MAMBO

MODERN APPROACHES TO THE MONITORING OF BIODIVERSITY

D4.2 Assessment of interoperability and maturity for upscaling habitat condition metrics

27/02/2024

Lead beneficiary: University of Amsterdam (UvA)

Authors: W. Daniel Kissling, Yifang Shi, Jinhu Wang, Agata Walicka, Charles George, Jesper E. Moeslund & France Gerard

Reviewer/s: Christophe Dominik & Alexis Joly

Funded by the European Union

Prepared under contract from the European Commission
Grant agreement No. 101060639, EU Horizon Europe Research and Innovation Action

Project acronym:	**MAMBO**
Project full title:	**Modern Approaches to the Monitoring of Biodiversity**
Project duration:	01.09.2022 – 31.08.2026 (48 months)
Project coordinator:	Dr. Toke Thomas Høye, Aarhus Universityy (AU)
Call:	HORIZON-CL6-2021-BIODIV-01
Deliverable title:	D4.2 Assessment of interoperability and maturity for upscaling habitat condition metrics
Deliverable n°:	D4.2
WP responsible:	WP4
Nature of the deliverable:	Report
Dissemination level:	Public
Lead beneficiary:	UvA
Due date of deliverable:	M18
Actual submission date:	M18

Deliverable status:

Version	Status	Date	Author(s)
0.1	First outline	10/09/2023	W. Daniel Kissling (UvA)
0.2	Input outline	09/10/2023	W. Daniel Kissling, Yifang Shi, Jinhu Wang (UvA)
0.3	Input outline	06/11/2023	Agata Walicka & Jesper E. Moeslund (AU)
0.4	Input outline	08/12/2023	France Gerard (UKCEH)
0.5	Updated outline	22/12/2023	W. Daniel Kissling (UvA)
0.6	Final outline	04/01/2024	W. Daniel Kissling (UvA)
0.6	Input material	30/01/2024	France Gerard & Charles George (UKCEH), Jinhu Wang (UvA)
0.7	Input material	05/02/2024	Agata Walicka (AU)
0.8	First draft	09/02/2024	W. Daniel Kissling (UvA)
0.8	Comments	14/02/2024	Jesper E. Moeslund (AU)
0.8	Review	18/02/2024	Christophe Dominik (UFZ)
0.9	Complete draft	19/02/2024	W. Daniel Kissling (UvA)
0.9	Review	22/02/2024	Alexis Joly (INRIA), Christophe Dominik (UFZ)
0.9	Input material	22/02/2024	Jinhu Wang (UvA)
1.0	Final version	27/02/2024	W. Daniel Kissling (UvA)

Table of Contents

1 Preface

This document is a report (deliverable D4.2) prepared for the Modern Approaches to the Monitoring of BiOdiversity (MAMBO) project (Høye *et al.*, 2023), funded by the European Commission through an EU Horizon Europe Research and Innovation Action grant (No. 101060639). The MAMBO project aims to support EU biodiversity policy and address knowledge gaps by providing solutions to biodiversity monitoring through the design and development of novel tools and technologies. This report identifies the challenges for upscaling habitat condition metrics derived from LiDAR point clouds collected with crewed aircraft through (sub)national airborne laser scanning (ALS) surveys as well as Red, Green and Blue (RGB) and Near-Infrared (NIR) imagery and LiDAR point clouds collected with affordable unmanned aerial vehicles (UAV, i.e. drones). Such datasets can be used to describe the condition of habitats for site-specific (e.g. Natura 2000) monitoring, including metrics related to vegetation structure, habitat openness, topographic relief, hydrology, plant species composition, or plant biomass and abundance. The report analyses the characteristics of datasets for upscaling to larger areas (e.g. pre-classifications, point densities, acquisition seasons and additional RGB/NIR information of open access ALS LiDAR datasets) and the challenges for using drone-collected imagery and UAV point clouds for a consistent EU-wide habitat monitoring (e.g. lack of standardized reporting of flight surveys, (meta)data and data pre-processing). Besides identifying the challenges for upscaling habitat condition metrics, the report also provides an outlook for how the upscaling of habitat condition metrics could benefit from the development of a cloud-based virtual research environment (VRE) that provides virtual labs, repositories, metadata catalogues and other digital infrastructure services for a user-friendly, scalable, and distributed processing of huge volumes of LiDAR point clouds and UAV imagery across the EU.

2 Executive summary

This document explores critical challenges associated with upscaling habitat condition metrics (e.g., related to the biotic, abiotic and landscape characteristics of Natura 2000 sites), specifically focusing on data obtained from (sub)national airborne laser scanning (LiDAR) surveys and site-focused unmanned aerial vehicle (UAV) remote sensing. Challenges encompass a lack of standardization in UAV flight surveys, (meta)data generation, and pre-processing, as well as inconsistent characteristics in (sub)national LiDAR datasets across the EU, involving varying point densities, acquisition seasons, and pre-classification levels. This can impact the robustness and transferability of habitat condition metrics related to vegetation structure, habitat openness, topography, hydrology, and plant species composition and biomass.

The integration of deep learning methods with UAV imagery and LiDAR point clouds for tasks like shrub and tree identification and plant species composition estimation introduces challenges in generalizability, with most models being site-specific and limited in spatial extent. The 3D nature of LiDAR point clouds adds complexity to deep learning applications, requiring careful consideration of projection-based or point-based models, pixel or voxel size choices, and challenges with inconsistent pre-classifications, platform- and sensor-specific settings or environmental and site differences in heterogeneous datasets.

The processing of habitat condition information from LiDAR point clouds and RGB/NIR imagery involves handling massive datasets, requiring substantial computational and engineering resources, parallel and distributed processing, big data storage, and large computing capacity. Challenges include the need for adjustments in processing workflows depending on data characteristics, processing parameters, and infrastructure-specific choices regarding memory allocation and distribution among cores and workers of virtual machines.

To address these challenges, the report proposes the development of a cloud-based virtual research environment (VRE) enabling the creation of virtual labs for upscaling habitat condition metrics from airborne LiDAR and UAV remote sensing, which could enhance data discovery, access, and workflow execution. This would require centralized repositories, metadata catalogues, and additional services for provenance tracking, workflow management, and cloud automation, with existing EU infrastructures like LifeWatch ERIC, the European Open Science Cloud (EOSC), and the European Grid Infrastructure (EGI) potentially supporting such a VRE.

Tackling the outlined challenges provides a strategic pathway for advancing habitat monitoring in the EU. Imperative is the development and application of standards and best practices for data collection and pre-processing, the compilation of standardized and machine-readable metadata, and the advancement of VRE services and free and open-source software. This will contribute to harmonizing habitat condition information and help to mitigate the degradation and unsustainable use of natural resources in the EU.

3 List of abbreviations

ALS	Airborne Laser Scanning
API	Application Programming Interface
ASPRS	American Society for Photogrammetry and Remote Sensing
CNN	Convolutional Neural Networks
CRS	Coordinate Reference System
EU	European Union
FAIR	Findability, Accessibility, Interoperability, and Reusability
FOSS	Free and Open-Source Software
GPS	Global Positioning System
HPC	High-Performance Computing
LAS	File format for the interchange and archiving of lidar point cloud data
LAZ	An open file format to compress LAS data
LiDAR	Light Detection and Ranging
MAMBO	Modern Approaches to the Monitoring of Biodiversity
MIF	Minimum Information Framework
NIR	Near-Infrared
RGB	Red, Green and Blue
RS	Remote Sensing
SEEA	System of Environmental-Economic Accounting
UAV	Unmanned Aerial Vehicle
UN	United Nations
ViT	Vision Transformer
VM	Virtual Machine
VRE	Virtual Research Environment

4 Introduction

4.1 Habitat condition

A good condition of habitats is essential for maintaining biodiversity because the diversity, distribution and abundance of animals, plants and other organisms is tightly linked to the biotic, abiotic and landscape characteristics of an ecosystem, including vegetation structure, soil properties, hydrology, terrain features, microclimates, hedges, tree lines and the amount of dead wood. In this context, a good habitat condition enables the conservation of natural habitats with their wild fauna and flora, especially threatened species. Yet, habitat extent and condition continue to decline at alarming rates, which contributes substantially to the ongoing loss of biodiversity (Díaz *et al.*, 2019). In the EU, the vast majority of protected habitats shows an unfavourable conservation status, with a high level of degradation and unsustainable use of natural resources (European Environment Agency, 2020). Although some habitats show improvements, progress has not been sufficient to meet the objectives of the EU Biodiversity Strategy to 2020 (European Environment Agency, 2020). Hence, habitats in the EU continue to decline and face deteriorating trends from changes in land use, eutrophication, unsustainable management practices and other human-induced pressures.

The new EU Biodiversity Strategy for 2030 strives to be more successful and ensure that ecosystems are healthy, resilient to climate change, rich in biodiversity and able to deliver essential ecosystem services (European Commission, 2021). This requires to accurately assess and monitor the condition of habitats. Of particular interest in the EU are habitats with endangered, vulnerable, rare and endemic animal and plant species, and natural and semi-natural habitat types that are in danger of disappearing. This includes habitats with a small distributional range and those that are characteristic for a specific biogeographical regions in the EU (European Commission, 2021). Deriving habitat condition indicators often requires combining various metrics that capture abiotic characteristics, biotic characteristics, and landscape-level characteristics (Maes *et al.*, 2023). This approach is also applied to the ecosystem condition typology of the United Nations (UN) System of Environmental-Economic Accounting (SEEA) which aims at regular and standardised stocktaking on the extent and condition of ecosystems and their services (Czúcz *et al.*, 2021). Measuring and monitoring habitat condition therefore requires quantitative metrics that describe the abiotic, biotic, and landscape-level characteristics of ecosystems.

4.2 Remote sensing

Habitat condition monitoring in the EU is still mainly based on qualitative indicators or expert judgements (European Environment Agency, 2020). For instance, habitat condition is reported at the EU level with the qualitative description of ‘good’, ‘not-good’ or ‘unknown’, derived from national reports of EU member states which are based on expert knowledge and limited field surveys (Röschel *et al.*, 2020). However, significant advances in remote sensing technologies are now creating opportunities for more effective ways of habitat monitoring. Since free and openly accessible remote sensing data from satellites are often spatially too coarse for monitoring habitat condition at the site level, airborne Light Detection and Ranging (LiDAR) data and drone imagery from affordable Unmanned Aerial Vehicles (UAVs) offer a promising alternative for habitat condition assessments at scales useful to land managers and decision makers.

LiDAR and UAV remote sensing provide data at high spatial resolution and are increasingly available for habitat condition assessments. To date, LiDAR point clouds collected with crewed aircraft through (sub)national airborne laser scanning (ALS) surveys are increasingly becoming available in Europe. Such point clouds can be used to derive biotic, abiotic and landscape-level characteristics (Table 1), including high-resolution metrics of topography and vegetation structure at a country-wide extent (Kissling *et al.*, 2023; Assmann *et al.*, 2022; Moeslund *et al.*, 2023). In addition, at the site level, flight campaigns with UAVs offer an increasing range of observations, including multispectral, hyperspectral and thermal imaging as well as LiDAR point clouds. The most affordable and cost-effective UAVs record imagery in the Red, Green and Blue (RGB) and Near-Infrared (NIR) spectrum at mm to cm resolution, enabling the mapping of specific plant species, vegetation structure, and abiotic characteristics (Table 1). Hence, both LiDAR point clouds and UAV imagery offer promising opportunities for assessing habitat condition in protected areas and other sites of community importance.

Table 1: Examples of using Light Detection and Ranging (LiDAR) and imaging data from Unmanned Aerial Vehicles (UAVs) to quantify habitat condition.

Habitat condition	Examples	References
LiDAR		
Biotic characteristics	Canopy height and cover, vertical and horizontal variability of vegetation, understory density	Bakx *et al.* (2019); Davies & Asner (2014); Koma *et al.* (2021)
Abiotic characteristics	Soil moisture, hydrology, elevation, aspect, slope	Moeslund *et al.* (2013); Assmann *et al.* (2022); (Davies & Asner, 2014)
Landscape characteristics	Hedges & tree lines, tree inventories, amount of dead wood, patchiness of open areas, edge extent	Lucas *et al.* (2019); Graham *et al.* (2019); Wang, Lindenbergh & Menenti (2018b); Abrego *et al.* (2021); Marchi, Pirotti & Lingua (2018); Martinuzzi *et al.* (2009); de Vries *et al.* (2021)
UAV		
Plant species mapping	Invasive, rare or protected species	Oldeland *et al.* (2021); Hill *et al.* (2017); Zhang *et al.* (2020); James & Bradshaw (2020)
Biotic characteristics	Understory biomass, woody plant distribution, vegetation height, woody plant volume and biomass	Zhang *et al.* (2022); Olariu *et al.* (2022); van Iersel *et al.* (2018); Cunliffe, Brazier & Anderson (2016)
Abiotic characteristics	Bare ground, micro-topography, soil moisture and wetness	Lendzioch *et al.* (2021); Ikkala *et al.* (2022); Eischeid *et al.* (2021); Barnas *et al.* (2019)

4.3 FAIR data principles

While hardware and software for collecting and processing LiDAR point clouds and UAV imagery have become much more accessible, an urgent need remains to improve data processing, interoperability, and the reuse of data. In this context, the FAIR data principles (Findability, Accessibility, Interoperability, and Reusability) act as a guideline for enabling the reusability of data and enhancing the ability of machines to automatically find and use the data (Wilkinson *et al.*, 2016). This includes, among other things, human- and machine-readable metadata, standard workflows for preparing, publishing, and preserving data products, and open and free access to raw data and derived data products (Hardisty *et al.*, 2019). Applying the FAIR data principles for an improved data management can lead to knowledge discovery and innovation, and to subsequent data and knowledge integration and reuse by the broader community.

