)	Sep 2021	Natural	Mixed	*P. abies*	Firebreak, high species diversity, sparse understorey
Plot 8	Vienna Woods, Austria (48.2331, 16.1780)	Sep 2021	Natural	Broadleaf	*F. sylvatica* and *F. excelsior*	Very dense canopy, standing dead trees and fallen trees, no understorey
Plot 9	Vienna Woods, Austria (48.2601, 16.2247)	Sep 2021	Natural	Mixed	*F.* sylvatica and *P. abies*	Natural regeneration, high species diversity
Plot 10	Vienna Woods, Austria (48.2441, 16.2091)	Sep 2021	Natural	Mixed	*A. alba* and *P. abies*	Multi-layer, uneven-aged, dense understorey
Plot 11	Loch Lomond, United Kingdom (56.0188, –4.5790)	Oct 2020	Natural	Broadleaf	*Q. robur*	Mature, sparse trees, thick branches, rocky terrain
Plot 12	Sycan Marsh, OR, USA (42.8450, –121.1644)	Jul 2019	Natural	Coniferous	*P. ponderosa*	Very sparse trees, natural regeneration, down wood
Plot 13	Sycan Marsh, OR, USA (42.8462, –121.1636)	Jul 2019	Natural	Coniferous	*P. ponderosa*	Low tree density, dense understorey, vertical continuity
Plot 14	Pebble Hill, GA, USA (30.7918, –84.0695)	Jun 2019	Plantation	Coniferous	*P. palustris*	Dense understorey, vertical discontinuity, flat terrain

Table 2. Quantitative description of the 14 plots in the SegmentedForests dataset. Details about how these metrics were computed are available in section 2.2 under point cloud processing.

	Area (m^2)	Density (stems/ha)	Dominant TH (m)	Dominant DBH (cm)	Canopy cover (%)	Understorey cover (%)	Slope (°)
Plot 1	650	3840	12–14	24–27	88	38	15–17
Plot 2	2100	280	26–28	85–90	95	1	13–14
Plot 3	1560	860	16–18	40–45	86	40	28–30
Plot 4	930	1620	20–24	25–30	69	81	24–26
Plot 5	680	3620	8–10	20–25	82	32	0–1
Plot 6	1300	530	26–29	38–41	90	48	12–14
Plot 7	1410	910	21–24	35–40	85	18	8–10
Plot 8	1390	2030	25–27	30–35	93	2	9–11
Plot 9	1460	580	36–40	68–78	92	4	5–7
Plot 10	1430	180	34–36	65–75	75	55	5–7
Plot 11	900	300	18–20	48–75	96	29	8–10
Plot 12	3590	150	18–22	50–60	18	51	0–1
Plot 13	3230	90	18–22	45–55	24	32	0–1
Plot 14	2490	330	23–26	45–55	45	65	2–3

Plots 6–10 (Austria)

Plots 6–10 are located in the Vienna Woods (Wienerwald), West of Vienna, Austria (Fig. 2). The Vienna Woods form a transitional zone between the Oroborealoid Bioclimatic Region to the South and the Nemoral Bioclimatic Region (Walter 1979) to the North and East. This area is characterized by a mix of climatic influences, including cooler, wetter conditions typical of Alpine zones and warmer, drier conditions associated with Continental Europe. Annual precipitation ranges from 700 mm to 1200 mm, and temperatures are variable, with cold winters and mild summers. Renowned for its ecological and cultural significance, the Vienna Woods are designated as a UNESCO Biosphere Reserve (UNESCO, n.d.), highlighting their importance for biodiversity conservation, sustainable land use, and research. These plots showcase a wide range of forest conditions, including variations in tree species composition, forest structure, and age classes. This reflects the ecological complexity and biodiversity typical of temperate continental forests.

Plot 6 covers ~1300 m^2 and is dominated by European beech (Fig. 5). It features an uneven-aged, managed deciduous forest. The canopy is relatively open with multi-layered structure typical of forests with complex age and size distributions. The

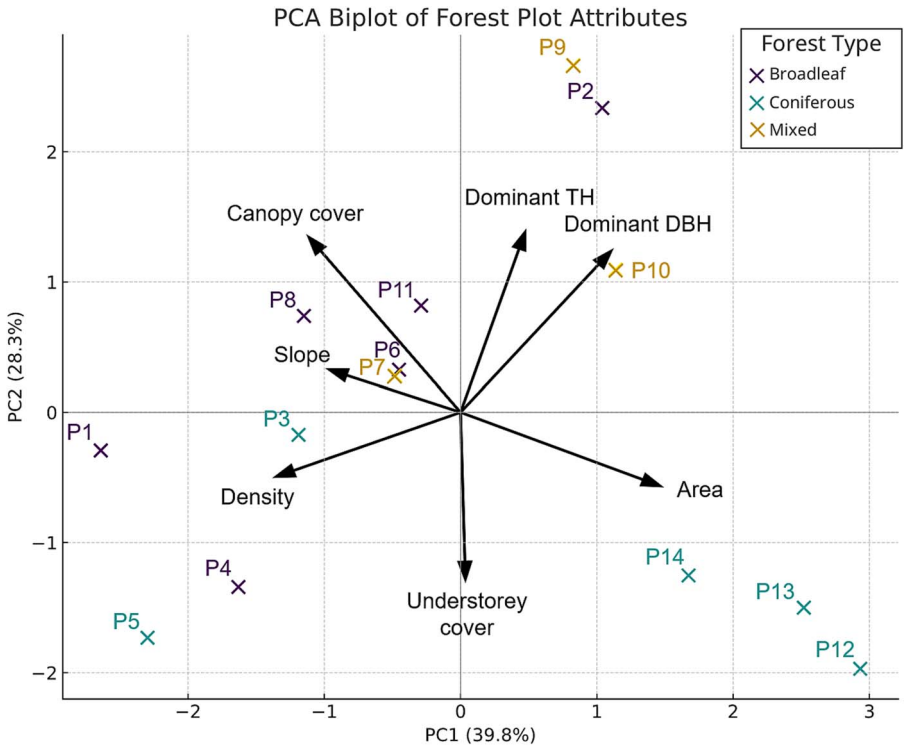

Figure 3. PCA biplot showing the structural diversity of the 14 forest plots included in the SegmentedForests dataset. Arrows represent the loadings of the original variables, scaled to indicate their relative contribution to the first two principal components (PC1 and PC2). Points are coloured by forest type (coniferous, broadleaf, or mixed), illustrating the distribution of structural characteristics across the dataset. Together, PC1 and PC2 explain 68.1% of the total variance in the dataset.

Table 3. Description of the labelled point clouds. Columns 'Ground + ground vegetation %', 'shrubs %', 'stems %', 'Branches + leaves %' and 'Other classes %' show the percentage of points in each point cloud that falls in each category. 'Other classes %' includes points assigned to classes other than 'ground+ground-vegetation', 'shrubs', 'stems' and 'branches+leaves' (Fig. 7). Totals (in bold) represent the sum of the total number of points and the mean class proportions.

	Scanning technology	Number of points (millions)	Ground + ground vegetation %	Shrubs %	Stems %	Branches + leaves %	Other classes%
Plot 1	MLS	20.8	23	8	29	40	0
Plot 2	MLS	23.9	28	0	8	64	0
Plot 3	MLS	42.4	10	16	22	50	2
Plot 4	MLS	32.5	5	38	20	33	4
Plot 5	MLS	30.7	18	11	24	47	0
Plot 6	TLS	77.1	24	7	19	43	7
Plot 7	TLS	59	31	5	28	33	3
Plot 8	TLS	280.5	51	1	12	27	9
Plot 9	TLS	72.6	37	0	27	34	2
Plot 10	TLS	42.5	12	16	28	40	4
Plot 11	MLS	19	16	7	10	63	4
Plot 12	TLS	61.1	36	19	2	37	6
Plot 13	TLS	64.7	30	21	3	43	3
Plot 14	TLS	95	4	16	7	69	4
	TOTALS:	**921.8**	**23.2**	**11.8**	**17.1**	**44.5**	**3.4**

understorey is abundant, contributing to the overall biodiversity and structural complexity of the stand. Additionally, large fallen trees and down wood are present.

Plot 7 covers 1410 m^2 of a dense, mixed forest, where Norway spruce (*Picea abies*) is the main species (Fig. 5). Additionally, the plot includes smaller populations of European beech, fir (*Abies alba*), pine (*Pinus spp.*), and European larch (*Larix decidua*), contributing to the species diversity. A firebreak is present in the plot, what causes a large gap in the canopy. The plot features abundant down wood and moderate understorey coverage. The DBH of the trees ranges from ~10 cm to 40 cm, reflecting variability in age and size classes. This diversity in species composition and tree dimensions, in addition to the firebreak, creates a structurally complex forest.

Plot 8 covers 1390 m^2. It is part of a very dense, deciduous forest, characterized by low species diversity and a dominance of European beech (Fig. 5). Whilst beech trees make up the majority of individuals, the plot also includes smaller populations of ash (*Fraxinus excelsior*), maple (*Acer pseudoplatanus*), and elm (*Ulmus glabra*), adding some diversity to the stand. The canopy

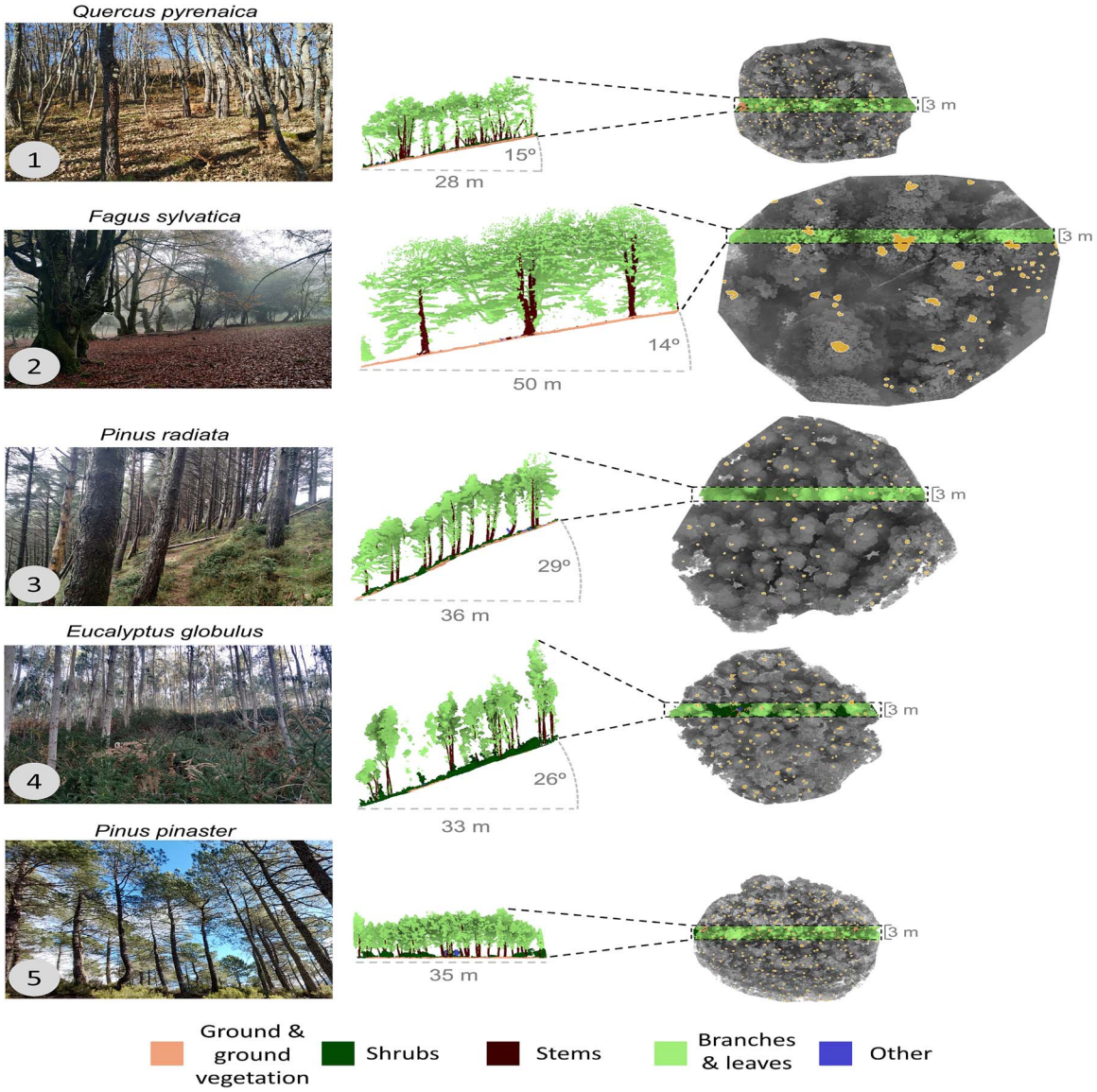

Figure 4. Plots in Spain. Left: pictures of the plots. Centre: cross-section (3 m wide) of the segmented point clouds acquired from the plots. The cross-sections are taken in the line of greatest slope (slope is not represented if it is <5°). The four main classes are represented in separate colours. Other classes have been merged for visualization purposes. Right: top view of the point clouds, showing the 3 m wide cross-section (green bar) and the stem map (golden points). Plot pictures reproduced with permission from Nelson Díaz Álvarez.

is extremely dense, which significantly limits light penetration to the forest floor. As a result, the understorey is almost non-existent, with minimal vegetation able to establish beneath the closed canopy. The plot features several standing dead trees and contains the highest volume of coarse woody debris across all plots in this dataset. This high volume of down wood is an important ecological feature, providing habitats for fungi, insects, and other decomposer organisms.

Plot 9 spans 1460 m^2 and features a natural regeneration, mixed forest, exhibiting a diverse composition of both broadleaf and coniferous tree species (Fig. 5). The species present are European beech, Norway spruce, black alder (*Alnus glutinosa*), fir, and ash. There is a strong mix of age classes and canopy structures within the stand, as reflected in the DBH of the trees, which ranges from 10 cm to 77 cm, approximately. As in Plot 8, the shrub layer is minimal. Plot 9 includes the tallest trees in the dataset, reaching heights of 40 m.

Lastly, Plot 10 covers 1430 m^2 of a multi-layered, mixed forest, dominated by fir and Norway spruce (Fig. 5), which form the upper canopy. These conifers are intermixed with broadleaf species, including pedunculate oak (*Quercus robur*) and European beech, which occupy a lower canopy layer beneath the dominant conifers. The plot is thus characterized by its tall, dominant trees and a highly stratified structure. The DBH of the trees ranges from ~20 cm to 74 cm, reflecting the mix of mature and dominant individuals. In addition to the two main canopy layers, the plot contains a large cohort of younger trees, forming additional layers between 1 m and 11 m in height. This younger cohort consists of hundreds of individuals, which, in addition to a dense understorey layer, contributes to the forest's structural complexity. Plot 10 is second to Plot 9 in terms of TH across the dataset, with heights ranging from 34 m to 36 m (Table 2).

Plot 11 (United Kingdom)

Plot 11 is located in Loch Lomond, Scotland, United Kingdom (Fig. 2), within the Atlantic Bioclimatic Region (European Environment Agency 2016). This region is characterized by a mild and wet

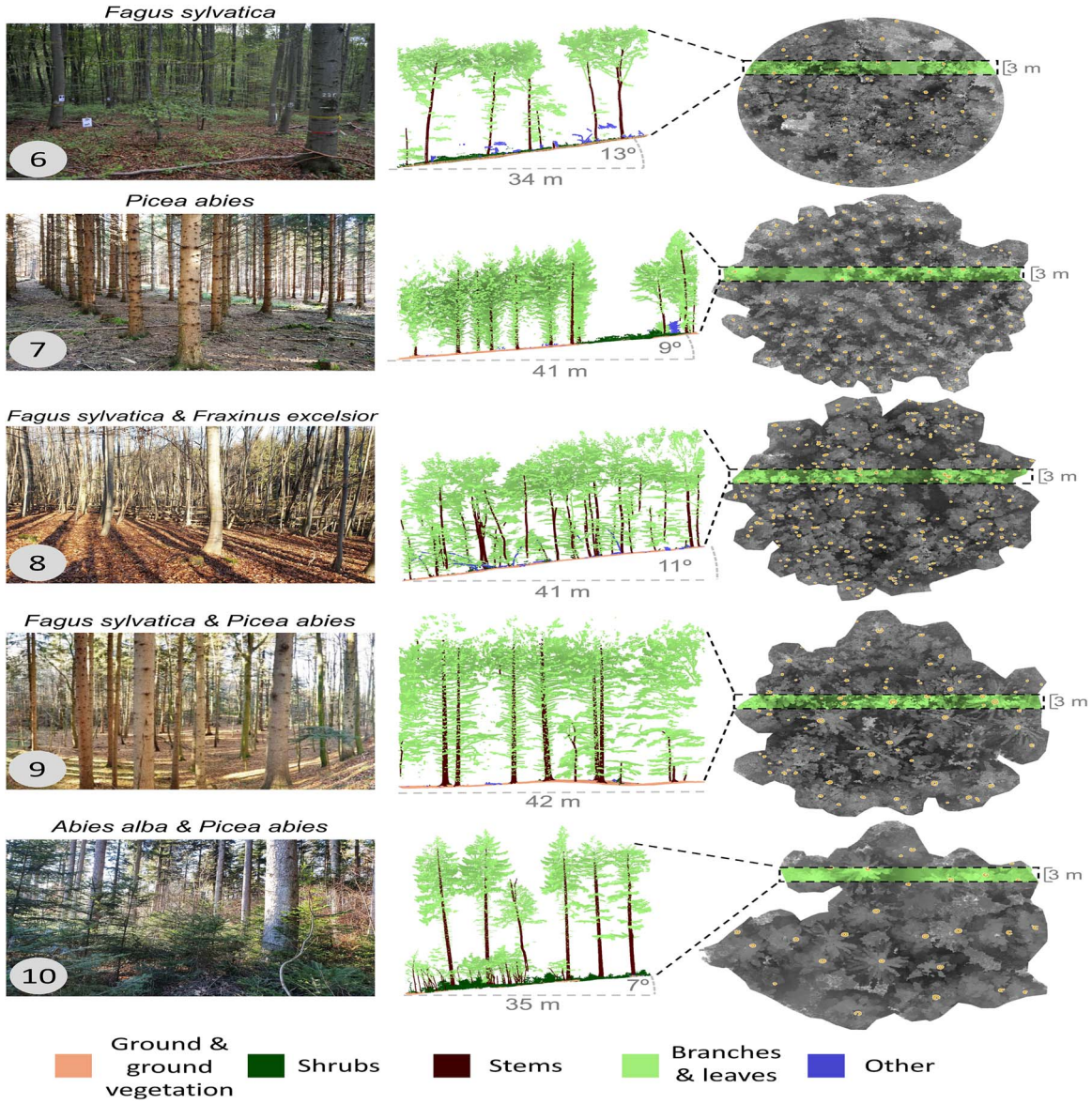

Figure 5. Plots in Austria. Left: pictures of the plots. Centre: cross-section (3 m wide) of the segmented point clouds acquired from the plots. The cross-sections are taken in the line of greatest slope (slope is not represented if it is <5°). The four main classes are represented in separate colours. Other classes have been merged for visualization purposes. Right: top view of the point clouds, showing the 3 m wide cross-section (green bar) and the stem map (golden points). Plot pictures reproduced with permission of Markus Hollaus and Yi-Chen Chen.

maritime climate, with high annual precipitation and relatively stable temperatures year-round due to the moderating influence of the nearby Atlantic Ocean. Loch Lomond itself is part of the Loch Lomond and The Trossachs National Park, a protected area renowned for its landscapes of woodlands, hills, and freshwater lochs. The forests in this national park are typically dominated by temperate broadleaf species, which thrive under the region's high humidity and abundant rainfall.