4.4 Aims and objectives

In this report, we synthesize key challenges for upscaling habitat condition metrics derived from LiDAR point clouds and drone imagery for an EU-wide habitat monitoring. This includes assessing the characteristics and FAIR data principles of open access LiDAR point clouds collected with airborne surveys, and RGB and NIR imagery and LiDAR point clouds collected with UAVs. We focus on the condition monitoring of natural and semi-natural habitats in the EU (e.g., Natura 2000 sites) which support endangered, vulnerable, rare, endemic and indicator animal and plant species. Besides identifying the challenges for upscaling habitat condition metrics, the report also provides an outlook for how those challenges could be addressed, proposing the development of a cloud-based virtual research environment (VRE) enabling the creation of virtual labs for upscaling habitat condition metrics from airborne LiDAR and UAV remote sensing.

5 Challenges for upscaling habitat condition metrics

5.1 Standardising drone surveys

For drone surveys, there is still much work to be done in developing standards or best practices that address the complex data pipeline typical of an UAV project, from raw data collection to derived products (Wyngaard *et al.*, 2019). Major challenges include the lack of standardisation of flight surveys, little guidance for standardised data generation, absence of standardised metadata, and the use of propriety software during the data pre-processing (Figure 1). For each flight survey with an UAV, decisions have to be made about the survey design and which sensor-platform technology to use (Figure 1). Sensor-platform technology is developing fast, and hardware and software of different sensor generations and platforms are rarely compatible (Zhang & Zhu, 2023). Moreover, decisions about sampling designs vary widely because the survey objectives and the sensor-platform technology vary among UAV campaigns. Currently, there is no standard approach or best practice for sensor use procedures, such as mounting requirements on different platforms, sample rates, altitude, flight patterns and ground observations for calibration.

Similarly, there is a lack of standards and guidelines for generating (meta)data describing calibration, training and validation procedures during the UAV campaign and how to report the flight parameters in a standardised way (Barbieri *et al.*, 2023). This potential lack in consistency in observations and calibration can significantly affect the quality of repeat

surveys and thus the monitoring of habitat condition. For instance, the parameters set by the pilot for the flight (e.g., flight line overlap, height, flight direction and terrain following) or other characteristics during the flight survey (e.g., sun light, flight time and cloud cover) can have large implications for multi-site and multi-temporal comparisons. Metadata of survey variables are currently not reported in a standardized and machine-readable way. A set of core metadata should include, for instance, wind speed, cloud cover, flight height, flight speed, flight pattern, sensor information, and the date and time. This would support the subsequent pre-processing and data analysis in a comparable way.

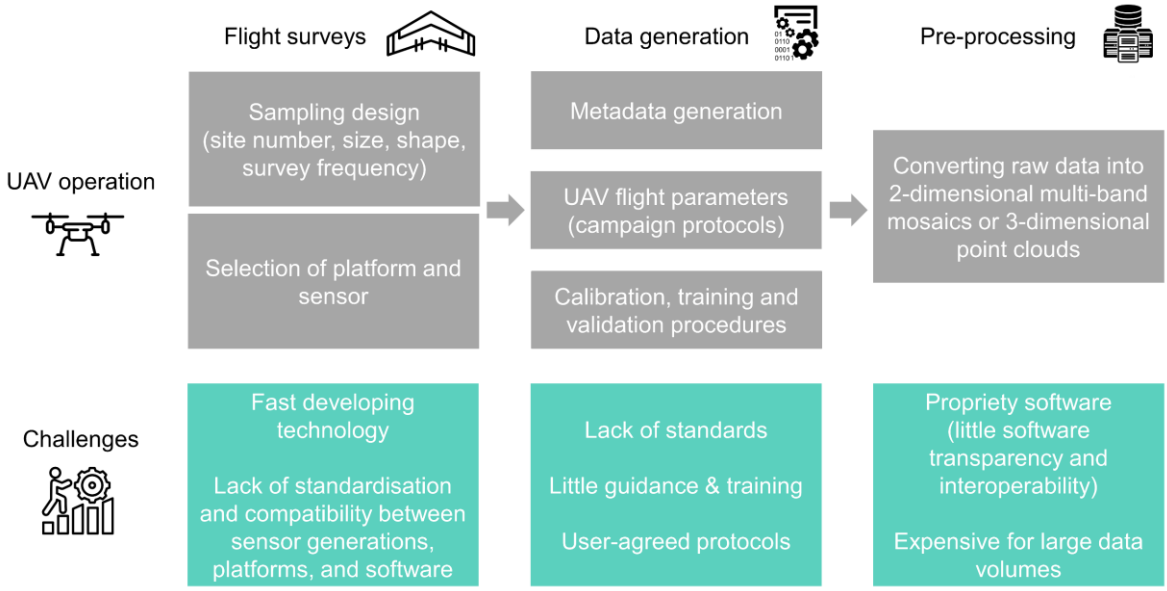

Figure 1: Key aspects of unmanned aerial vehicle (UAV) operation (grey boxes) and related challenges for standardisation (green boxes) during flight surveys, data generation and pre-processing.

For LiDAR point clouds collected with UAV, each LiDAR sensor manufacturer stores the data on the instrument in its own way, which is then converted to a standard format using the company's own software. In case of RGB and NIR imagery, Structure from Motion (when a large number of highly overlapping image frames are combined to deliver 2-dimensional image mosaics and 3-dimensional point clouds) often relies on propriety software provided by the sensor-platform provider. Several of these software packages at least allow some control (Jiang, Jiang & Jiang, 2020), although much of the modelling is still a black-box. For cloud-based solutions, generally very little control is given to the user.

5.2 Characteristics of (sub)national LiDAR point clouds

In contrast to UAV-collected data, many national and subnational LiDAR point clouds collected with airplanes through ALS surveys are becoming freely available across Europe (Figure 2a,b). These LiDAR point clouds are often made accessible via a website or an institutional repository within a country (Annex Table A1). The point clouds are typically provided as LAS or (compressed) LAZ files, i.e., a standard open file format designed for the interchange and archiving of LiDAR point clouds. This file format has been developed by the American Society for Photogrammetry and Remote Sensing (ASPRS) and adopted as a standard by the Open Geospatial Consortium (OGC) (ASPRS, 2019). The LAS/LAZ format provides standardized information on the Coordinate Reference System (CRS), data types,

header block, and several attributes of the point cloud (ASPRS, 2019). For instance, attributes such as X, Y, and Z values of each point, intensity values (i.e., the magnitude of the pulse return), the pulse return number and number of returns, scan direction and scan angle, GPS time, and the edge of the flight line are reported. In addition, a classification field can provide information on the attributes of each point in case a pre-classification has been performed before the data provisioning to the public. Examples of such pre-classification attributes are 'ground', 'bridge', 'water', 'building', and 'unclassified' (Figure 2c). The coding and definitions of these attributes typically follow the ASPRS standard point classes (ASPRS, 2019). In addition, information on RGB or NIR image channel values or waveforms can be reported if they are simultaneously recorded by the deployed sensor.

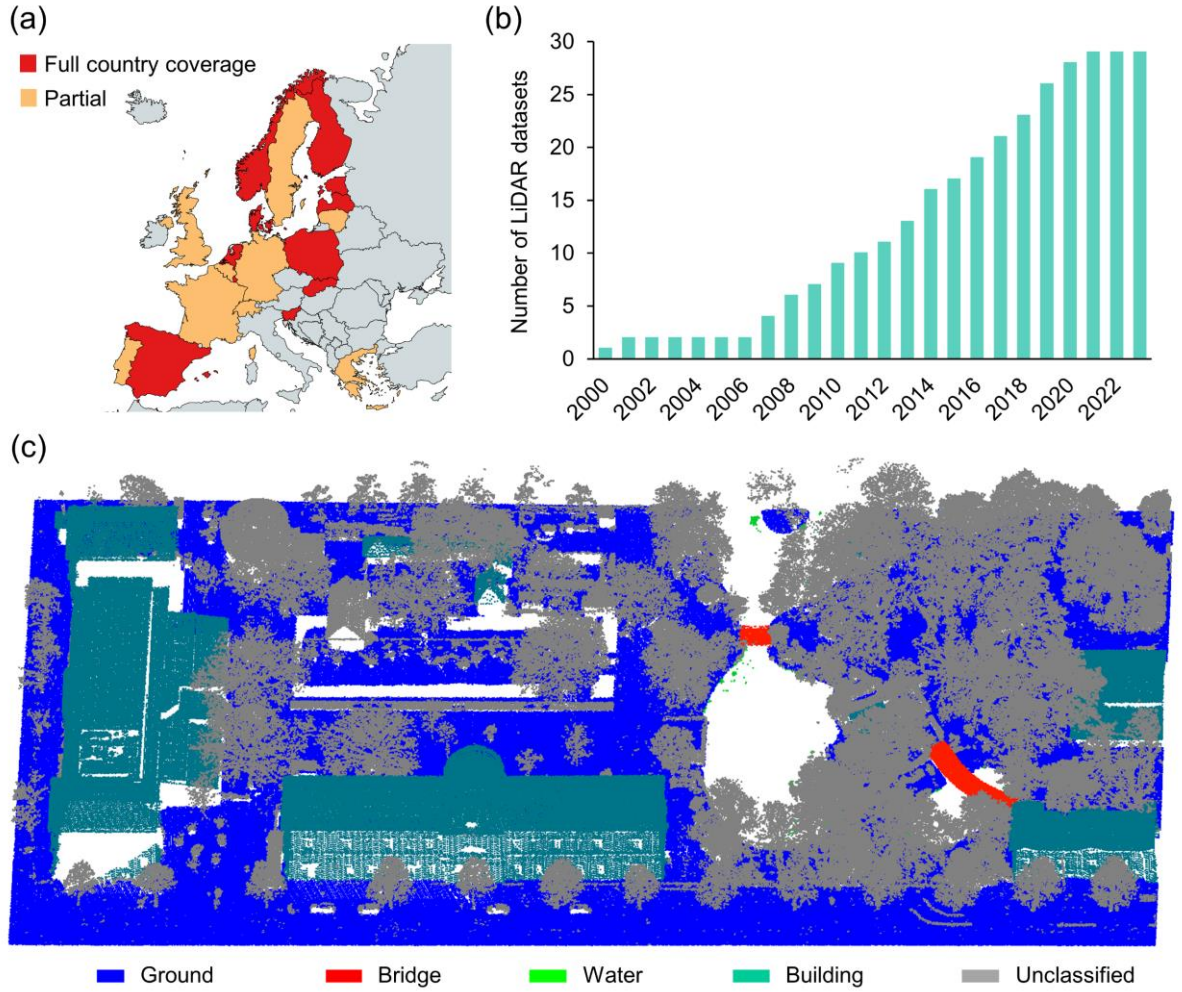

Figure 2: Coverage of open-access (sub)national LiDAR point clouds in Europe. (a) Full or partial LiDAR coverage per country (see details in Annex Table A1). (b) Cumulative number of datasets across Europe (datasets listed in Annex Table A1). (c) Example of a LiDAR point cloud from the Dutch airborne laser scanning survey (AHN4) showing the coverage of pre-classification attributes (i.e., standard point classes). White areas indicate areas with no data (e.g., laser pulse absorption in water, building fronts, below cars).

While the LAS/LAZ format provides standardized metadata for national and subnational LiDAR point clouds, the datasets from different countries are obtained with different budgets, for different purposes and with different requirements depending on each

country's needs. Consequently, such datasets still come with a wide variety of characteristics (Figure 3). For instance, (sub)national LiDAR datasets vary widely in terms of point densities, season of data acquisition (i.e., whether collected in the leaf-on or leaf-off season) and whether they provide additional RGB/NIR information or not (Figure 3a–c). This creates challenges for the consistent calculation of habitat condition metrics with different LiDAR datasets, e.g., datasets from different countries or from different flight campaigns within the same country. Point densities from open-access (sub)national LiDAR point clouds in Europe range from 1–30 points/m^2, with the majority having a density range around 5–10 points/m^2 (Figure 3a). Most datasets are collected during the leaf-off season (i.e., autumn, winter and early spring), but some also during the leaf-on season (Figure 3b). Only a few open-access LiDAR point clouds provide additional RGB or NIR information (Figure 3c). Sometimes this information is recorded as RGB and sometimes as external bytes, and different versions of LAS data formats can provide different information (ASPRS, 2019). In many cases, the pre-processing of the LiDAR point clouds and the quality control procedures are not accurately documented because they are done by companies and not described in scientific publications. For instance, information about the dates of data acquisition and the equipment used are often not found in the documentations, and the accuracy and exact methodology of the pre-classification is usually not reported.