Plot 11 spans ~900 m^2 and features a monospecific stand dominated by pedunculate oaks (Fig. 6), a species well-suited to the temperate and humid conditions of the area. The trees are sparse and mature, with thick stems and large, sprawling branches, forming a very dense canopy that limits light penetration to the forest floor. The size and diameter of the trees is highly variable, as DBH varies from 13 cm to 74 cm (Table 2). A distinctive feature of this plot is the high density of large rocks on the ground, adding a unique structural and ecological element to the site.

Plots 12–14 (United States)

The dataset includes three plots from the United States (Fig. 2), representing distinct forest ecosystems in two major bioclimatic regions. Plot 12 and Plot 13 are located in Sycan Marsh, a high-elevation wetland in south central Oregon that falls within the North-Western Forested Mountains Bioclimatic Region (Commission for Environmental Cooperation 1997). Lastly, Plot 14 is located in Pebble Hill, Georgia, an example of an Eastern Temperate Forest (Commission for Environmental Cooperation 1997). Plots 12 and 13 feature a mixed-age forest structure dominated by ponderosa pine (*Pinus ponderosa*), with some Western juniper (*Juniperus occidentalis*) also present. Ponderosa pine is the most widely distributed pine in North America (Lowery 1984, Fryer 2018), ranging from Canada to Mexico and providing critical habitat for numerous wildlife species. It is one of the most commercially valuable tree species in the United States, playing a significant role in the forestry industry.

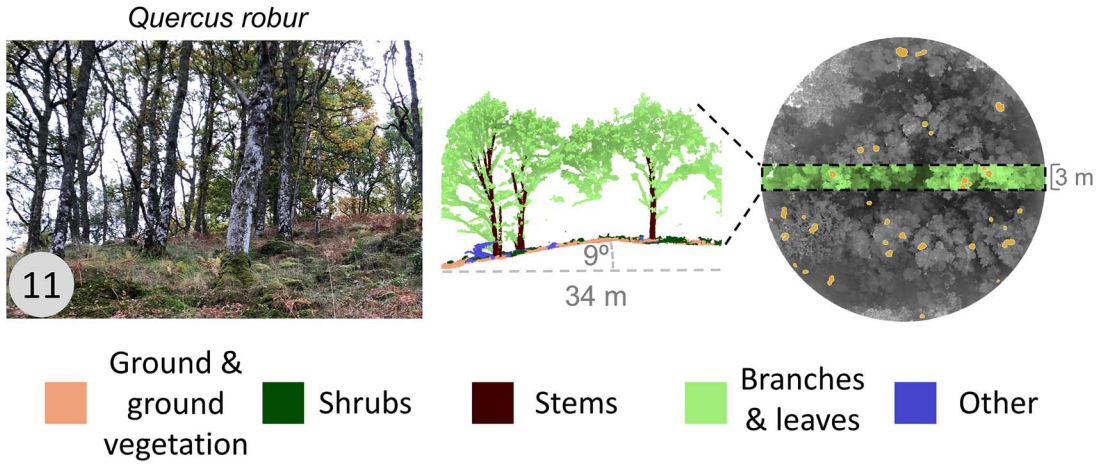

Figure 6. Plot in United Kingdom. Left: picture of the plot. Centre: cross-section (3 m wide) of the segmented point cloud acquired from the plot. The cross-section is taken in the line of greatest slope. The four main classes are represented in separate colours. Other classes have been merged for visualization purposes. Right: top view of the point cloud, showing the 3 m wide cross-section (green bar) and the stem map (golden points).

Plot 12 covers 3590 m^2 and features mature, very sparse trees ranging in height from 18 m to 22 m, that do not form a connected canopy (Fig. 7). The understorey is relatively open, interspersed with cohorts of young trees that contribute to the structural diversity.

Plot 13 covers 3230 m^2 and exhibits a similar distribution of sparse, mature trees to Plot 12 but it has a denser understorey, with shrubs covering ∼50% of the ground (Fig. 7). Additionally, the plot features a significantly larger cohort of young trees, indicative of more active regeneration processes. This, combined with the fact that the branches of mature trees often start at or near the base of the stems, creates minimal vertical gaps between the understorey and the tree crowns. This arrangement results in a nearly continuous vertical structure, enhancing vertical fuel continuity and increasing the potential for crown fire propagation, whist also providing habitat connectivity for species dependent on the understorey.

Finally, Plot 14 spans over 2600 m^2 and is dominated by a sparse plantation of longleaf pine (*Pinus palustris*) (Fig. 7). This species is iconic to the South-Eastern United States, historically covering vast areas of the coastal plains and now a focus of conservation and restoration efforts due to its ecological importance. Nowadays, there are ∼2.1 million ha of longleaf pine forests in the United States (America's Longleaf Restoration Initiative 2023), with notable populations in Florida, Georgia, and Alabama, as well as isolated pockets in other South-Eastern states. The dominant longleaf pines in this plot do not form a dense canopy. However, several layers of smaller, younger pines are present, in addition to a shrub layer that covers ∼65% of the ground. The terrain is nearly flat.

Point cloud acquisition and processing

Point cloud acquisition

The SegmentedForests dataset includes point clouds acquired using various technologies to enhance the variability across forest plots. All plots were scanned on dry days with minimal to no wind conditions to reduce the presence of artefacts. The plots in Spain (Plots 1–5, Fig. 4) and the plot in the United Kingdom (Plot 11, Fig. 6) were scanned during 2021 and 2022 using a GeoSLAM ZEB Horizon MLS (GeoSLAM 2023). The GeoSLAM ZEB Horizon features a range of 100 m, a field of view (FOV) of 360° × 270°, and can capture 300 000 points per second with a relative accuracy of up to 6 mm.

The plots in Austria (Plots 6–10, Fig. 5) were scanned with two different TLS devices and in two different campaigns. Plot 6 was acquired during the 'Benchmarking of close-range photogrammetry methods for forestry applications' project (gis.tuzvo.sk/benchcrp/). It was scanned in 2017 using a Riegl VZ-2000i TLS scanner (RIEGL Laser Measurement Systems GmbH 2023a) from 16 scanning positions. Further information about Plot 6 can be found in Piermattei et al. (2019), where it is referred to as 'Plot 2'. Plots 7, 8, 9 and 10 are part of the 'SilviLaser 2021 Benchmark Dataset—Terrestrial Challenge' dataset (Hollaus and Chen 2023). They correspond to SL21BM_TER_038, SL21BM_TER_040, SL21BM_TER_042 and SL21BM_TER_044, respectively. These plots were scanned in 2021 using a Riegl VZ-400i TLS device (Riegl Laser Measurement Systems GmbH 2023b), with the number of scan positions varying depending on plot characteristics. The Riegl VZ-2000i offers a FOV of 100° × 360°, captures up to 500 000 points per second, has a relative accuracy of up to 5 mm, and a range of up to 2500 m. The Riegl VZ-400i has similar specifications but a reduced range of up to 800 m and a higher relative accuracy of up to 3 mm.

The plots in the United States (Plots 12–14, Fig. 7) were also scanned using the two TLS systems mentioned above. The Sycan Marsh plots (Plots 12 and 13) were scanned in 2021 with a RIEGL VZ-400i as part of a larger survey in which the scanner was placed every 50 m. In contrast, the Pebble Hill plot (Plot 14) was scanned in 2017 using a RIEGL VZ-2000i from eight scan positions.

Point cloud processing

All MLS point clouds were registered using GeoSLAM Hub software (GeoSLAM 2021), whist all TLS point clouds were registered using RiSCAN Pro 2.0 (RIEGL Laser Measurement Systems 2022). Minor registration errors can be presumed, although no visible misalignments or artefacts were detected during the semantic annotation process. Each point cloud contains millions of points, capturing the fine details of the forest structure. No global target density was enforced for the dataset. The plots from the United States underwent subsampling to achieve better point distribution uniformity due to sparse coverage in some areas of plots: plots 12 and 13 were subsampled to a 1 cm resolution, whereas Plot 14 was subsampled to a 0.5 cm resolution. The point clouds from Spain, Austria, and the United Kingdom are provided at their original point densities. Therefore, the spatial resolution of the point clouds varies ranging from millimetre to centimetre-level.

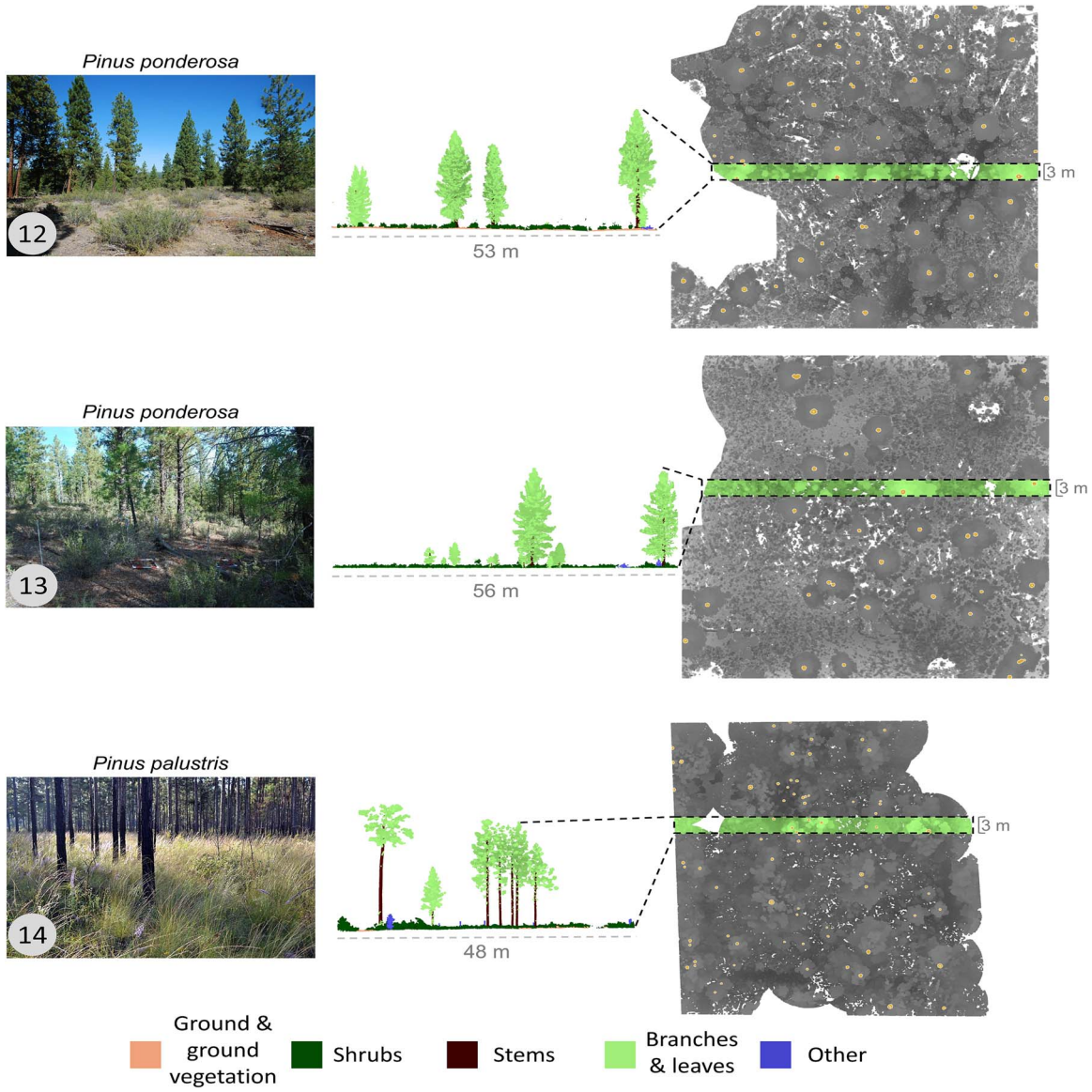

Figure 7. Plots in the United States. Left: pictures of the plots. Centre: cross-section (3 m wide) of the segmented point clouds acquired from the plots. The cross-sections are taken in the line of greatest slope (slope is not represented if it is <5°). The four main classes are represented in separate colours. Other classes have been merged for visualization purposes. Right: top view of the point clouds, showing the 3 m wide cross-section (green bar) and the stem map (golden points).

After registration and density adjustment, point clouds were further processed to derive structural descriptors. CloudCompare (v2.13.2), a widely used point cloud processing free-access software (Girardeau-Montaut 2024) was employed. In CloudCompare, plots were cropped to their final shapes and the quantitative descriptors shown in Table 2 were computed. TH and DBH were computed using the 3DFin software (v0.4.1), freely available within CloudCompare as a plug-in (Laino et al. 2024). From its outputs, the top 10% TH and DBH were selected, and their approximate range was used to determine the values of the dominant TH and DBH. Tree density was computed using the number of detected stems divided by the estimated plot area, both also provided by 3DFin. To compute canopy and understorey cover, 3DFin-generated normalized height values (Z_0) were used to vertically project the classified points. Using CloudCompare's 'Rasterize' tool, the horizontal (XY) coverage of relevant classes was mapped using a 10 cm resolution grid. Canopy cover was defined as the proportion of plot area covered by rasterized 'branches + leaves' points. Understorey cover was similarly computed using the merged 'shrub' and 'small tree branch' classes. 3DFin configuration files, which contain the values for all parameters used, are provided alongside the dataset.

Point cloud labelling

To prepare the semantic labels in SegmentedForests, a systematic approach for manual segmentation of the point clouds in CloudCompare was followed (Fig. 8). Initially, as each point cloud had been processed using the software 3DFin (Fig. 8, step 1), the scalar fields that it generated (normalized height (Z_0) and distance to the nearest tree axis) were used to enhance the visual representation of the point clouds in CloudCompare. Both variables provide visual cues that aid the operator in distinguishing between different vegetation structures within the unlabelled versions of the point clouds.

Subsequently, each point cloud was divided into horizontal slices every 2 metres along the vertical axis (Fig. 8, step 2) by

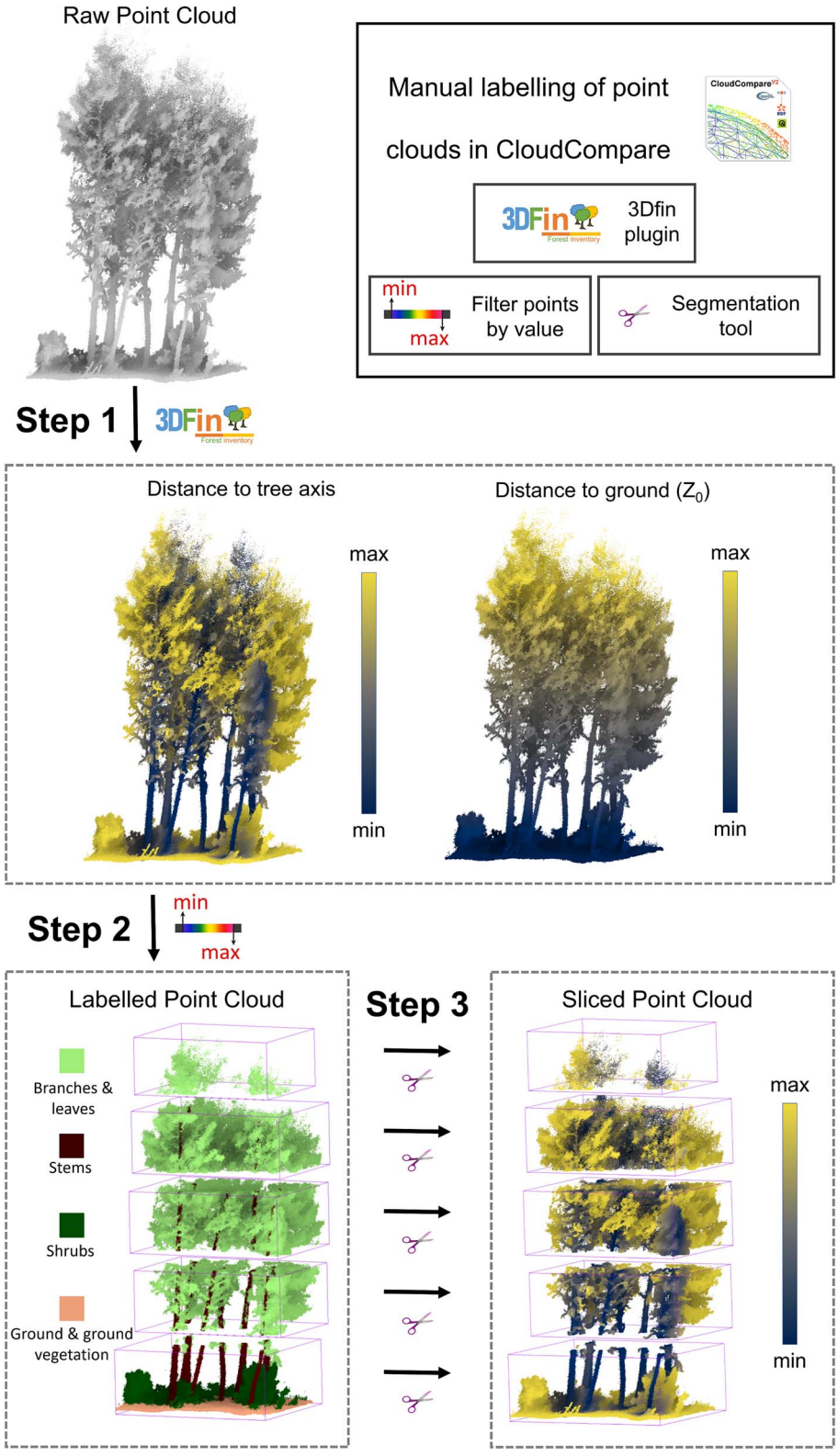

Figure 8. Workflow for the manual segmentation of the point clouds in CloudCompare. Step 1: The point clouds were processed with the 3DFin plugin, to compute distance to the ground and distance to the nearest tree axis. Step 2: The distance to ground is used to slice the point clouds every 2 metres along the vertical axis. Step 3: Points within each slice are manually annotated using the 'segment' tool.

filtering the value of Z_0. This slicing allowed annotators to systematically examine and label smaller sections of the plot at a time, improving accuracy and reducing annotation complexity. Manual segmentation was then performed slice by slice using the "Segment" tool within CloudCompare (Fig. 8, step 3). This tool provides a polygonal line that defines a 3D polygon, which can be used to enclose and label points within the point cloud. This workflow was applied consistently across all plots to ensure annotation quality and consistency.