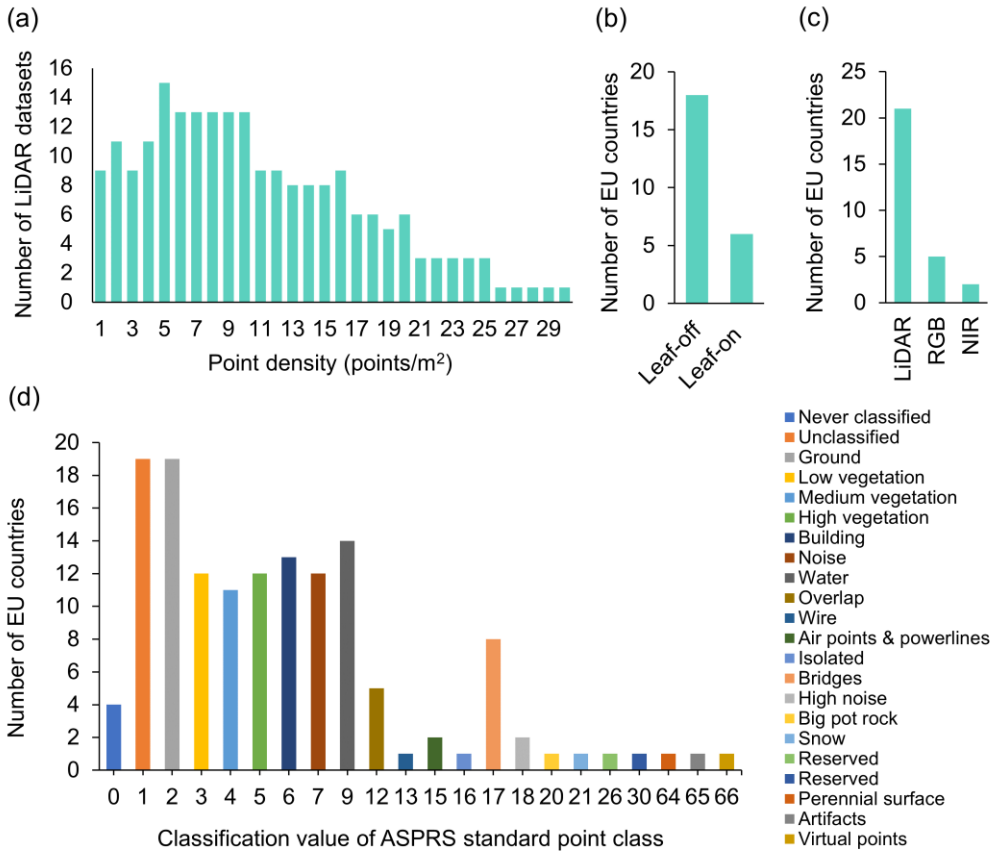

Figure 3: Variation in characteristics of LiDAR point clouds collected through (sub)national airborne laser scanning surveys in Europe (see details in Annex Table A1). (a) Range of available point densities. (b) Season of data acquisition (leaf-on vs. leaf-off). (c) Available information on LiDAR and simultaneously collected RGB or NIR. (d) Available pre-classifications following the ASPRS standard point classes. See Annex Table A1 for details of datasets.

Almost all publicly available LiDAR datasets from ALS surveys provide some pre-classification for each point (Annex Table A1), but the level of detail in the pre-classification varies widely among the different datasets (Figure 3d). Two of the classes ('ground' and 'unclassified') are almost always provided (Figure 3d) because terrain mapping is the intended purpose of many national ALS flight campaigns. This can be sufficient for quantifying habitat condition in relation to abiotic characteristics such as elevation, aspect and slope, although the 'ground' class may also unintentionally include short-stature vegetation such as grasses (Figure 4). A few other classes (e.g., vegetation, building, noise and water) are also commonly provided (i.e., in >50% of the datasets, Figure 3d), but the details and accuracy of those classes vary widely. This can influence habitat condition mapping, e.g. the calculation of vegetation structure metrics derived from national ALS datasets (Kissling *et al.*, 2023). For instance, Finland categorizes vegetation into low, medium, and high; Estonia only identifies tall vegetation; Switzerland includes a general vegetation class; and Scotland, Belgium, and the Netherlands do not provide any vegetation class (Annex Table A1). When vegetation points are not provided in a specific class, they are typically included in the class 'unclassified' which can also contain points from other objects such as transmission towers, traffic signs, trucks and cars (Figure 4). This can introduce biases and errors when quantifying vegetation structure if one assumes that the class 'unclassified' only contains vegetation points (Kissling *et al.*, 2023). For instance, powerlines and transmission towers are only rarely pre-classified (Figure 3d) and may cause erroneously high vegetation height values in LiDAR metrics capturing the canopy height of ecosystems (Shi & Kissling, 2023).

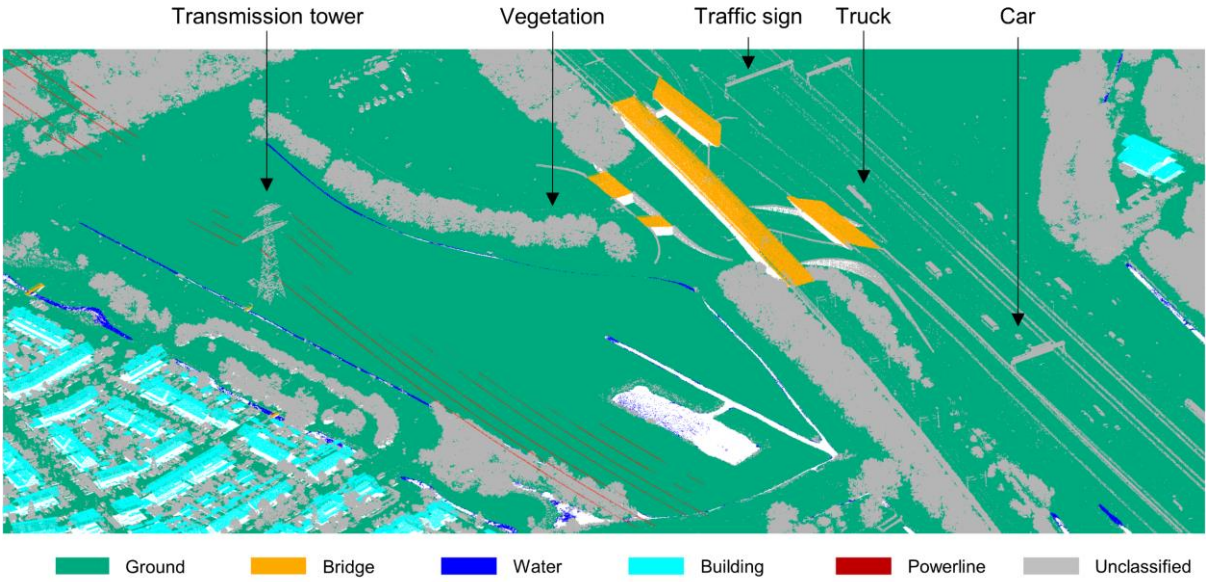

Figure 4: Example of a LiDAR point cloud (AHN4 dataset from the Netherlands) in which vegetation and non-vegetation objects (e.g., transmission towers, traffic signs, trucks and cars) are included in the same class ('unclassified', grey). The 'ground' class (green) probably includes not only ground points (i.e., terrain) but also some short-stature vegetation (e.g., grasses). White areas indicate areas with no data (e.g., laser pulse absorption in water, building fronts, below bridges).

5.3 Metric robustness and transferability

One of the most important characteristics of a LiDAR point cloud is its point density. Point densities of publicly available LiDAR datasets in Europe vary widely (Figure 3a) due to

differences in scanning geometry (multiple flight lines and different scanning directions), capability of the used laser scanner, and the season of data acquisition (leaf-on or leaf-off). Point densities also vary between ALS surveys from different time periods in the same country because laser scanner technology is rapidly developing and new ALS surveys tend to have improved sensors with better scanning capabilities (i.e., higher point densities). This can affect the calculation of biotic, abiotic and landscape characteristics and the monitoring of other habitat condition over time (i.e., change detection). Nevertheless, not all metrics might be affected in a similar way by changes in point densities (Figure 5). Moreover, the robustness and transferability of metrics might also depend on the volume geometry at which they are calculated, such as the spatial resolution (e.g., 1 × 1 m or 10 × 10 m) of the grid cell that defines the neighbourhood for metric calculation (Meijer *et al.*, 2020).

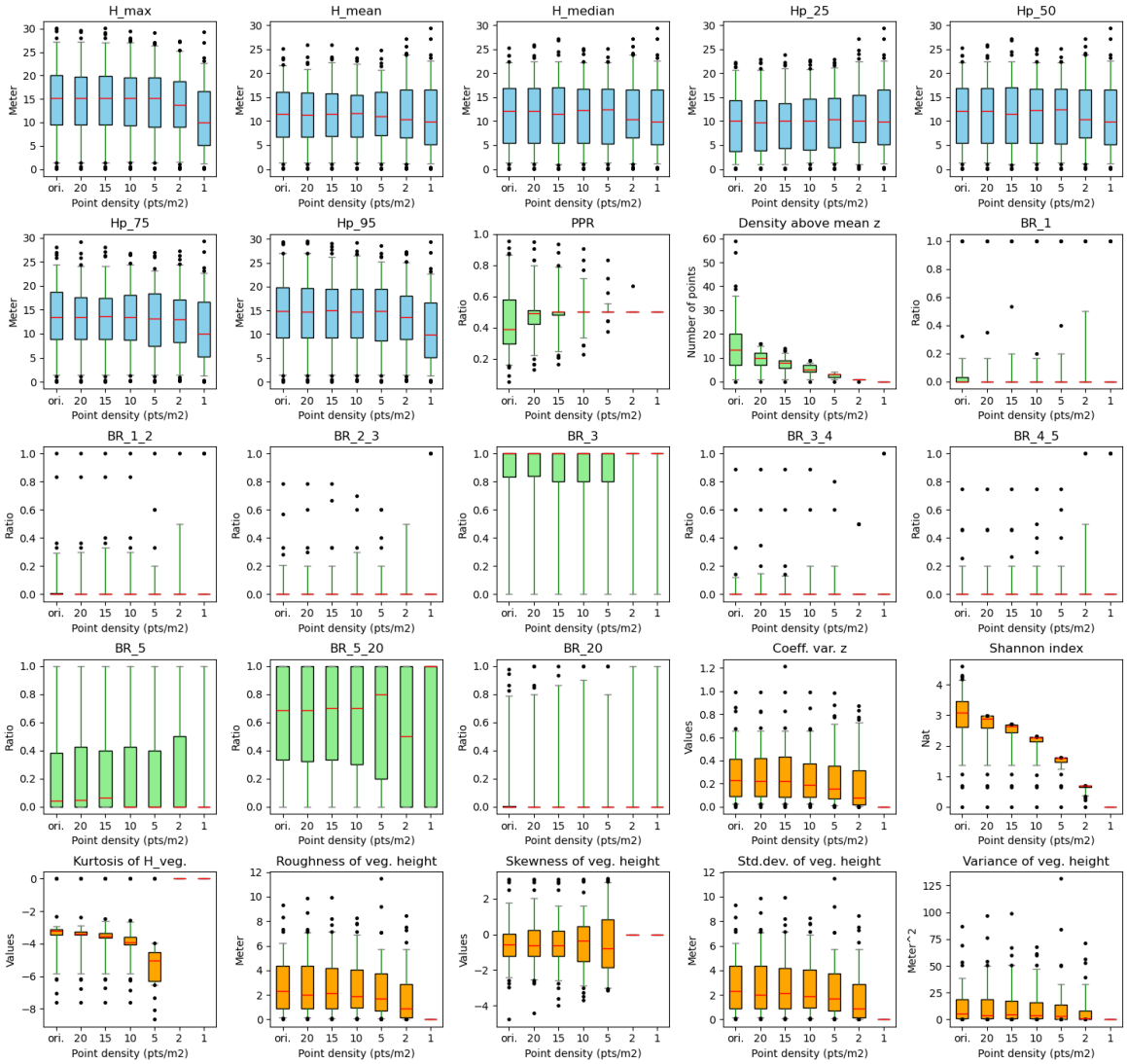

Figure 5: Twenty-five LiDAR vegetation metrics (blue: height metrics, green: cover metrics, orange: structural complexity metrics) calculated with different point densities across plots of 1 × 1 m resolution (*n* = 94) in woodland habitats of the Netherlands. Point densities were down-sampled from the original Dutch AHN4 dataset ('ori.') to six lower point densities (note the inverted x-axes). Boxes represent the interquartile range, horizontal red lines the medians, whiskers extend to the 5th and 95th percentiles, and outliers are plotted as dots. See Annex Text A1 for methodological details and Annex Table A2 for metric explanations.

To assess how different point densities may affect the robustness and transferability of metrics, a large diversity of LiDAR vegetation metrics and terrain features could be calculated (Bakx *et al.*, 2019; Davies & Asner, 2014; Assmann *et al.*, 2022; Moeslund *et al.*, 2019). To exemplify the effect, we focus here on a set of 25 LiDAR vegetation metrics which represent different aspects of vegetation height, vegetation cover and structural complexity (e.g., vertical variability) of vegetation (Kissling *et al.*, 2023; Kissling & Shi, 2023). The 25 LiDAR metrics were calculated with a workflow that combines shapefiles from Natura 2000 sites with LAZ files from LiDAR point clouds (Annex Figure A1). We selected Natura 2000 sites with woodlands in the Netherlands (Annex Figure A2) and calculated metrics using the original point density of the Dutch AHN4 dataset (20–30 points/m^2) as well as six randomly down-sampled point clouds (20, 15, 10, 5, 2 and 1 points/m^2, respectively) for the same locations (see methods in Annex Text A1). Results revealed that the 25 LiDAR metrics are not equally robust to varying point densities (Figure 5). Vegetation-height-related metrics were mostly robust to varying point densities (blue plots in Figure 5) whereas vegetation-cover-related metrics and structural complexity metrics were often sensitive to low point densities (green and orange plots in Figure 5).

Especially vegetation cover metrics derived from LiDAR point clouds might be strongly affected by varying point densities because they depend per definition on the density of vegetation points (Annex Table A2). Hence, a low point density affects the robustness of those metrics (green plots in Figure 5). For example, the pulse penetration ratio (PPR, Figure 5), a metric to quantify the openness of vegetation, tends to imply that woodlands are more open if point densities decrease. The density above the mean vegetation height, a measure of the density of the upper vegetation layer, is also strongly decreasing if point densities decrease (Figure 5). Other vegetation cover metrics, such as height layer ratios (measuring the density of vegetation in certain height layers), vary in their response to point densities depending on which vegetation layer is considered, especially at point densities <5 points/m^2. Besides cover metrics, metrics of vegetation structural complexity (capturing the vertical distribution of biomass) can also be affected by varying point densities (orange plots in Figure 5). Some structural complexity metrics are relatively robust if point densities are >5 points/m^2, including the coefficient of variation, roughness, skewness, standard deviation and variance of vegetation height (Figure 5). Other structural complexity metrics, such as the Shannon index (measuring the evenness of vertical vegetation distribution), strongly decrease with lower point densities. The majority of European open-access LiDAR point clouds derived from ALS surveys have a point density of 5–10 points/m^2 (Figure 3a). However, it is important to realize that many points are related to ground, water, buildings or other human infrastructures, so that the actual density of vegetation points is much lower. Calculating the metrics at a coarser resolution (e.g. 10 × 10 m instead of 1 × 1 m) will result in a higher number of points available for metric calculations and thus increases the robustness of metrics to varying point densities (see Annex Figure A3 for examples of the 25 metrics calculated at 10 × 10 m resolution). However, this also comes at the cost of losing details at high resolution, such as the fine-scale distribution of shrubs and open patches.

5.4 Generalizability of deep learning methods

As in other research fields, deep learning methods are increasingly applied to UAV imagery and LiDAR point clouds in the context of ecological applications, e.g., for identifying and classifying individual shrubs and trees (Allen *et al.*, 2022; Allen *et al.*, 2023; Fan *et al.*, 2023),

mapping tree crowns (Weinstein *et al.*, 2020), estimating forest tree species composition (Murray *et al.*, 2024), for semantic class labelling (e.g. tree, shrub, grass, low vegetation) in the context of point cloud classification (Zhao *et al.*, 2021; Yousefhussien *et al.*, 2018; Widyaningrum *et al.*, 2021), or when removing powerline noise for calculating vegetation metrics (Shi & Kissling, 2023). While these developments are promising, most of the development and testing of these algorithms is done at single sites or with single datasets, and only few models are trained and evaluated with datasets from multiple sites or across broad spatial extents (Figure 6a). Hence, current model training and testing is insufficient for generalizing deep learning methods for ecological applications and habitat condition monitoring, partly because a large amount of training data is currently not easily (or openly) accessible.

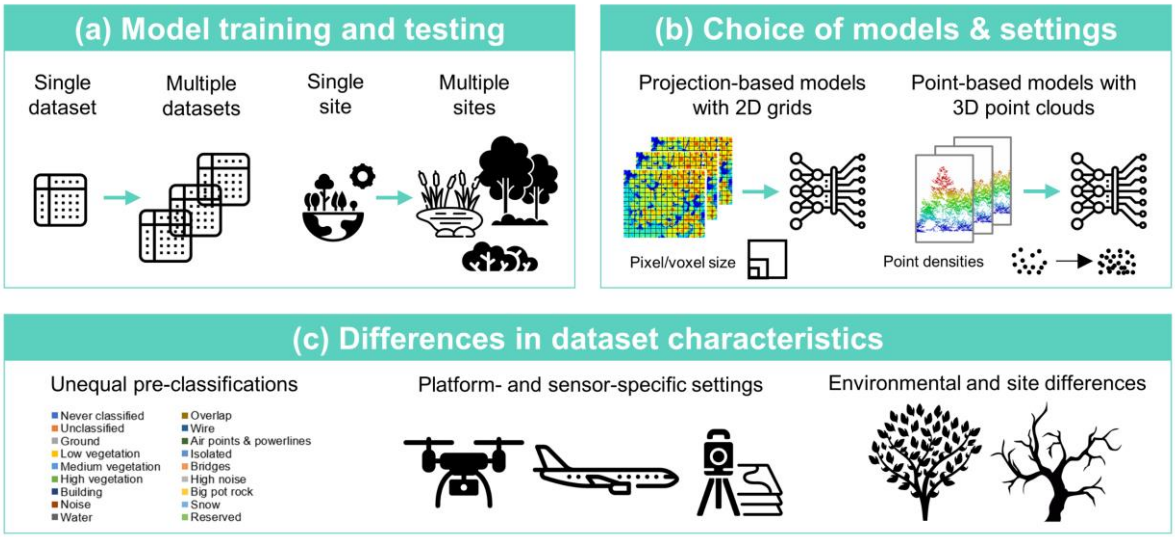

Figure 6: Challenges for generalizing deep learning methods to large-scale or multi-site habitat condition applications. (a) Lack of training and testing data from multiple datasets or multiple sites. (b) Choice of deep learning model type and settings. (c) Differences in dataset characteristics.

Another key challenge for generalizing deep learning methods is the type of deep learning model that is applied (Figure 6b). The 3D nature of LiDAR point clouds means that the data cannot be directly processed using standard convolutional neural networks (CNNs) or vision transformers (ViTs) that are deployed for 2D grid-based datasets (e.g. RGB images). The LiDAR data have to be either regularized by projecting the 3D point cloud into 2D pixels or voxels (projection-based models) or the CNNs and ViTs have to be designed specifically for 3D data processing (point-based models). For projection-based models, the choice of the pixel or voxel size (which depends on available point densities) has to be defined before model training and might subsequently affect the accuracy of the deep learning models when applied to other datasets (Xi *et al.*, 2020). For point-based models, differences in point densities between datasets can affect the generalizability of deep learning methods from one dataset to another (Chen, Chen & Liu, 2021).

Besides point densities, other differences in dataset characteristics can influence the generalizability of deep learning models (Figure 6c). For instance, using the pre-classification of an ALS point cloud dataset from one country for model training and applying the model to a dataset from another country can be highly accurate for a range of classification categories

(e.g., ground, vegetation, powerlines and buildings), but may lead to misclassifications in other categories that vary in shape among countries (e.g., transmission towers) (Shi & Kissling, 2023). Similarly, platform- and sensor-specific settings or environmental and site differences will affect the generalizability of deep learning models (Figure 6c). Scanner settings (e.g., scanning angles, scanner maximum range, scanner pulse repetition rate) can affect basic characteristics of LiDAR point clouds (Brede *et al.*, 2022), differences in the normalization of intensity values can affect deep learning models (Murray *et al.*, 2024), and data from different seasons might require to remove specific details (e.g. leaves of deciduous trees) before model training (Wang *et al.*, 2018a). Data acquired with different sensor platforms —such as airborne, mobile, terrestrial or UAV laser scanners— will differ not only in their data structures (e.g., level of details captured from trees), but also most likely deviate in the acquisition season. The generalizability of species recognition models also suffers from geographic differences in species composition. Most deep learning models focus on the recognition of a few species that are specific for the location of the training data (Fan *et al.*, 2023). Such deep learning models will not successfully generalize to other species that are absent in the training samples.

5.5 Computational challenges

To derive biotic, abiotic and landscape characteristics from LiDAR point clouds or RGB/NIR imagery from UAVs, massive (e.g., multi-terabyte) datasets need to be processed, especially for country-wide assessments (Assmann *et al.*, 2022; Kissling *et al.*, 2023). Given the large data volumes (examples in Annex Table A1), the processing is computationally demanding and requires parallel and distributed processing (Meijer *et al.*, 2020). As an example, processing of the country-wide LiDAR data (AHN3) of the Netherlands required handling ~16 TB data volume with ~700 billion points (Kissling *et al.*, 2022). Processing this dataset into 25 metrics of vegetation structure at 10 m resolution across the whole Netherlands took 294 days of total central processing unit (CPU) time (i.e., 14 days total wall-time), using a high-throughput workflow on a cluster of virtual machines (VMs) with fast CPUs and high memory nodes within the Dutch national IT infrastructure SURF (Kissling *et al.*, 2022). This workflow includes typical steps for LiDAR point cloud processing, such as re-tiling, normalization, feature extraction and rasterization (Figure 7). To perform the processing, all 1,367 LAZ files (each of 5 km × 6.25 km size and with an individual data volume between 0.3 MB to 6 GB) first had to be downloaded from the national AHN repository and then split and re-tiled into 37,457 tiles of 1 km × 1 km size, due to memory constrains for the VMs (Kissling *et al.*, 2022). Processing such massive datasets therefore comes with many computational challenges that are typical for big data (Figure 7).

Handling massive datasets requires big data storage, sufficient computing and engineering resources, scheduling of virtual machines, and parallelization and distribution of tasks (Kissling *et al.*, 2022; Meijer *et al.*, 2020). Not every IT-infrastructure could provide sufficient data storage and computing capacity for such a workflow. Depending on the available infrastructure, data storage could differ between remote and local storage, and the speed/connection of accessing the data storage is crucial for processing efficiency. It is generally recommended that the data storage is close to the processing. Besides data storage, the way of parallelization and distribution within the workflow may also differ from one infrastructure to another. For instance, deploying the workflow with a cluster of VMs on a national IT infrastructure requires to define how the tasks are distributed among workers of the cluster (Kissling *et al.*, 2022). When deploying the same workflow on a remote cloud

infrastructure, the parallelization has to be done using additional modules to avoid performance bottlenecks (Zhao *et al.*, 2022; Wang *et al.*, 2022). Different IT-infrastructures will also provide different computing capacity and resources. Three key factors of the capacity are the number of workers/VMs, the available cores per worker/VM, and the memory capacity. Configuration adjustments have to be made based on the input data (e.g., volume), required output (e.g., spatial resolution, number of metrics), and the availability of computing resources within a given IT infrastructure. This requires a good knowledge of the input data, the workflow, and the infrastructure, as well as thorough testing of the (time) efficiency of each workflow step during the processing.

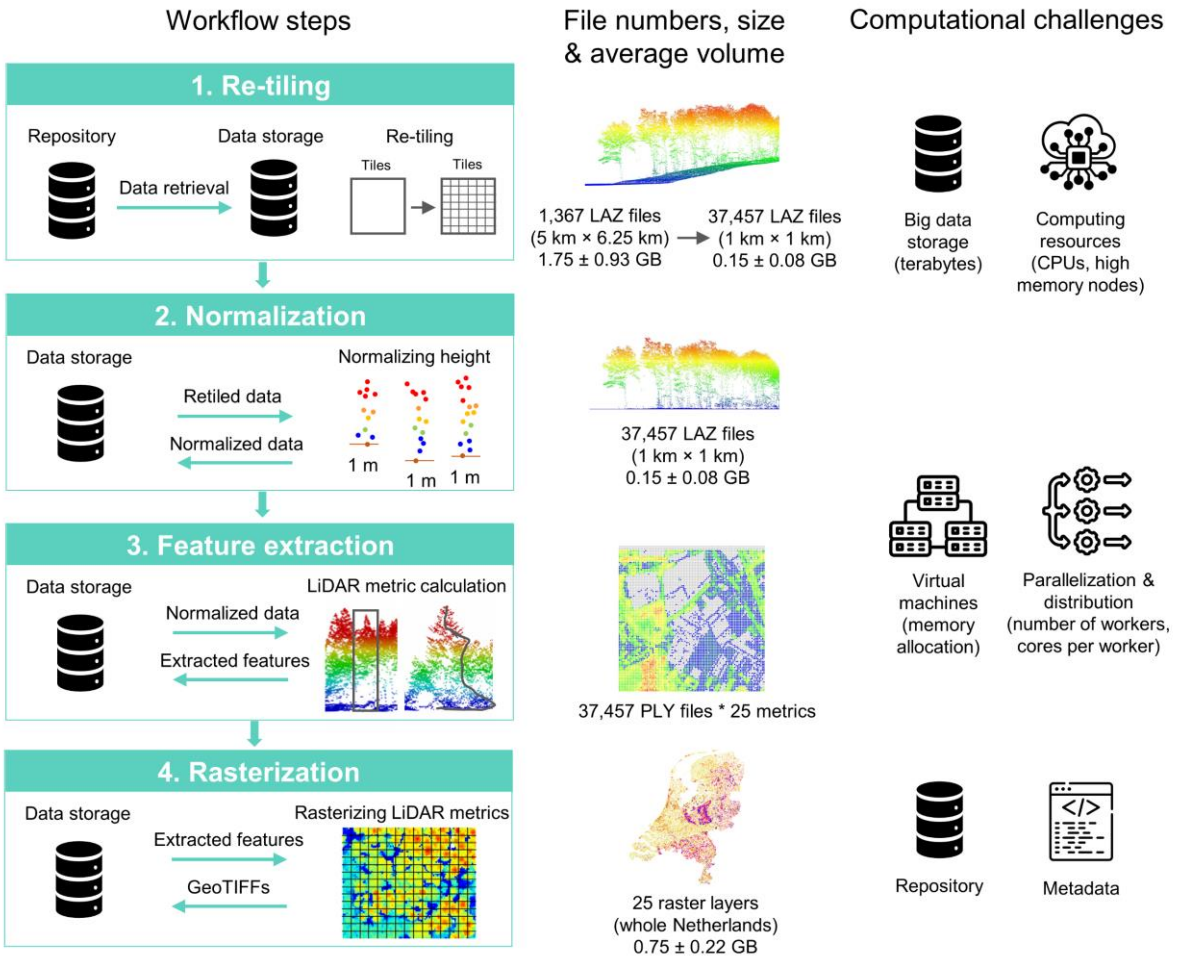

Figure 7: Example of a high-throughput LiDAR workflow for generating geospatial data products of vegetation structure from airborne laser scanning point clouds. The number, sizes and volumes of files creates various computational challenges, illustrated with the processing of a country-wide airborne laser scanning dataset of the Netherlands (see details in Kissling *et al.*, 2022).