Despite these precautions, the manual segmentation process remains challenging, as discerning which class a point belongs to is not always straightforward, even for the human eye. This

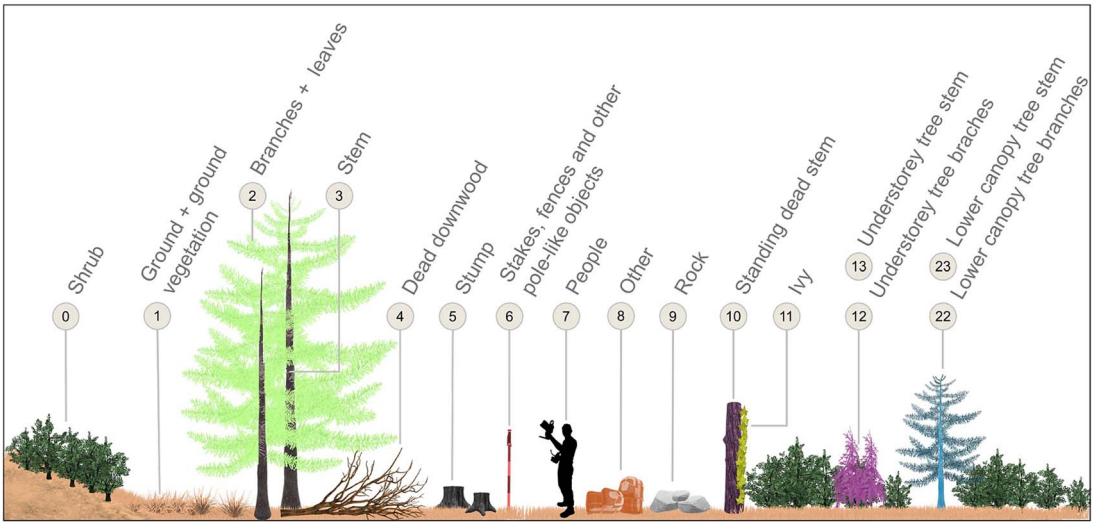

Figure 9. Illustration of the 16 semantic classes present in SegmentedForests.

complexity arises from several factors, including the structural ambiguity of forest components, occlusion effects and variability in point cloud density and quality between and within the plots. These challenges are particularly evident in areas where point density is low or where points are unevenly distributed, making it difficult to establish clear boundaries between classes. For instance, the transition between tree trunks, larger branches, smaller branches, and foliage is often gradual, particularly for broadleaf species, where there is no distinct demarcation between these features (Krisanski et al. 2021, Kaijaluoto et al. 2022). This is especially relevant in complex forest environments where shading or occlusion can obscure features. Moreover, noise in the data or low resolution from the scanner can exacerbate these challenges. In addition, branches may in some cases be confused with trunk points due to their position and proximity (Ma et al. 2023). This issue becomes increasingly pronounced in the upper canopy, where sensor limitations such as beam divergence and occlusion further degrade scene reconstruction quality, resulting in ambiguous point classifications (Krisanski et al. 2021). Similarly, distinguishing between ground and low vegetation is problematic in dense understorey conditions, where ground-level points may overlap with low vegetation, leading to potential misclassification (Kaijaluoto et al. 2022).

To address these challenges, a set of consistent criteria was applied across the workflow. Semantic labels were created following two general guidelines. First, if a structure is clearly visible and well-defined in the point cloud, it must be segmented, even if doing so requires introducing a new label. An example of this is ivy, which was clearly visible over some trunks in some of the TLS point clouds and thus it was assigned a specific label. Second, for points that are difficult to assign a class, the final criterion to label them must be decided and applied consistently, both within the same plot and between different plots. For example, irregular structures near the ground surface are labelled as 'shrub' when in doubt. These guidelines provided a structured framework to handle the inherent variability and ambiguity associated to this task, and resulted in the assignment of 16 different semantic labels. These are illustrated in Fig. 9.

Semantic labels are encoded in the point cloud in a scalar field named 'Class'. Labels 6–9 correspond to non-vegetation structures found within the point clouds, whist all other labels refer to vegetation structures. 'Ground + ground vegetation' was defined to include all points lying on or immediately above the ground surface, encompassing elements such as low herbaceous layers. Understorey vegetation was generally assigned the 'shrub' label, unless it was clear that the structure represented a small tree. These distinctions were often based on structural and spatial criteria, such as height, branching patterns, and the presence of a clear stem (even if thin). Two types of understorey trees were identified: the first type ('understorey tree') shared the shrub layer's height but was still structurally different, whist the second type ('lower canopy tree') was distinctly taller than the shrub layer (Fig. 9). Stem was generally the most easily distinguishable class, defined as the main, vertical axis of tree growth. However, this definition often became difficult to apply in practice. Figure 10 illustrates the criteria followed to label stem points. In Plot 2 and Plot 11, for example, many trees lacked a clearly defined vertical growth axis. In such cases, main branches were identified, and only those growing vertically or near-vertically were assigned the 'stem' label (Fig. 10b). This approach was preferred over leaving certain trees without any assigned stem. Similarly, bifurcate and trifurcate trunks were also labelled as stems (Fig. 10c). Plot 4, for instance, contains several trees with bifurcations, which were treated consistently under this criterion. On occasion, some branches were found to grow nearly vertically (Fig. 10d), which made finding the stem harder than initially expected.

Additional vegetation classes, such as 'down wood', 'stump', and 'ivy', were used in some cases and only when the features were clearly distinguishable. Similarly, non-vegetation classes were labelled where present to ensure a comprehensive and accurate representation of all point cloud elements.

Ten different operators contributed to the manual segmentation of the SegmentedForests dataset, all following the labelling criteria described above. However, in cases where boundaries between classes were ambiguous, operator discretion played a critical role in ensuring consistency. To minimize variability in interpretation, the entire dataset was thoroughly revisited by a single operator (DL, first author of this study), ensuring uniform application of the labelling criteria across all plots. Overall, the process of manual segmentation of the dataset was extremely time-consuming, as has been noted by other authors (Kaijaluoto et al. 2022, Lines et al. 2022a, Ma et al. 2023, Van den Broeck et al. 2023). Labelling each plot required between 60 and 180 hours, depending on its size, complexity, and point density. On average,

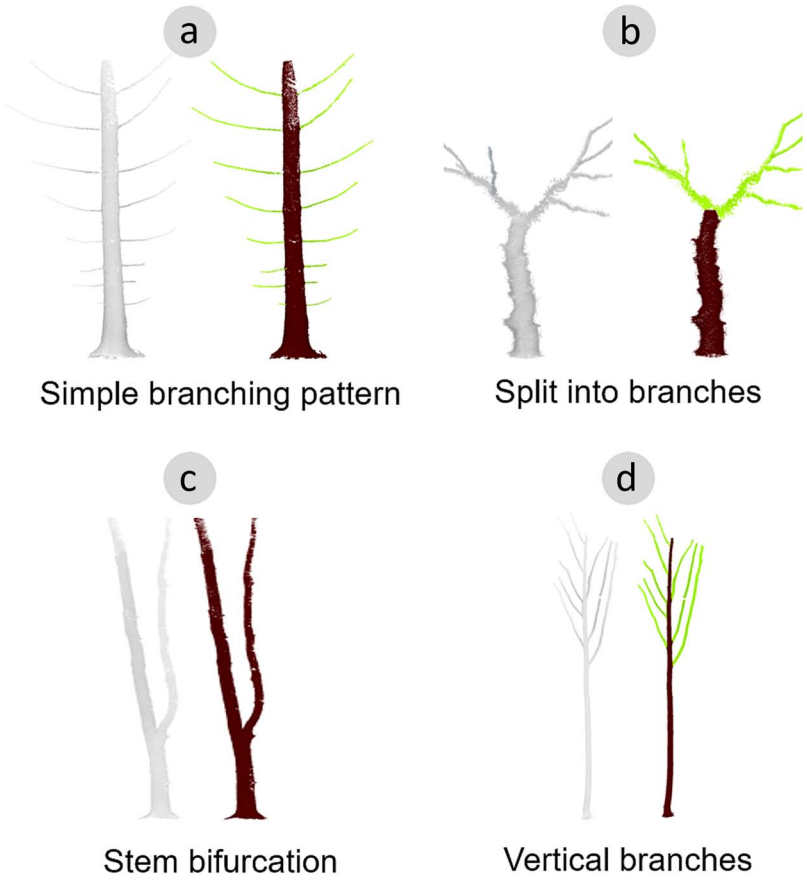

Figure 10. Different branching patterns found whist manually segmenting the point clouds. Each requires a defined criterion to differentiate between 'stem' and 'branch + leaves' labels. (a) Represents an easy case where the stem grows vertically and the branches grow laterally. The stem is noticeably wider than the branches. (b) Represents a tree where the transition from stem to branches is not so clear, and there are branches that become main axes of growth. (c) Represents a bifurcate stem. Each of the two stems in a bifurcate tree have been labelled as 'stem'. (d) Represents a tree where several branches grow almost vertically. These branches were found to be harder to tell apart from the stem.

segmenting each plot required the equivalent of 21 full working days, assuming an 8-hour workday. In total, >920 million points were manually annotated across the SegmentedForests dataset. Table 3 shows information about the segmented point clouds.