Different data characteristics (e.g., data volumes of tiles, data format, point density, pre-classification, and coordinate system) from different data sources require different processing steps and parameters. For instance, when a single raw file is too large to be handled for a given cluster/worker, splitting the big chunk of files into smaller ones should be considered as the first step of the process. Moreover, the parameters in the workflow have to be adjusted based on the specific pre-classification code of the target data source.

For example, for the Dutch AHN3 data, vegetation belongs to the classification code 'unclassified', while the Spanish ALS data have the vegetation classified as 'low vegetation', 'medium vegetation', and 'high vegetation'. Similarly, during the re-tiling step, the employed grid setting (X, Y coordinate boundaries) need to be adjusted for the specific coordinate system and the coverage of the target data source.

6 Outlook

6.1 Cloud-based virtual research environments

The upscaling of habitat condition metrics for a consistent EU-wide habitat monitoring would benefit from a cloud-based virtual research environment (VRE) that enables the creation of application-specific virtual laboratories (e.g., airborne LiDAR or UAV virtual labs), providing user-centric support for data discovery and access and for composing and executing workflows (Figure 8). Such a VRE would include a catalogue of research assets (to search datasets, software and algorithms), a workflow management system, a data management framework, and tools for enabling user collaboration (Zhao *et al.*, 2022). The development of virtual labs for airborne LiDAR and UAV remote sensing would enable users to scale up the consistent retrieval of habitat conditions metrics from multi-national LiDAR point clouds and multi-site drone data (LiDAR and imagery) collected across the EU. Building a VRE will most likely be a success if the implemented tools and programming languages are familiar to the users (e.g., ecologists, data scientists and remote sensing experts). For instance, a range of free and open-source software (FOSS) tools for LiDAR and UAV image processing are now available, including CloudCompare (https://www.danielgm.net/cc/), OpenDroneMap (https://www.opendronemap.org/), the Point Data Abstraction Library (PDAL, https://pdal.io/en/2.6.0/), the Geospatial Data Abstraction Library (GDAL, https://gdal.org/), the R package LidR (Roussel *et al.*, 2020), the Python tool Laserchicken (Meijer *et al.*, 2020), and Jupyter Notebooks of the high-throughput Laserfarm workflow (Kissling *et al.*, 2022). When developing a VRE, such notebook environments (e.g., R-Studio and Jupyter) and popular programming languages (e.g., Python, R, C++ and Julia) should be taken into account (Zhao *et al.*, 2022). For less-programming oriented users, simplified and user-friendly interfaces should be available for workflow execution, e.g., by encapsulating different workflow steps as dockerized services with file-based input and output (Zhao *et al.*, 2022).

6.2 Repositories

An important step is to make existing LiDAR point clouds and drone imagery across the EU better findable and accessible. For ALS LiDAR, a recent European Commission report (Kakoulaki, Martinez & Florio, 2021) has summarized available LiDAR point clouds, digital surface models (DSMs) and digital terrain models (DTMs) in Europe, but the availability of datasets changes quickly as more flight campaigns are conducted and more countries decide to publish their (sub)national datasets (see Annex Table A1). For LiDAR point clouds, there is currently no central data repository in place. Instead, all national datasets are stored on separate repositories or websites, usually without machine-readable access to their interfaces (e.g., no standardized and stable APIs). National websites are usually in the local language and poorly documented which generates additional barriers for data re-use in a European context. There are also only few data repositories available (national or global) for sharing raw and pre-processed UAV data (RGB/NIR imagery and LiDAR point clouds). One example is the OpenAerialMap repository (https://openaerialmap.org/) which provides access to openly licensed imagery and map layer services. Much UAV data, however, are

available in general-purpose open repositories such as Zenodo. This makes them often very hard to find. More broadly, there is no overview or inventory of UAV datasets yet available, and information on accessibility (e.g., access protocols, licences, ownership, citation), data extent (e.g., spatial and temporal extent), and interoperability (e.g., standards and formats) are not available in a machine-readable way. Hence, creating a data repository in a VRE (or APIs to exchange data with existing repositories) would be an important infrastructure service (Figure 8).

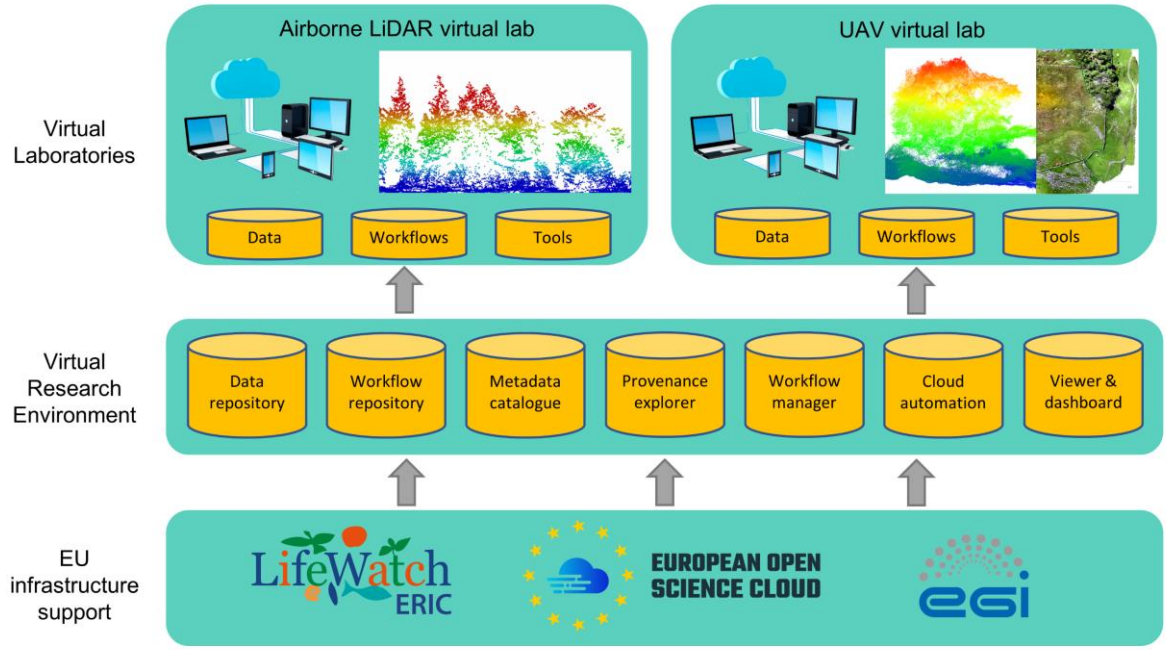

Figure 8: A simplified illustration of a Virtual Research Environment (VRE) that enables the creation of application-specific virtual labs for deriving habitat condition metrics, e.g., for processing airborne LiDAR or UAV remote sensing data, with underlying support from available EU infrastructure services.

Similar to LiDAR and UAV imagery, the processing workflows and deep learning models are not easy to find and access. Some machine learning models are stored on online repositories such as the Hugging Face's open-source platform (https://huggingface.co/), but few researchers make use of this for workflows that produce habitat condition metrics from LiDAR and UAV imagery. Other processing workflows are also stored in general-purpose open repositories such as Zenodo, and thus hard to find. For software that has been developed in the Python programming language, the Python Package Index (PyPI) provides a software repository. However, many processing workflows, even if programmed in Python, are not easy to find. Hence, a workflow repository in a VRE (Figure 8) with relevant R and Python scripts, Jupyter notebooks and other open-source software tools would facilitate the upscaling of habitat condition metrics from airborne LiDAR and UAV remote sensing across the EU.

6.3 Metadata catalogues

A metadata catalogue would allow making airborne LiDAR point clouds, RGB/NIR imagery from UAVs, and their processing workflows findable and accessible in a VRE (Figure 8). This

may include mapping services that harvest metadata of relevant assets (datasets, software and algorithms) from other platforms and research environments (Martin *et al.*, 2019; Zhao *et al.*, 2022). Such a catalogue should include basic information on metadata (e.g., general info, geographic and temporal information, LiDAR-specific attributes), how data are stored (databases, single files, file formats, etc.), and how they can be accessed (e.g., via open data platforms, institutional repositories or websites). The LAS/LAZ format (ASPRS, 2019) already provides a good basis for the standardized description of LiDAR raw data (e.g., point clouds from national ALS surveys). However, once LiDAR point clouds have been processed into LiDAR metrics, they are typically made available as raster files (Assmann *et al.*, 2022; Kissling *et al.*, 2022; Roussel *et al.*, 2020). To describe these generated data products in a standardized way (for publishing, sharing, and reuse), metadata templates need to include also relevant metadata specifications for geospatial and environmental datasets, e.g., based on INSPIRE, ISO, and EML (Hardisty *et al.*, 2019). For datasets captured with UAVs, there is currently no standardised way of describing their metadata. A Minimum Information Framework (MIF) therefore needs to be developed and implemented to describe UAV platforms and flight plans (Barbieri *et al.*, 2023), including best practice protocols for campaign flying and data pre-processing. Similar solutions are needed for describing the metadata of processing workflows for UAV imagery and the resulting data products that are generated from them.

6.4 Other infrastructure services

Besides repositories and metadata catalogues, a cloud-based VRE would require additional digital infrastructure services (Figure 8). For instance, a provenance explorer would provide an interface for users to interactively explore the system logs and the provenance of workflows and to identify the anomalies and reproduce the workflows or problems (Zhao *et al.*, 2022). A workflow manager (Figure 8) would help to design specific workflows, to compose the logical relations and dependencies among components, to configure the parameters (as well as input and output), and to execute the processing. Tools for cloud automation would enable the execution of workflows on remote infrastructures, including planning, automation and configuration of VMs and computing clusters, and the scheduling of workflow execution (Zhao *et al.*, 2022). Finally, online dashboards and web viewers would visualize the progress of workflow execution and provide a comprehensive data view, including maps, graphics, and summary statistics.

A VRE would take advantage of existing EU infrastructures (Figure 8). For instance, LifeWatch ERIC is an e-Science European infrastructure for biodiversity and ecosystem research, which provides various ICT tools and services, including functionalities for VREs (https://www.lifewatch.eu/). The European Open Science Cloud (EOSC) is a virtual environment for hosting and processing research data to support EU open science, e.g., with open and seamless services for storage, management, analysis and re-use of research data (https://eosc-portal.eu/). The European Grid Infrastructure (EGI) provides access to high-throughput computing resources across Europe using grid computing techniques, including high-throughput and cloud computing, storage and data management (https://www.egi.eu/). Hence, the basic IT support could come from existing EU infrastructures whereas the specific infrastructure services of the VRE would have to be developed.

7 Conclusions

This report summarizes critical challenges for upscaling habitat condition metrics from airborne LiDAR and UAV remote sensing for an EU-wide monitoring. While the provided

overview is probably not exhaustive, the encountered challenges are exemplary for all 'four Vs' of big data that are typical in Earth system science (i.e., volume, velocity, variety and veracity; Reichstein *et al.*, 2019). They also reflect the main challenges encountered when applying the FAIR Guiding Principles for scientific data management (e.g., capturing standardized, richly described and machine-actionable (meta)data; Wilkinson *et al.*, 2016). It is therefore imperative to advance the development and application of standards and best practices for data collection and pre-processing, the compilation of standardized and machine-readable metadata, and the advancement of free and open-source software (FOSS). Dealing with heterogeneous datasets (e.g., different point densities and pre-classifications of national LiDAR datasets), applying and developing deep learning algorithms and making use of cloud and HPC computing infrastructures further requires interdisciplinary collaborations (e.g., between ecologists, machine learning engineers, and software developers) and a high degree of specialization for their basic processing. Advancing the adoption of Open Science principles (open data, open source and open methods) and FAIR Guiding Principles (Findability, Accessibility, Interoperability, and Reusability) (Gallagher *et al.*, 2020; Wilkinson *et al.*, 2016) will help to better assess and monitor the condition of habitats, and thereby contribute to reverse the degradation and unsustainable use of natural resources in the EU.

8 Acknowledgements

We thank Christophe Dominik (Helmholtz-Centre for Environmental Research - UFZ) and Alexis Joly (INRIA) for feedback on this document and Zhiming Zhao (University of Amsterdam) for insights into cloud-based virtual research environments.

9 References

Abrego, N., Roslin, T., Huotari, T., Ji, Y., Schmidt, N. M., Wang, J., Yu, D. W. & Ovaskainen, O. (2021). Accounting for species interactions is necessary for predicting how arctic arthropod communities respond to climate change. *Ecography* **44**(6), 885-896.

Allen, M. J., Grieve, S. W. D., Owen, H. J. F. & Lines, E. R. (2023). Tree species classification from complex laser scanning data in Mediterranean forests using deep learning. *Methods in Ecology and Evolution* **14**(7), 1657-1667.

Allen, N., Cooksley, H., Buchmann, C., Schurr, F. & Pagel, J. (2022). Automated mapping and identification of shrub individuals in South Africa's Fynbos biome using drone imagery and deep learning. In *Proceedings of the 7th bwHPC Symposium*, pp. 11–16. Open Access Repositorium der Universität Ulm und Technischen Hochschule Ulm, Ulm University.

ASPRS. (2019). LAS Specification 1.4 - R15, pp. 50. American Society for Photogrammetry & Remote Sensing, Maryland, USA.

Assmann, J. J., Moeslund, J. E., Treier, U. A. & Normand, S. (2022). EcoDes-DK15: high-resolution ecological descriptors of vegetation and terrain derived from Denmark's national airborne laser scanning data set. *Earth Syst. Sci. Data* **14**(2), 823-844.

Bakx, T. R. M., Koma, Z., Seijmonsbergen, A. C. & Kissling, W. D. (2019). Use and categorization of Light Detection and Ranging vegetation metrics in avian diversity and species distribution research. *Diversity and Distributions* **25**(7), 1045-1059.

Barbieri, L., Wyngaard, J., Swanz, S. & Thomer, A. K. (2023). Making drone data FAIR through a community-developed information framework. *Data Science Journal* **22**(1), 1-9.

Barnas, A. F., Darby, B. J., Vandeberg, G. S., Rockwell, R. F. & Ellis-Felege, S. N. (2019). A comparison of drone imagery and ground-based methods for estimating the extent

of habitat destruction by lesser snow geese (*Anser caerulescens caerulescens*) in La Pérouse Bay. *Plos One* **14**(8), e0217049.

Brede, B., Bartholomeus, H. M., Barbier, N., Pimont, F., Vincent, G. & Herold, M. (2022). Peering through the thicket: Effects of UAV LiDAR scanner settings and flight planning on canopy volume discovery. *International Journal of Applied Earth Observation and Geoinformation* **114**, 103056.

Chen, J., Chen, Y. & Liu, Z. (2021). Classification of typical tree species in laser point cloud based on deep learning. *Remote Sensing* **13**(23), 4750.

Cunliffe, A. M., Brazier, R. E. & Anderson, K. (2016). Ultra-fine grain landscape-scale quantification of dryland vegetation structure with drone-acquired structure-from-motion photogrammetry. *Remote Sensing of Environment* **183**, 129-143.

Czúcz, B., Keith, H., Driver, A., Jackson, B., Nicholson, E. & Maes, J. (2021). A common typology for ecosystem characteristics and ecosystem condition variables. *One Ecosystem* **6**, e58218.

Davies, A. B. & Asner, G. P. (2014). Advances in animal ecology from 3D-LiDAR ecosystem mapping. *Trends in Ecology & Evolution* **29**(12), 681–691.

de Vries, J. P. R., Koma, Z., WallisDeVries, M. F. & Kissling, W. D. (2021). Identifying fine-scale habitat preferences of threatened butterflies using airborne laser scanning. *Diversity and Distributions* **27**(7), 1251-1264.

Díaz, S., Settele, J., Brondízio, E. S., Ngo, H. T., Agard, J., Arneth, A., Balvanera, P., Brauman, K. A., Butchart, S. H. M., Chan, K. M. A., Garibaldi, L. A., Ichii, K., Liu, J., Subramanian, S. M., Midgley, G. F., Miloslavich, P., Molnár, Z., Obura, D., Pfaff, A., Polasky, S., Purvis, A., Razzaque, J., Reyers, B., Chowdhury, R. R., Shin, Y.-J., Visseren-Hamakers, I., Willis, K. J. & Zayas, C. N. (2019). Pervasive human-driven decline of life on Earth points to the need for transformative change. *Science* **366**(6471), eaax3100.

Eischeid, I., Soininen, E. M., Assmann, J. J., Ims, R. A., Madsen, J., Pedersen, Å. Ø., Pirotti, F., Yoccoz, N. G. & Ravolainen, V. T. (2021). Disturbance mapping in Arctic tundra improved by a planning workflow for drone studies: advancing tools for future ecosystem monitoring. *Remote Sensing* **13**(21), 4466.

European Commission. (2021). EU Biodiversity Strategy for 2030: Bringing nature back into our lives. Publications Office of the European Union, Luxembourg.

European Environment Agency. (2020). State of nature in the EU: Results from reporting under the nature directives 2013-2018. Publications Office of the European Union, Luxembourg.

Fan, Z., Wei, J., Zhang, R. & Zhang, W. (2023). Tree species classification based on PointNet++ and airborne laser survey point cloud data enhancement. *Forests* **14**(6), 1246.

Gallagher, R. V., Falster, D. S., Maitner, B. S., Salguero-Gómez, R., Vandvik, V., Pearse, W. D., Schneider, F. D., Kattge, J., Poelen, J. H., Madin, J. S., Ankenbrand, M. J., Penone, C., Feng, X., Adams, V. M., Alroy, J., Andrew, S. C., Balk, M. A., Bland, L. M., Boyle, B. L., Bravo-Avila, C. H., Brennan, I., Carthey, A. J. R., Catullo, R., Cavazos, B. R., Conde, D. A., Chown, S. L., Fadrique, B., Gibb, H., Halbritter, A. H., Hammock, J., Hogan, J. A., Holewa, H., Hope, M., Iversen, C. M., Jochum, M., Kearney, M., Keller, A., Mabee, P., Manning, P., McCormack, L., Michaletz, S. T., Park, D. S., Perez, T. M., Pineda-Munoz, S., Ray, C. A., Rossetto, M., Sauquet, H., Sparrow, B., Spasojevic, M. J., Telford, R. J., Tobias, J. A., Violle, C., Walls, R., Weiss, K. C. B., Westoby, M., Wright, I. J. & Enquist, B. J. (2020). Open Science principles for accelerating trait-based science across the Tree of Life. *Nature Ecology & Evolution* **4**, 294–303.

Graham, L., Broughton, R. K., Gerard, F. & Gaulton, R. (2019). Remote sensing applications for hedgerows. In *The ecology of hedgerows and field margins* (ed J. W. Dover). Taylor & Francis, Routledge.

Hardisty, A. R., Michener, W. K., Agosti, D., Alonso García, E., Bastin, L., Belbin, L., Bowser, A., Buttigieg, P. L., Canhos, D. A. L., Egloff, W., De Giovanni, R., Figueira, R., Groom, Q., Guralnick, R. P., Hobern, D., Hugo, W., Koureas, D., Ji, L., Los, W., Manuel, J., Manset, D., Poelen, J., Saarenmaa, H., Schigel, D., Uhlir, P. F. & Kissling, W. D. (2019). The Bari Manifesto: An interoperability framework for essential biodiversity variables. *Ecological Informatics* **49**, 22-31.

Hill, D. J., Tarasoff, C., Whitworth, G. E., Baron, J., Bradshaw, J. L. & Church, J. S. (2017). Utility of unmanned aerial vehicles for mapping invasive plant species: a case study on yellow flag iris (*Iris pseudacorus* L.). *International Journal of Remote Sensing* **38**(8-10), 2083-2105.

Høye, T. T., August, T., Balzan, M. V., Biesmeijer, K., Bonnet, P., Breeze, T. D., Dominik, C., Gerard, F., Joly, A., Kalkman, V., Kissling, W. D., Metodiev, T., Moeslund, J., Potts, S., Roy, D. B., Schweiger, O., Senapathi, D., Settele, J., Stoev, P. & Stowell, D. (2023). Modern Approaches to the Monitoring of Biodiversity (MAMBO). *Research Ideas and Outcomes* **9**, e116951.

Ikkala, L., Ronkanen, A.-K., Ilmonen, J., Similä, M., Rehell, S., Kumpula, T., Päkkilä, L., Klöve, B. & Marttila, H. (2022). Unmanned aircraft system (UAS) structure-from-motion (SfM) for monitoring the changed flow paths and wetness in minerotrophic peatland restoration. *Remote Sensing* **14**(13), 3169.

James, K. & Bradshaw, K. (2020). Detecting plant species in the field with deep learning and drone technology. *Methods in Ecology and Evolution* **11**(11), 1509-1519.

Jiang, S., Jiang, C. & Jiang, W. (2020). Efficient structure from motion for large-scale UAV images: A review and a comparison of SfM tools. *ISPRS Journal of Photogrammetry and Remote Sensing* **167**, 230-251.

Kissling, W. D. & Shi, Y. (2023). Which metrics derived from airborne laser scanning are essential to measure the vertical profile of ecosystems? *Diversity and Distributions* **29**(10), 1315-1320.

Kissling, W. D., Shi, Y., Koma, Z., Meijer, C., Ku, O., Nattino, F., Seijmonsbergen, A. C. & Grootes, M. W. (2022). Laserfarm – A high-throughput workflow for generating geospatial data products of ecosystem structure from airborne laser scanning point clouds. *Ecological Informatics* **72**, 101836.

Kissling, W. D., Shi, Y., Koma, Z., Meijer, C., Ku, O., Nattino, F., Seijmonsbergen, A. C. & Grootes, M. W. (2023). Country-wide data of ecosystem structure from the third Dutch airborne laser scanning survey. *Data in Brief* **46**, 108798.

Koma, Z., Grootes, M. W., Meijer, C. W., Nattino, F., Seijmonsbergen, A. C., Sierdsema, H., Foppen, R. & Kissling, W. D. (2021). Niche separation of wetland birds revealed from airborne laser scanning. *Ecography* **44**(6), 907-918.

Lendzioch, T., Langhammer, J., Vlček, L. & Minařík, R. (2021). Mapping the groundwater level and soil moisture of a montane peat bog using UAV monitoring and machine learning. *Remote Sensing* **13**(5), 907.

Lucas, C., Bouten, W., Koma, Z., Kissling, W. D. & Seijmonsbergen, A. C. (2019). Identification of linear vegetation elements in a rural landscape using LiDAR point clouds. *Remote Sensing* **11**(3), 292.

Maes, J., Bruzón, A. G., Barredo, J. I., Vallecillo, S., Vogt, P., Rivero, I. M. & Santos-Martín, F. (2023). Accounting for forest condition in Europe based on an international statistical standard. *Nature Communications* **14**(1), 3723.

Marchi, N., Pirotti, F. & Lingua, E. (2018). Airborne and terrestrial laser scanning data for the assessment of standing and lying deadwood: current situation and new perspectives. *Remote Sensing* **10**(9), 1356.

Martin, P., Remy, L., Theodoridou, M., Jeffery, K. & Zhao, Z. (2019). Mapping heterogeneous research infrastructure metadata into a unified catalogue for use in a generic virtual research environment. *Future Generation Computer Systems* **101**, 1-13.

Martinuzzi, S., Vierling, L. A., Gould, W. A., Falkowski, M. J., Evans, J. S., Hudak, A. T. & Vierling, K. T. (2009). Mapping snags and understory shrubs for a LiDAR-based assessment of wildlife habitat suitability. *Remote Sensing of Environment* **113**(12), 2533-2546.

Meijer, C., Grootes, M. W., Koma, Z., Dzigan, Y., Gonçalves, R., Andela, B., van den Oord, G., Ranguelova, E., Renaud, N. & Kissling, W. D. (2020). Laserchicken—A tool for distributed feature calculation from massive LiDAR point cloud datasets. *SoftwareX* **12**, 100626.

Moeslund, J. E., Arge, L., Bøcher, P. K., Dalgaard, T., Odgaard, M. V., Nygaard, B. & Svenning, J.-C. (2013). Topographically controlled soil moisture is the primary driver of local vegetation patterns across a lowland region. *Ecosphere* **4**(7), art91.

Moeslund, J. E., Clausen, K. K., Dalby, L., Fløjgaard, C., Pärtel, M., Pfeifer, N., Hollaus, M. & Brunbjerg, A. K. (2023). Using airborne lidar to characterize North European terrestrial high-dark-diversity habitats. *Remote Sensing in Ecology and Conservation* **9**(3), 354-369.

Moeslund, J. E., Zlinszky, A., Ejrnæs, R., Brunbjerg, A. K., Bøcher, P. K., Svenning, J.-C. & Normand, S. (2019). Light detection and ranging explains diversity of plants, fungi, lichens and bryophytes across multiple habitats and large geographic extent. *Ecological Applications* **29**(5), e01907.

Murray, B. A., Coops, N. C., Winiwarter, L., White, J. C., Dick, A., Barbeito, I. & Ragab, A. (2024). Estimating tree species composition from airborne laser scanning data using point-based deep learning models. *ISPRS Journal of Photogrammetry and Remote Sensing* **207**, 282-297.

Olariu, H. G., Malambo, L., Popescu, S. C., Virgil, C. & Wilcox, B. P. (2022). Woody plant encroachment: evaluating methodologies for semiarid woody species classification from drone images. *Remote Sensing* **14**(7), 1665.

Oldeland, J., Revermann, R., Luther-Mosebach, J., Buttschardt, T. & Lehmann, J. R. K. (2021). New tools for old problems — comparing drone- and field-based assessments of a problematic plant species. *Environmental Monitoring and Assessment* **193**(2), 90.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N. & Prabhat. (2019). Deep learning and process understanding for data-driven Earth system science. *Nature* **566**(7743), 195-204.

Röschel, L., Noebel, R., Stein, U., Naumann, S., Romão, C., Tryfon, E., Gaudillat, Z., Roscher, S., Moser, D., Ellmauer, T., Löhnertz, M., Halada, L., Staneva, A. & Rutherford, C. (2020). State of Nature in the EU - Methodological paper. Methodologies under the Nature Directives reporting 2013-2018 and analysis for the State of Nature 2000. In *ETC/BD Technical paper N° 2/2020*. European Topic Centre on Biological Diversity, Paris cedex, France.

ROUSSEL, J.-R., AUTY, D., COOPS, N. C., TOMPALSKI, P., GOODBODY, T. R. H., MEADOR, A. S., BOURDON, J.-F., DE BOISSIEU, F. & ACHIM, A. (2020). lidR: An R package for analysis of Airborne Laser Scanning (ALS) data. *Remote Sensing of Environment* **251**, 112061.

SHI, Y. & KISSLING, W. D. (2023). Performance, effectiveness and computational efficiency of powerline extraction methods for quantifying ecosystem structure from light detection and ranging. *GIScience & Remote Sensing* **60**(1), 2260637.

VAN IERSEL, W., STRAATSMA, M., ADDINK, E. & MIDDELKOOP, H. (2018). Monitoring height and greenness of non-woody floodplain vegetation with UAV time series. *ISPRS Journal of Photogrammetry and Remote Sensing* **141**, 112-123.