Data splitting

To facilitate model development, evaluation, and benchmarking, each point cloud in the SegmentedForests dataset includes a predefined partitioning. A scalar field named 'Splits' is provided for every point, encoding its assignment to an integer code: 0 for training, 1 for validation, and 2 for testing. For the dataset release, we adopted a reference split aiming for ~50%, 25%, and 25% of the points in each category, respectively. Figure 11 illustrates this spatial subdivision. It is important to note that this ratio reflects a practical choice rather than a strict rule, and users are free to define alternative splits if needed. For instance, the provided splits can easily be aggregated to create alternative schemes, such as 75% training (i.e. by merging training and test splits), 25% validation sets and test with whole, unseen plots in a leave-one-out cross-testing to evaluate model generalization.

Although every effort was made to balance these subsets, achieving a precise 50/25/25 split for all properties (such as area, number of points, and class proportions) in each plot was not always feasible, due to the inherent spatial irregularities in forest environments. Particularly, during the splitting process, we strived to keep entire objects (especially trees) intact within a single subset, i.e., if a tree happens to cross a split boundary, all its points are assigned to the subset containing its base.

Discussion, limitations, and future work

Comparing our dataset to similar products is difficult due to the scarcity of publicly available alternatives. However, when doing so, the SegmentedForests dataset stands out for its diversity and level of annotation detail. A summary comparison to TreeScope (Cheng et al. 2024) and FOR-Instance (Puliti et al. 2023) is shown in Table 4. For example, Cheng et al. (2024) provides annotated MLS data from 3.2 ha of forests, and 46 ha of orchards. Whilst the total area covered is extensive, the semantic labels in that dataset are less detailed. This limited labelling restricts its use for applications requiring a detailed understanding of forest structures. In contrast, the SegmentedForests dataset includes annotated point clouds from a smaller total area (2.3 ha) but spans diverse countries, regions and multiple ecological conditions. Additionally, even though the area covered by Cheng et al.'s (2024) dataset is significantly bigger in area, it includes 1860 trees, which is not far from the 1620 trees covered in SegmentedForests.

The FOR-Instance dataset (Puliti et al. 2023) is the second notable alternative. While this dataset includes a larger number of plots (29), a significant proportion of its labelled area (~42%) is classified as 'out-points', reducing its effective labelled area. In contrast, the SegmentedForests dataset covers a slightly smaller

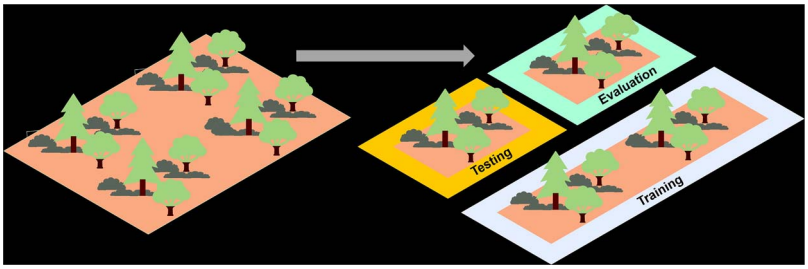

Figure 11. Illustration of the 50/25/25 splitting that has been performed in the point clouds.

Table 4. Comparison of the main characteristics of SegmentedForests and freely-available comparable datasets.

	SegmentedForests	**TreeScope**	**FOR-Instance**
Scanning technology	MLS and TLS	MLS and ULS	ULS
Perspective	**Ground-based**	**Ground-based and Aerial**	Aerial
Countries covered	4	1	5
Total area (ha) forest | plantation	2.3 1.7 | 0.6	49.3 3.2 | 49	2.8 2.5 | 0.3
Tree count	∼1600	∼1900	∼1100
Semantic classes	**16**	3	6

total area (∼2.3 ha) but provides denser semantic labels and a higher total tree count (∼1100 trees in FOR-Instance vs. ∼1600 trees in SegmentedForests). The main difference however lies in the data acquisition method. The point clouds in FOR-Instance are captured with ULS, which, whist relatively dense, is not as detailed as TLS or MLS data and represents a fundamentally different perspective. Ground-based point clouds, such as those in SegmentedForests, provide detailed structural information from below the canopy, capturing elements like ground vegetation and lower tree layers more effectively. These differences make SegmentedForests and FOR-Instance complementary resources, addressing distinct research needs in forest monitoring.

In addition to its primary purpose of supporting semantic segmentation models, the SegmentedForests dataset holds significant potential for other applications in forest research and management (Xiang et al. 2024). The diversity in tree species, canopy structures, ground vegetation, and management practices across the plots makes it an ideal resource for testing and developing a wide range of models. For instance, the dataset could potentially be used to build models that predict stem diameter and volume, compute shrub coverage, analyze canopy gap fraction, or assess light microenvironments within forest stands. These models could address key challenges in forest-related applications, such as quantifying biomass, monitoring forest health, and studying structural and functional forest dynamics.

Whilst the SegmentedForests dataset presented here represents a significant step forward in providing annotated point clouds for forest research, it still faces limitations. The dataset encompasses substantial variability in forest structure and semantic labels, but does not yet represent all forested ecosystems. For instance, alpine forests, tropical rainforests, and woodland savannahs are absent. These forested ecosystems are critical to global biodiversity and carbon dynamics, and their inclusion in future versions of the dataset would expand its applicability and impact. Such an expansion would make the dataset more representative of global forest diversity and enable broader ecological and forestry research. In this regard, we plan on releasing an updated version of the dataset in the near future with coverage of boreal forests and more downed woody debris.

The manual segmentation process, though systematic and carefully supervised, is inherently susceptible to small errors due to the sheer volume of points and the subjectivity of human operators (Van den Broeck et al. 2023). An additional limitation of the dataset is that no field validation has been performed to assess the accuracy of the semantic segmentation of the point clouds. This is because the original purpose of the field campaigns where these point clouds were obtained was not that of manual segmentation. Moreover, colour was not available in the point clouds, which would have made the process more reliable. The labelling is based on features that are visually identifiable in the 3D data, such as point density, spatial arrangement, and relative position. Whilst this approach provides a reasonable approximation to forest structure, it inherently relies on the subjective judgment of operators and the resolution of the point clouds, which may lead to occasional misclassifications. Even though every effort was made to minimize these inconsistencies, minor variations in the application of criteria may still exist. These limitations reflect the broader challenges of manual point cloud annotation and underscore the need for automated tools and consistent methodologies in future dataset development. All this being said, human-in-the-loop segmentation also brings several potential benefits when dealing with large datasets (Jain et al. 2019) that help to balance these limitations.

Lastly, the use of the "branches+leaves" label, which groups both photosynthetic material (leaves) and non-photosynthetic material (branches) into a single category, should be pointed out as a specific limitation. This grouping is not ideal, as it limits the dataset's applicability for studies focusing on parameters like leaf area index (LAI), respiration rates, or photosynthesis-related functions, all of which would greatly benefit from precise separation between these components. Additionally, separating leaves from branches would allow merging stem and branch classes into a single 'wood' class. This would address another challenge faced during the manual point cloud segmentation: the difficulty in distinguishing stems from branches. In many cases, the boundary between thick branches and stems is ambiguous, especially for deciduous species with complex branching patterns. By merging the stem and branch labels into a single 'wood' class, whist separating leaves into their own category, this issue could be effectively mitigated. Such a refinement could improve both consistency across point clouds and the accuracy of the models developed using SegmentedForests. The distinction

between leaves and branches, however, has not been included in this version of the dataset due to the sheer complexity of the task. Nevertheless, this enhancement is currently a work in progress, and future versions of the dataset will aim to provide this additional level of detail.

Conclusions

The SegmentedForests dataset represents a substantial advancement in the field of forest point cloud research, offering a first-of-its-kind resource for the development and benchmarking of semantic segmentation models. Its combination of size, diversity, and annotation detail positions it as a valuable tool for researchers and practitioners seeking to explore new approaches to forest structure analysis. With over 920 million points, covering 2.3 ha of forest and encompassing >1600 trees across 16 species, the dataset provides a robust foundation for advancing DL-based segmentation models in forestry.

One of the most impactful aspects of SegmentedForests is that it is freely available, which ensures accessibility for the broader research community. This openness is aimed at fostering collaboration, innovation, and the creation of new models and methodologies, benefiting a wide range of users. Whilst its primary goal is to support semantic segmentation, the dataset's diversity and richness may inspire alternative uses, such as testing models that compute tree metrics (e.g. height, diameter, volume) or quantifying vegetation coverage. These possibilities underscore the dataset's potential to drive progress across multiple domains within forest ecology and management.