WANG, D., BRUNNER, J., MA, Z., LU, H., HOLLAUS, M., PANG, Y. & PFEIFER, N. (2018a). Separating tree photosynthetic and non-photosynthetic components from point cloud data using dynamic segment merging. *Forests* **9**(5), 252.

WANG, J., LINDENBERGH, R. & MENENTI, M. (2018b). Scalable individual tree delineation in 3D point clouds. *The Photogrammetric Record* **33**(163), 315-340.

WANG, Y., KOULOUZIS, S., BIANCHI, R., LI, N., SHI, Y., TIMMERMANS, J., KISSLING, W. D. & ZHAO, Z. (2022). Scaling notebooks as re-configurable cloud workflows. *Data Intelligence* **4**(2), 409-425.

WEINSTEIN, B. G., MARCONI, S., BOHLMAN, S. A., ZARE, A. & WHITE, E. P. (2020). Cross-site learning in deep learning RGB tree crown detection. *Ecological Informatics* **56**, 101061.

WIDYANINGRUM, E., BAI, Q., FAJARI, M. K. & LINDENBERGH, R. C. (2021). Airborne laser scanning point cloud classification using the DGCNN deep learning method. *Remote Sensing* **13**(5), 859.

WILKINSON, M. D., DUMONTIER, M., AALBERSBERG, I. J., APPLETON, G., AXTON, M., BAAK, A., BLOMBERG, N., BOITEN, J.-W., DA SILVA SANTOS, L. B., BOURNE, P. E., BOUWMAN, J., BROOKES, A. J., CLARK, T., CROSAS, M., DILLO, I., DUMON, O., EDMUNDS, S., EVELO, C. T., FINKERS, R., GONZALEZ-BELTRAN, A., GRAY, A. J. G., GROTH, P., GOBLE, C., GRETHE, J. S., HERINGA, J., 'T HOEN, P. A. C., HOOFT, R., KUHN, T., KOK, R., KOK, J., LUSHER, S. J., MARTONE, M. E., MONS, A., PACKER, A. L., PERSSON, B., ROCCA-SERRA, P., ROOS, M., VAN SCHAIK, R., SANSONE, S.-A., SCHULTES, E., SENGSTAG, T., SLATER, T., STRAWN, G., SWERTZ, M. A., THOMPSON, M., VAN DER LEI, J., VAN MULLIGEN, E., VELTEROP, J., WAAGMEESTER, A., WITTENBURG, P., WOLSTENCROFT, K., ZHAO, J. & MONS, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* **3**, 160018.

WYNGAARD, J., BARBIERI, L., THOMER, A., ADAMS, J., SULLIVAN, D., CROSBY, C., PARR, C., KLUMP, J., RAJ SHRESTHA, S. & BELL, T. (2019). Emergent challenges for science sUAS data management: fairness through community engagement and best practices development. *Remote Sensing* **11**(15), 1797.

XI, Z., HOPKINSON, C., ROOD, S. B. & PEDDLE, D. R. (2020). 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* **168**, 1-16.

YOUSEFHUSSIEN, M., KELBE, D. J., IENTILUCCI, E. J. & SALVAGGIO, C. (2018). A multi-scale fully convolutional network for semantic labeling of 3D point clouds. *ISPRS Journal of Photogrammetry and Remote Sensing* **143**, 191-204.

ZHANG, C., ATKINSON, P. M., GEORGE, C., WEN, Z., DIAZGRANADOS, M. & GERARD, F. (2020). Identifying and mapping individual plants in a highly diverse high-elevation ecosystem using UAV imagery and deep learning. *ISPRS Journal of Photogrammetry and Remote Sensing* **169**, 280-291.

ZHANG, Y., ONDA, Y., KATO, H., FENG, B. & GOMI, T. (2022). Understory biomass measurement in a dense plantation forest based on drone-SfM data by a manual low-flying drone under the canopy. *Journal of Environmental Management* **312**, 114862.

ZHANG, Z. & ZHU, L. (2023). A review on Unmanned Aerial Vehicle remote sensing: platforms, sensors, data processing methods, and applications. *Drones* **7**(6), 398.

ZHAO, P., GUAN, H., LI, D., YU, Y., WANG, H., GAO, K., MARCATO JUNIOR, J. & LI, J. (2021). Airborne multispectral LiDAR point cloud classification with a feature reasoning-based graph convolution network. *International Journal of Applied Earth Observation and Geoinformation* **105**, 102634.

ZHAO, Z., KOULOUZIS, S., BIANCHI, R., FARSHIDI, S., SHI, Z., XIN, R., WANG, Y., LI, N., SHI, Y., TIMMERMANS, J. & KISSLING, W. D. (2022). Notebook-as-a-VRE (NaaVRE): From private notebooks to a collaborative cloud virtual research environment. *Software: Practice and Experience* **52**(9), 1947-1966.

MAMBO

MODERN APPROACHES TO THE MONITORING OF BIODIVERSITY

www.mambo-project.eu

Project partners

AARHUS UNIVERSITY

Naturalis Biodiversity Center

UFZ

University of Reading

Inria

PENSOFT

UK CEH

cirad

ECOSTACK INNOVATIONS LIMITED

10 Annex

10.1 Table A1 Examples of European open-access LiDAR point clouds

Table A1: Examples of European open-access LiDAR point clouds derived from airborne laser scanning (ALS) surveys over a (sub)national extent (as at 2023-10-28). Such raw datasets (point clouds) enable the quantification of vegetation structure and terrain properties at high spatial (e.g. 1–10 m) resolution by processing the multi-terabyte ALS dataset into raster layers, capturing height, cover or structural complexity of vegetation.

Country	Dataset name	Survey period	Point density (pt/m^2)	RGB NIR	Data volume*	Pre-classification	Leaf-on/leaf-off	Coordinate system	Coverage	Download
Finland	Laser scanning data 5 p	2008–2020	0.5–5	-	4 TB	Unclassified (1), Ground (2), Low vegetation (3), Medium vegetation (4), High vegetation (5), Low error points (7), Overlap area (12), Air points (15), Isolated (16), Fault points (17)	Leaf-off	ETRS89 / TM35FIN(E,N) (EPSG:3067); N2000 height (EPSG:3900)	Complete	Link
Sweden	Laser data Download, forest/NH	2018–2022	1–2	-	5 TB	Unclassified (1), Ground (2), Low point (noise) (7), Water (9), Bridges (17), High noise (18)	Leaf-off	SWEREF 99 TM (EPSG:3006)	Partially	Link
Norway	Laser scanning for national detailed height model (NDH)	2016–2023	0.2–10	-	6 TB	Unclassified (1), Ground (2), Low vegetation (3), Medium vegetation (4), High vegetation (5), Building (6), Low point (7), Water (9), Overlap (12), Bridge (17),	Leaf-on & Leaf-off	EUREF89 UTM32/33/35	Complete	Link

						Big pot rock (20), Snow (21)				
Denmark	Danish elevation model (DHM)	2007 2018-2023	4.5–25	-	10 TB	Unclassified (1), Ground (2), Low vegetation (3), Medium vegetation (4), High vegetation (5), Building (6), Water (9)	Leaf-off	ETR89 UTM 32N (EPSG: 25832)	Complete	Link
Estonia	Estonian Topographic Database	2008–2021	0.2–18	RGB	30 TB	Unclassified (1), Terrain (2), Tall vegetation (5), Buildings (6), Noise (7), Water (9), Terrain under bridges (17), Noise (18)	Leaf-off	Estonian Coordinate System of 1997 (EPSG:3301)	Complete	Link
England	National LiDAR Programme Point Cloud	2016–2021	0.5–16	-	45 TB	Unclassified (1), Ground (2), Low vegetation (3), Medium vegetation (4), High vegetation (5), Building (6), Noise (7) Water (9)	Leaf-off	OSGB 1936 / British National Grid	Complete	Link
Scotland	LiDAR for Scotland Phase 1-6 LAS	2009–2022	1–16	-	8 TB	Unclassified (1), Ground (2)	Leaf-off	ETRS89 (EPSG:4258)	Complete	Link
Netherlands	AHN1–4	1996–2002, 2007–2012, 2014–2019, 2020–2022	0.2–1, 6–10, 10–16, 20–30	-	1TB, 7 TB, 16 TB, 40 TB	Never classified (0), Unclassified (1), Ground (2), Building (6), Water (9), Reserved (26)	Leaf-off	EPSG:28992 Amersfoort / RD New	Complete	Link
Belgium	Open LiDAR data Digital Elevation Model	2001–2004, 2013–2015	16–20	RGB	25 TB	Unclassified (1), Ground (2), Water (9)	Leaf-off	Belge 1972 (EPSG:31370)	Partially	Link
Germany	Digital terrain models (DGM)	2010–2013 2014–2019 2020–2023	4–10	-	10 TB	Unclassified (1), Ground (2), Water (9), Wire (13),	Leaf-off	EPSG: 25833	Partially	Link

						Reserved (30)				
France	National HD LiDAR programme Point Cloud	2021–2023	1–10	-	20 TB	Unclassified (1), Ground (2), Low vegetation (3), Medium vegetation (4), High vegetation (5), Building (6), Water (9), Bridge (17), Perennial Surface (64), Artifacts (65), Virtual Points (66)	Leaf-on/leaf-off	RGF93 / Lambert-93	Partially	Link
Switzerland	swissSURFACE3D	2017–2024	5–20	-	25 TB	Unclassified (1), Ground (2), Vegetation (3), Building (6), Water (9), Bridge (17)	Leaf-off	LV95 – LN02	Partially	Link
Luxembourg	LiDAR 2019 - 3D survey	2019	15–20	RGB	2 TB	Unclassified (1), Ground (2), Low vegetation (3), Medium vegetation (4), High vegetation (5), Building (6), Noise (7), Water (9), Bridge (13), Powerlines (15)	Leaf-off	ETRS89	Complete	Link
Poland	ISOK Laser scanning (LIDAR) data– entral level	2010-2011 2011-2023	4–12	-	12 TB	Never classified (0), Ground (2), Low vegetation (3), Medium vegetation (4), High vegetation (5), Building (6), Noise (7), Water (9), Overlap (12)	Leaf-off/Leaf-on	ETRS89 / Poland CS92	Complete	Link
Spain	LIDAR 2ª Cobertura	2015–2021	0.5–2	RGB NIR	5 TB	Unclassified (1), Ground (2), Low vegetation (3),	Leaf-off	ETRS89	Complete	Link

						Medium vegetation (4), High vegetation (5), Building (6), Noise (7), Water (9), Overlap (12), Bridge (17)				
Portugal	Projeto áGIL - Dados LiDAR	2019–2020	5–14	RGB NIR	5 TB	-	Leaf-on	PT-TM06/ETRS89	Partially	Link
Greece	Santorini LiDAR data	2012	2–5	-	2 TB	Unclassified (1), Noise (7)	Leaf-on	HGRS87	Partially	Link
Slovenia	Slovenian LiDAR data	2014–2015	2–5	-	2.5 TB	Never classified (0), Unclassified (1), Ground (2), Low vegetation (3), Medium vegetation (4), High vegetation (5), Building (6), Noise (7)	Leaf-on	EPSG:5514	Complete	Link
Latvia	Basic data of the digital elevation model	2013–2019	1.5–4	-	10 TB	Unclassified (1), Ground (2), Low vegetation (3), Medium vegetation (4), High vegetation (5), Building (6), Noise (7), Water (9)	Leaf-off	LKS-92 TM	Complete	Link
Slovakia	Airborne Laser Scanning 1st project cycle	2017–2023	5–10	-	8 TB	Unclassified (1), Ground (2)	Leaf-off	S-JTSK(JTSK03)	Complete	Link
Lithuania	Lidar_DR_LT	2019–2021	5–25	-	6 TB	Never classified (0), Unclassified (1), Ground (2), Low vegetation (3), Medium vegetation (4), High vegetation (5), Building (6), Noise (7), Overlap (12)	Leaf-off	LKS94	Partially	Link

*Data volume represents how much data storage is needed. It is estimated based on the number of files available in each download portal and the average size of each file.

10.2 Table A2 LiDAR metrics of vegetation structure

Table A2: Twenty-five LiDAR metrics capturing vegetation structure in three key dimensions (i.e. vegetation height, vegetation cover and structural complexity of vegetation). All metrics were calculated with the normalized point cloud of the Dutch AHN4 dataset (see Annex Table A1), using the Laserfarm high-throughput workflow (Kissling *et al.*, 2022).