Future updates of the SegmentedForests dataset will further enhance its utility by addressing current limitations, such as refining the distinction between photosynthetic and non-photosynthetic materials and expanding coverage to include additional types of forest ecosystems. Moreover, future work by the authors will provide semantic segmentation models trained on SegmentedForests, which will also be made publicly available. These models, together with the dataset, will expand the toolkit available for the community, enabling researchers to tackle complex challenges in forest monitoring and modelling.

Acknowledgements

The authors wish to thank Álvaro Irurozqui Sicilia, María Pérez Vallejo and Víctor Hernández Nicolás for their contribution in point cloud labelling and Forest4F (https://forest4f.com). They are also gratefully to Louise Loudermilk, Markus Hollaus, Martin Mokros, Nelson Díaz Álvarez, and Yi-Chen Chen for providing some of the raw point clouds and pictures of the plots.

Conflict of interest: The authors declare no conflict of interest. The funders had no role in the design of the study, in the collection, analyses, or interpretation of data, in the writing of the manuscript, or in the decision to publish the results. This research was supported (in part) by the US Department of Agriculture, Forest Service. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the US government.

Funding

This work was supported by the UK NERC project (NE/T001194/1): 'Advancing 3D Fuel Mapping for Wildfire Behaviour and Risk Mitigation Modelling', the Spanish Knowledge Generation project (PID2021-126790NB-I00): 'Advancing carbon emission estimations from wildfires applying artificial intelligence to 3D terrestrial point clouds', (grant PRE2022-104159) funded by MCIN/AEI/10.13039/501100011033 and FSE+, research grant 'FIREPROs' (IDE/2024/000780) funded by the Principality of Asturias Government (Spain), COST Action 3DForEcoTech CA20118 supported by COST (European Cooperation in Science and Technology) and the US DoD SERDP/ESTCP projects (RC20-1025), (RC23-7626), and (RC20-1046).

Data availability

The dataset is available under a MIT license at https://zenodo.org/records/17396681.

References

Alonso-Rego C, Arellano-Pérez S, Cabo C. *et al.* Estimating fuel loads and structural characteristics of shrub communities by using terrestrial laser scanning. *Remote Sens* 2020;**12**:1–21. https://doi.org/10.3390/rs12223704.

Alvites C, Santopuoli G, Hollaus M. *et al.* Terrestrial laser scanning for quantifying timber assortments from standing trees in a mixed and multi-layered mediterranean forest. *Remote Sens* 2021;**13**:4265. https://doi.org/10.3390/rs13214265.

America's Longleaf Restoration Initiative. Range-Wide Conservation Plan (2025–2040). Alexandria, VA: America's Longleaf Restoration Initiative; 2023. Accessed November 16, 2024. https://www.americaslongleaf.org.

Béland M, Baldocchi DD, Widlowski JL. *et al.* On seeing the wood from the leaves and the role of voxel size in determining leaf area distribution of forests with terrestrial LiDAR. *Agric For Meteorol* 2014;**184**:82–97. https://doi.org/10.1016/j.agrformet.2013.09.005.

Belton D, Moncrieff S, Chapman J. Processing tree point clouds using Gaussian mixture models. ISPRS annals of the photogrammetry, remote sensing and spatial. *Inform Sci* 2013;**2**:43–8. https://doi.org/10.5194/isprsannals-II-5-W2-43-2013.

Breiman L. Statistical modeling: the two cultures (with comments and a rejoinder by the author). *Statistical Science* 2001;**16**:199–231. https://doi.org/10.1214/ss/1009213726.

Bütler R, Lachat T, Larrieu L, Paillet Y. Habitat trees: Key elements for forest biodiversity. *Integrative Approaches as an Opportunity for the Conservation of Forest Biodiversity*. Joensuu, Finland: European Forest Institute; 2013, 84–91.

Cabo C, Ordóñez C, López-Sánchez CA, Armesto J. Automatic dendrometry: Tree detection, tree height and diameter estimation using terrestrial laser scanning. *International Journal of Applied Earth Observation and Geoinformation* 2018;**69**:164–74. https://doi.org/10.1016/j.jag.2018.01.011.

Calders K. Terrestrial laser scans—Riegl VZ400, individual tree point clouds and cylinder models, Rushworth Forest. *Version 1 Terrestrial Ecosystem Research Network (Dataset)* 2014. https://doi.org/10.4227/05/542B766D5D00D.

Calders K, Adams J, Armston J. *et al.* Terrestrial laser scanning in forest ecology: expanding the horizon. *Remote Sens Environ* 2020;**251**:112102. https://doi.org/10.1016/j.rse.2020.112102.

Cheng D, Cladera F, Prabhu A. *et al.* TreeScope: An agricultural robotics dataset for LiDAR-based mapping of trees in forests and orchards. In: *2024 IEEE International Conference on Robotics and Automation (ICRA)*. Piscataway, NJ: IEEE; 2024, 14860–14866. https://doi.org/10.1109/ICRA57147.2024.10611103.

Chianucci F, Puletti N, Grotti M. *et al.* Nondestructive tree stem and crown volume allometry in hybrid poplar plantations derived from terrestrial laser scanning. *Forest Science* 2020;**66**:737–46. https://doi.org/10.1093/forsci/fxaa021.

Commission for Environmental Cooperation. *Ecological Regions of North America: Toward a Common Perspective*. Montreal, Quebec, Canada: Commission for Environmental Cooperation, 1997.

Disney MI, Boni Vicari M, Burt A. *et al.* Weighing trees with lasers: advances, challenges and opportunities. *Interface Focus* 2018;**8**:20170048. https://doi.org/10.1098/rsfs.2017.0048.

Ehbrecht M, Schall P, Ammer C. *et al.* Quantifying stand structural complexity and its relationship with forest management, tree species diversity, and microclimate. *Agric For Meteorol* 2017;**242**:1–9. https://doi.org/10.1016/j.agrformet.2017.04.012.

Engel N, Belagiannis V, Dietmayer K. Point transformer. *IEEE Access* 2021;**9**:134826–40. https://doi.org/10.1109/ACCESS.2021.3116304.

European Environment Agency. *Biogeographical Regions in Europe*. Copenhagen, Denmark: European Environment Agency. Accessed November 28, 2024. https://www.eea.europa.eu/en/analysis/maps-and-charts/biogeographical-regions-in-europe-2?activeTab=8a280073-bf94-4717-b3e2-1374b57ca99d.

Fassnacht FE, White JC, Wulder MA. *et al.* Remote sensing in forestry: current challenges, considerations and directions. *Forestry: An International Journal of Forest Research* 2023;**97**:11–37. https://doi.org/10.1093/forestry/cpad024.

Food and Agriculture Organization of the United Nations (FAO). *Global Forest Resources Assessment 2020: Main Report*. Rome, Italy: FAO; 2020. https://doi.org/10.4060/ca9825en.

Fryer JL. *Pinus Ponderosa Var. Benthamiana, P. p. var. ponderosa: Ponderosa pine. In Fire Effects Information System [Online]*. Forest Service, Rocky Mountain Research Station, Missoula Fire Sciences Laboratory (Producer). Retrieved November 11, 2024, from: U.S. Department of Agriculture, 2018. https://www.fs.usda.gov/database/feis/plants/tree/pinponp/all.html.

GeoSLAM (A FARO Technologies, Inc. Company). (2021). *GeoSLAM hub* [computer software]. *GeoSLAM Ltd*. Retrieved December 05, 2024, from: https://knowledge.faro.com/Software/GeoSlam/GeoSLAM_Hub.

GeoSLAM (A FARO Technologies, Inc. Company). *Geoslam ZEB Horizon RT* 2023. Retrieved December 05, 2024, from: https://geoslam.com/solutions/zeb-horizon-rt/.

Girardeau-Montaut D. *CloudCompare (Version 2.13.2) [Computer Software]*. 2024. Retrieved 08 August, 2024, from: http://www.cloudcompare.org.

Guo MH, Cai JX, Liu ZN. *et al.* PCT: Point cloud transformer. *Comput Vis Media*. 2021;**7**:187–99. https://doi.org/10.1007/s41095-021-0229-5.

Halperin D, Eisl N. Point cloud based scene segmentation: a survey. *arXiv* 2025. https://arxiv.org/abs/2503.12595.

Hollaus M, Chen Y-C. SilviLaser 2021 benchmark dataset—terrestrial challenge (1.1) [data set]. *TU Wien* 2023. https://doi.org/10.48436/kndye-egv02.

Hyyppä E, Kukko A, Kaijaluoto R. *et al.* Accurate derivation of stem curve and volume using backpack mobile laser scanning. *ISPRS Journal of Photogrammetry and Remote Sensing* 2020;**161**:246–62. https://doi.org/10.1016/j.isprsjprs.2020.01.018.

Jain S, Munukutla S, Held D. Few-shot point cloud region annotation with human in the loop. *arXiv* 2019. https://arxiv.org/abs/1906.04409.

Kaijaluoto R, Kukko A, el Issaoui A. *et al.* Semantic segmentation of point cloud data using raw laser scanner measurements and deep neural networks. *ISPRS Open Journal of Photogrammetry and Remote Sensing* 2022;**3**:100011. https://doi.org/10.1016/j.ophoto.2021.100011.

Krisanski S, Taskhiri MS, Gonzalez Aracil S. *et al.* Sensor agnostic semantic segmentation of structurally diverse and complex forest point clouds using deep learning. *Remote Sens* 2021;**13**:1413. https://doi.org/10.3390/rs13081413.

Kulicki M, Cabo C, Trzciński T. *et al.* Artificial intelligence and terrestrial point clouds for forest monitoring. *Current Forestry Reports* 2024;**11**:5. https://doi.org/10.1007/s40725-024-00234-4.

Laino D, Cabo C, Prendes C. *et al.* 3DFin: a software for automated 3D forest inventories from terrestrial point clouds. *Forestry: An International Journal of Forest Research* 2024;**97**:479–96. https://doi.org/10.1093/forestry/cpae020.

Liang X, Qi H, Deng X. *et al.* ForestSemantic: a dataset for semantic learning of forest from close-range sensing. *Geo-Spatial Information Science* 2024;**28**:185–211. https://doi.org/10.1080/10095020.2024.2313325.

Lines ER, Allen M, Cabo C. *et al.* AI applications in forest monitoring need remote sensing benchmark datasets. In: *2022 IEEE International Conference on Big Data (Big Data)*. Piscataway, NJ: IEEE; 2022b, 4528–33. https://doi.org/10.1109/BigData55660.2022.10020772.

Lines ER, Fischer FJ, Owen HJF. *et al.* The shape of trees: reimagining forest ecology in three dimensions with remote sensing. *J Ecol* 2022a;**110**:1730–45.

Lowery DP, United States Forest Service. *Ponderosa Pine (American Wood)*. Forest Service, U.S: Department of Agriculture, 1984.

Luyssaert S, Inglima I, Jung M. *et al.* CO_2 balance of boreal, temperate, and tropical forests derived from a global database. *Glob Chang Biol* 2007;**13**:2509–37. https://doi.org/10.1111/j.1365-2486.2007.01439.x.

Ma L, Zheng G, Eitel JUH. *et al.* Improved salient feature-based approach for automatically separating photosynthetic and nonphotosynthetic components within terrestrial Lidar point cloud data of Forest canopies. *IEEE Trans Geosci Remote Sens* 2016;**54**:679–96. https://doi.org/10.1109/TGRS.2015.2459716.

Ma Z, Dong Y, Zi J. *et al.* Forest-PointNet: a deep learning model for vertical structure segmentation in complex forest scenes. *Remote Sens* 2023;**15**:4793. https://doi.org/10.3390/rs15194793.

McElhinny C, Gibbons P, Brack C. *et al.* Forest and woodland stand structural complexity: its definition and measurement. *For Ecol Manage* 2005;**218**:1–24. https://doi.org/10.1016/j.foreco.2005.08.034.

Mead DJ. *Sustainable Management of Pinus Radiata Plantations (FAO Forestry Paper No. 170)*. Rome: Food and Agriculture Organization of the United Nations (FAO); 2013. https://doi.org/10.13140/2.1.5173.0885.

Nguyen A, Le B. 3D point cloud segmentation: A survey. In: *IEEE Conference on Robotics, Automation and Mechatronics (RAM) Proceedings*. Piscataway, NJ: IEEE; 2013, 225–230. https://doi.org/10.1109/RAM.2013.6758588.

Oviedo de la Fuente M, Cabo C, Roca-Pardiñas J. *et al.* 3D point cloud semantic segmentation through functional data analysis. *J Agric Biol Environ Stat* 2024;**29**:723–44. https://doi.org/10.1007/s13253-023-00567-w.

Piermattei L, Karel W, Wang D. *et al.* Terrestrial Structure from Motion Photogrammetry for Deriving Forest Inventory Data. *Remote Sensing* 2019;**11**:950. https://doi.org/10.3390/rs11080950.

Prendes Pérez C, Cabo C, Ordóñez C. *et al.* An algorithm for the automatic parametrization of wood volume equations from terrestrial laser scanning point clouds: application in *Pinus pinaster*. *GIScience & Remote Sensing* 2021;**58**:1–21. https://doi.org/10.1080/15481603.2021.1972712.

Puletti N, Grotti M, Scotti R. Evaluating the eccentricities of poplar stem profiles with terrestrial laser scanning. *Forests* 2019;**10**:239. https://doi.org/10.3390/f10030239.

Puliti S, Pearse G, Surový P. *et al.* FOR-instance: a UAV laser scanning benchmark dataset for semantic and instance segmentation of individual trees (Versión 1) [data set]. *Zenodo* 2023. https://doi.org/10.5281/zenodo.8287792.

Qi CR, Su H, Mo K, Guibas LJ. PointNet: Deep learning on point sets for 3D classification and segmentation. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Piscataway, NJ: IEEE; 2017, 77–85. https://doi.org/10.1109/CVPR.2017.16.

Qiu H, Zhang H, Lei K. *et al.* Forest digital twin: A new tool for forest management practices based on spatio-temporal data, 3D simulation engine, and intelligent interactive environment. *Computers and Electronics in Agriculture* 2023;**215**:108416. https://doi.org/10.1016/j.compag.2023.108416.

Rehush N, Abegg M, Waser LT. *et al.* Identifying tree-related microhabitats in TLS point clouds using machine learning. *Remote Sens* 2018;**10**:1735. https://doi.org/10.3390/rs10111735.

RIEGL Laser Measurement Systems. (2022). RisCAN pro 2.0 [computer software]. RIEGL laser measurement systems GmbH. https://www.riegl.com.

RIEGL Laser Measurement Systems GmbH. RIEGL VZ-2000i. 2023a. Accessed December 5, 2024. http://www.riegl.com/nc/products/terrestrial-scanning/produktdetail/product/scanner/58/.

RIEGL Laser Measurement Systems GmbH. RIEGL VZ-2000i. 2023b. Accessed December 5, 2024. http://www.riegl.com/nc/products/terrestrial-scanning/produktdetail/product/scanner/48/.

Robert D, Raguet H, Landrieu L. Efficient 3D semantic segmentation with Superpoint Transformer. In: *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. Piscataway, NJ: IEEE; 2023, 17149–58. https://doi.org/10.1109/ICCV51070.2023.01577.

Schraik D, Hovi A, Rautiainen M. Crown level clumping in Norway spruce from terrestrial laser scanning measurements. *Agric For Meteorol* 2021;**296**:108238. https://doi.org/10.1016/j.agrformet.2020.108238.

Shen X, Huang Q, Wang X. *et al.* A deep learning-based method for extracting standing wood feature parameters from terrestrial laser scanning point clouds of artificially planted forest. *Remote Sens* 2022;**14**:3842. https://doi.org/10.3390/rs14153842.

Tao S, Guo Q, Xu S. *et al.* A geometric method for wood-leaf separation using terrestrial and simulated lidar data. *Photogrammetric Engineering and Remote Sensing* 2015;**81**:767–76. https://doi.org/10.14358/PERS.81.10.767.

Tian J, Li H, Sun X. *et al.* Quality assessment of shrub observation data based on TLS: a case of revegetated shrubland, southern Qinghai-Tibetan plateau. *Land Degradation and Development* 2023;**34**:1570–81. https://doi.org/10.1002/ldr.4554.

Tockner A, Gollob C, Ritter T. *et al.* LAUTx—individual tree point clouds from Austrian forest inventory plots [data set]. *Zenodo* 2022. https://doi.org/10.5281/zenodo.6560112.

UNESCO. *Wienerwald Biosphere Reserve*. Paris, France: UNESCO Publishing; Accessed November 26, 2024. https://www.unesco.org/en/mab/wienerwald.

Van den Broeck WAJ, Terryn L, Cherlet W. *et al.* Three-dimensional deep learning for leaf-wood segmentation of tropical tree point clouds. *International archives of the photogrammetry. Remote Sensing and Spatial Information Sciences* 2023;**XLVIII-1**:765–70. https://doi.org/10.5194/isprs-archives-XLVIII-1-W2-2023-765-2023.

Walter H. *Vegetation of the Earth and Ecological Systems of the Geobiosphere*. 2nd ed. Berlin, Germany: Springer-Verlag; 1979.

Weiser H, Schäfer J, Winiwarter L. *et al.* Terrestrial, UAV-borne, and airborne laser scanning point clouds of central European forest plots, Germany, with extracted individual trees and manual forest inventory measurements [data set]. *PANGAEA* 2022. https://doi.org/10.1594/PANGAEA.942856.

White JC, Coops NC, Wulder MA. *et al.* Remote sensing Technologies for Enhancing Forest Inventories: a review. *Can J Remote Sens* 2016;**42**:619–41. https://doi.org/10.1080/07038992.2016.1207484.

Xi Z, Hopkinson C, Rood SB. *et al.* See the forest and the trees: effective machine and deep learning algorithms for wood filtering and tree species classification from terrestrial laser scanning. *ISPRS Journal of Photogrammetry and Remote Sensing* 2020;**168**:1–16. https://doi.org/10.1016/j.isprsjprs.2020.08.001.

Xi Z, Chasmer L, Hopkinson C. Delineating and reconstructing 3D forest fuel components and volumes with terrestrial laser scanning. *Remote Sensing* 2023;**15**:4778. https://doi.org/10.3390/rs15194778.

Xiang B, Wielgosz M, Kontogianni T. *et al.* Automated forest inventory: analysis of high-density airborne LiDAR point clouds with 3D deep learning. *Remote Sens Environ* 2024;**305**:114078. https://doi.org/10.1016/j.rse.2024.114078.

Zhao H, Jiang L, Jia J, Torr P, Koltun V. Point Transformer. In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. Piscataway, NJ: IEEE; 2022, 16239–48.

Forestry: An International Journal of Forest Research, 2026, **99**, 1–17
https://doi.org/10.1093/forestry/cpaf062
Dataset paper