Abbreviation	Name	Formula	Description	Ecological relevance
Vegetation height				
H_max	Maximum vegetation height	z_{max}	Maximum of normalized z within a grid cell	Height of the vegetation canopy surface and tree tops
H_mean	Mean of vegetation height	$\frac{1}{N} \times \sum z_i$ where N is the number of normalized z values and $\sum z_i$ the sum of all normalized z values in a grid cell	Mean of normalized z within a grid cell	Average height of vegetation (e.g. mean tree and shrub height in forests)
H_median	Median of vegetation height	z_{median}	Median of normalized z within a grid cell	Average height and vertical distribution of vegetation
Hp_25	25th percentile of vegetation height	$n = \left(\frac{25}{100}\right) \times N$, where N = number of normalized z values (sorted from smallest to largest), and n = ordinal rank of a given value	Capturing the 25th percentile of normalized z within a grid cell	Density of vegetation in the low stratum
Hp_50	50th percentile of vegetation height	$n = \left(\frac{50}{100}\right) \times N$, where N = number of normalized z values (sorted from smallest to largest), and n = ordinal rank of a given value. This corresponds to the Hmedian	Capturing the 50th percentile of normalized z within a grid cell	Average height and vertical distribution of vegetation
Hp_75	75th percentile of vegetation height	$n = \left(\frac{75}{100}\right) \times N$, where N = number of normalized z values (sorted from smallest to largest), and n = ordinal rank of a given value	Capturing the 75th percentile of normalized z within a grid cell	Density of vegetation in the upper stratum
Hp_95	95th percentile of vegetation height	$n = \left(\frac{95}{100}\right) \times N$, where N = number of normalized z values (sorted from smallest to largest), and n = ordinal rank of a given value	Capturing the 95th percentile of normalized z within a grid cell	Height of the vegetation canopy surface and tree tops, accounting for the effect of outliers
Vegetation cover				
PPR	Pulse penetration ratio	$\frac{N_{ground}}{N_{total}}$	Ratio of number of ground points (N_{ground}) to the total number of points (N_{total}) within a grid cell	Openness of vegetation, canopy fractional cover, laser penetration index
Density above mean z	Canopy cover above mean height	$100 \times \sum[z_i > \bar{z}]/N$ where z_i are all normalized z values that are larger than the	Number of returns above mean height within a grid cell	Density of upper vegetation layer

		mean vegetation height $\bar{z}$ within a grid cell, and N the total number of normalized z values		
BR_1	Density of vegetation points below 1 m	$N_{z<1}/N_{total}$	Ratio of number of vegetation points (<1 m) to the total number of vegetation points within a grid cell	Density of vegetation <1 m
BR_1_2	Density of vegetation points between 1–2 m	$N_{1<z<2}/N_{total}$	Ratio of number of vegetation points (between 1–2 m) to the total number of vegetation points within a grid cell	Density of vegetation in 1–2 m layer
BR_2_3	Density of vegetation points between 2–3 m	$N_{2<z<3}/N_{total}$	Ratio of number of vegetation points (between 2–3 m) to the total number of vegetation points within a grid cell	Density of vegetation in 2–3 m layer
BR_3	Density of vegetation points above 3 m	$N_{z>3}/N_{total}$	Ratio of number of vegetation points (>3 m) to the total number of vegetation points within a grid cell	Density of vegetation above 3 m
BR_3_4	Density of vegetation points between 3–4 m	$N_{3<z<4}/N_{total}$	Ratio of number of vegetation points (between 3–4 m) to the total number of vegetation points within a grid cell	Density of vegetation in 3–4 m layer
BR_4_5	Density of vegetation points between 4–5 m	$N_{4<z<5}/N_{total}$	Ratio of number of vegetation points (between 4–5 m) to the total number of vegetation points within a grid cell	Density of vegetation in 4–5 m layer
BR_5	Density of vegetation points below 5 m	$N_{z<5}/N_{total}$	Ratio of number of vegetation points (<5 m) to the total number of vegetation points within a grid cell	Density of vegetation in understory layer (<5 m)
BR_5_20	Density of vegetation points between 5–20 m	$N_{5<z<20}/N_{total}$	Ratio of number of vegetation points (between 5–20 m) to the total number of vegetation points within a grid cell	Density of vegetation in 5–20 m layer
BR_20	Density of vegetation points above 20 m	$N_{z>20}/N_{total}$	Ratio of number of vegetation points (>20 m) to the total number of vegetation points within a grid cell	Density of vegetation above 20 m
Vegetation structural complexity				
Coeff. var. z	Coefficient of variation of vegetation height	$\frac{1}{\bar{z}} \times \sqrt{\sum \frac{(z_i-\bar{z})^2}{N-1}}$ where $\bar{z}$ is the mean vegetation height, z_i all normalized z values in a grid cell, and N the number of normalized z values	Coefficient of variation of normalized z within a grid cell	Vertical variability of vegetation distribution (ratio of the standard deviation to the mean)
Shannon index	Shannon index based on entropy	$-\sum_i p_i \times log_2 p_i$ where $p_i = N_i / \sum_j N_j$ and N_i the points in bin i	The negative sum of the proportion of points within 0.5 m height layers multiplied with the logarithm of the proportion of points within 0.5 m height layers within a grid cell	Complexity and evenness of vertical vegetation distribution, sometimes referred to as foliage height diversity

Kurtosis of H_veg.	Kurtosis of vegetation height	$\frac{1}{\sigma^4} \times \sum(z_i - \bar{z})^4 / N$ where z_i are the normalized z values in a grid cell, $\bar{z}$ the mean of normalized z values, and N the total number of normalized z values	Kurtosis of normalized z within a grid cell	Vertical distribution ('tailedness') of vegetation
Roughness of veg. height	Roughness of vegetation	$\sqrt{\sum(R_i - \bar{R})^2/(N-1)}$ where R_i are the residual after plane fitting, and $\bar{R}$ the mean of residuals	Standard deviation of the residuals of a locally fitted plane within a cylinder	Small-scale roughness and variability of vegetation
Skewness of veg. height	Skewness of vegetation height	$\frac{1}{\sigma^3} \times \sum(z_i - \bar{z})^3 / N$ where z_i are the normalized z values in a grid cell, $\bar{z}$ the mean of normalized z values, and N the total number of normalized z values	Skewness of normalized z within a grid cell	Vertical distribution (asymmetry) of vegetation
Std. dev. of veg. height	Standard deviation of vegetation height	$\sqrt{\sum \frac{(z_i - \bar{z})^2}{N-1}}$ where $\bar{z}$ is the mean vegetation height, z_i all normalized z values in a grid cell, and N the number of normalized z values	Standard deviation of normalized z within a grid cell	Vertical variability (i.e. amount of variation around mean) of vegetation distribution
Variance of veg. height	Variance of vegetation height	$\sum \frac{(z_i - \bar{z})^2}{N-1}$ where $\bar{z}$ is the mean vegetation height, z_i all normalized z values in a grid cell, and N the number of normalized z values	Variance of normalized z within a grid cell	Vertical variability of vegetation distribution (dispersion around mean height)

10.3 Text A1: Assessing the robustness of LiDAR vegetation metrics

To assess how LiDAR vegetation metrics might be affected by varying point densities, we calculated a set of 25 LiDAR vegetation metrics representing different aspects of vegetation height, vegetation cover and structural complexity (Annex Table A2). The metrics were calculated with the Laserchicken software (Meijer *et al.*, 2020) using the Laserfarm workflow (Kissling *et al.*, 2022) and the Dutch AHN4 dataset from the years 2020–2022 with a point density of 20–30 points/m^2 (Annex Table A1). The implementation of the Laserfarm workflow and the metric calculation was similar to the country-wide processing of the AHN3 point cloud (Kissling *et al.*, 2023; Kissling *et al.*, 2022), except that the metrics were only calculated for a selected number of plots.

We developed a workflow that combined Natura 2000 shapefiles from the European Environmental Agency and LAZ files from the AHN4 dataset (Annex Figure A1). We randomly placed 100 plots (i.e., squared polygons around centre points) of 1 × 1 m resolution or 10 × 10 m resolution within Dutch Natura 2000 sites. Focus was on Natura 2000 sites which predominantly contain woodland habitats (Annex Figure A2). To do this, we downloaded the Natura 2000 vector layer (version 2021 revision 1, Oct. 2022) from the European Environmental Agency (https://www.eea.europa.eu/en/datahub/datahubitem-view/6fc8ad2d-195d-40f4-bdec-576e7d1268e4) with its associated metadata and used the available habitat information (i.e., habitat codes, their percentage and description) to map Natura 2000 sites in the Netherlands (Annex Figure A2a). The habitat 'woodland' was dominant in 42 (21%) of the 198 recognized Natura 2000 sites in the Netherlands. This represented broad-leaved deciduous woodlands (habitat code N16) and coniferous woodlands (N17), but also mixed woodlands (N19) and one plantation (N20) (Annex Figure A2b). Woodlands were the most prevalent habitat class in those sites (i.e., largest percentage cover), with the majority of sites (90%) having a woodland cover of >38% (Annex Figure A2b).

Random placement of plots within the woodland Natura 2000 sites of the Netherlands was done using the "Create Random Points" tool of ArcGIS Pro 3.0 (ESRI, Redlands, California). The centre of each plot was then exported from ArcGIS and imported into the R programming software using the R package 'sf'. The R package 'dplyr' was used to generate square polygons of the desired resolutions (either 1 × 1 m resolution or 10 × 10 m) around the centre points. The square polygons were subsequently saved as ESRI shapefiles. These were then used to clip the LiDAR point clouds from AHN4. We acquired the AHN4 point cloud dataset from the repository of the PDOK webservices (https://app.pdok.nl/viewer/). The square polygons were loaded into memory and their spatial position was compared with the bounding box of the AHN4 LAZ files (Annex Figure A1), using two C++ libraries (i.e., 'shapefile' and 'LasTools'). When the bounding box intersected with a specific square polygon, only points inside the polygon were parsed and clipped. In case the square polygon was fully inside the bounding box (no intersection), all points within the polygon were clipped. Clipped points were exported as a LAZ file and used to calculate the 25 LiDAR vegetation metrics with the Laserfarm workflow. Since not all locations of the 100 randomly placed plots had points, the actual sample sizes were slightly smaller than 100, i.e., 94 plots for the 1 × 1 m resolution and 96 plots for the 10 × 10 m resolution, respectively.

Metrics were calculated with the original point density of the Dutch AHN4 dataset (20–30 points/m^2, see Annex Table A1) and with six randomly down-sampled point clouds

for the same plots (with 20, 15, 10, 5, 2 and 1 points/m^2, respectively). All points of each plot were first reshuffled and then the relevant number of points was randomly picked (e.g., 500 points in a 10 × 10 m plot to obtain a point density of 5 points/m^2). This was done using the C++ library 'algorithm' with the random seeds generator. The 25 LiDAR vegetation metrics were then calculated for the 1 × 1 m plots (Figure 5 main text) and the 10 × 10 m plots (Annex Figure A3).

10.4 Figure A1: Workflow for calculating LiDAR metrics within Natura 2000 sites

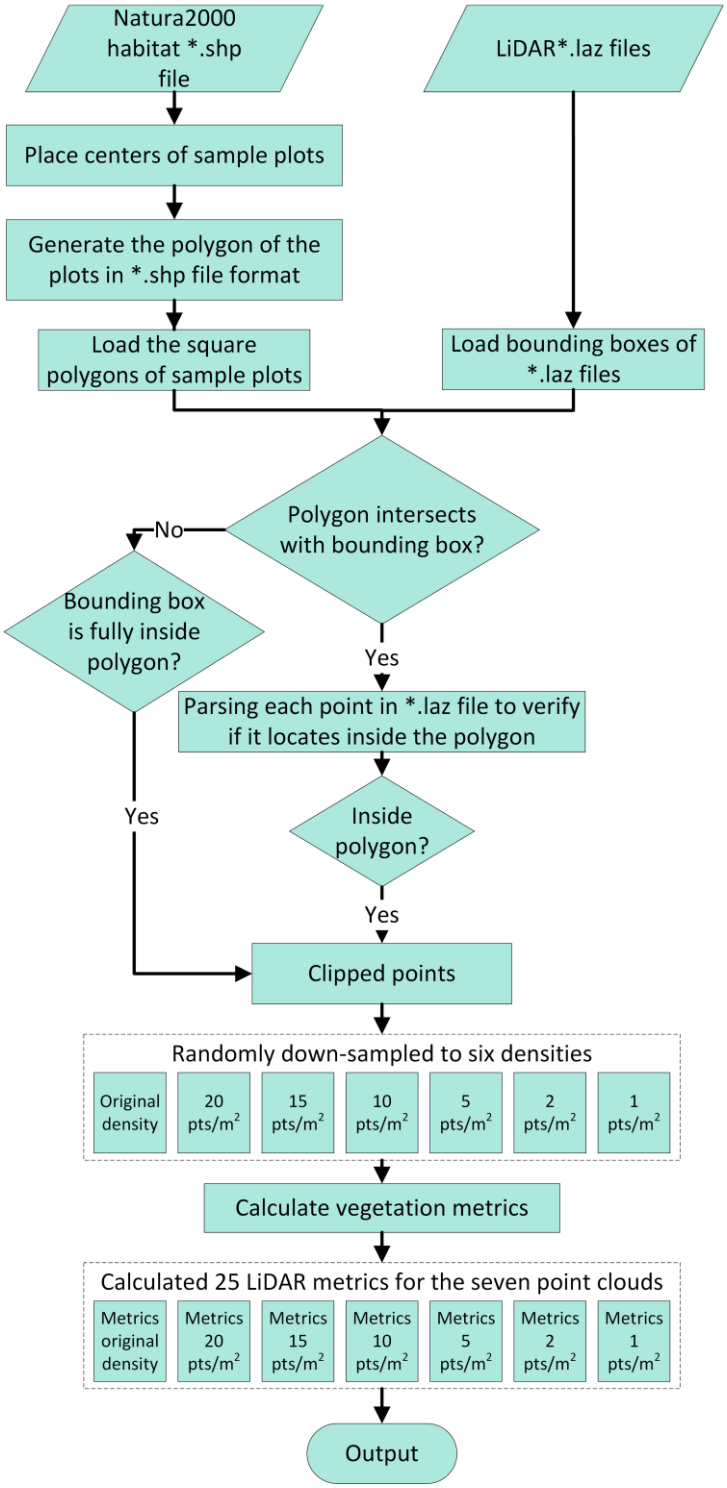

Figure A1: Workflow for calculating LiDAR metrics within specific plots that are randomly located within Natura 2000 sites. The workflow combines shapefiles from Natura 2000 sites with LAZ files from LiDAR point clouds, and then clips the points that fall inside specific polygons (e.g., plots of 1 × 1 m or 10 × 10 m resolution). The clipped points are then used to calculate LiDAR vegetation metrics for the original point density and for six down-sampled point densities.

10.5 Figure A2: Dutch Natura 2000 sites with woodland habitats

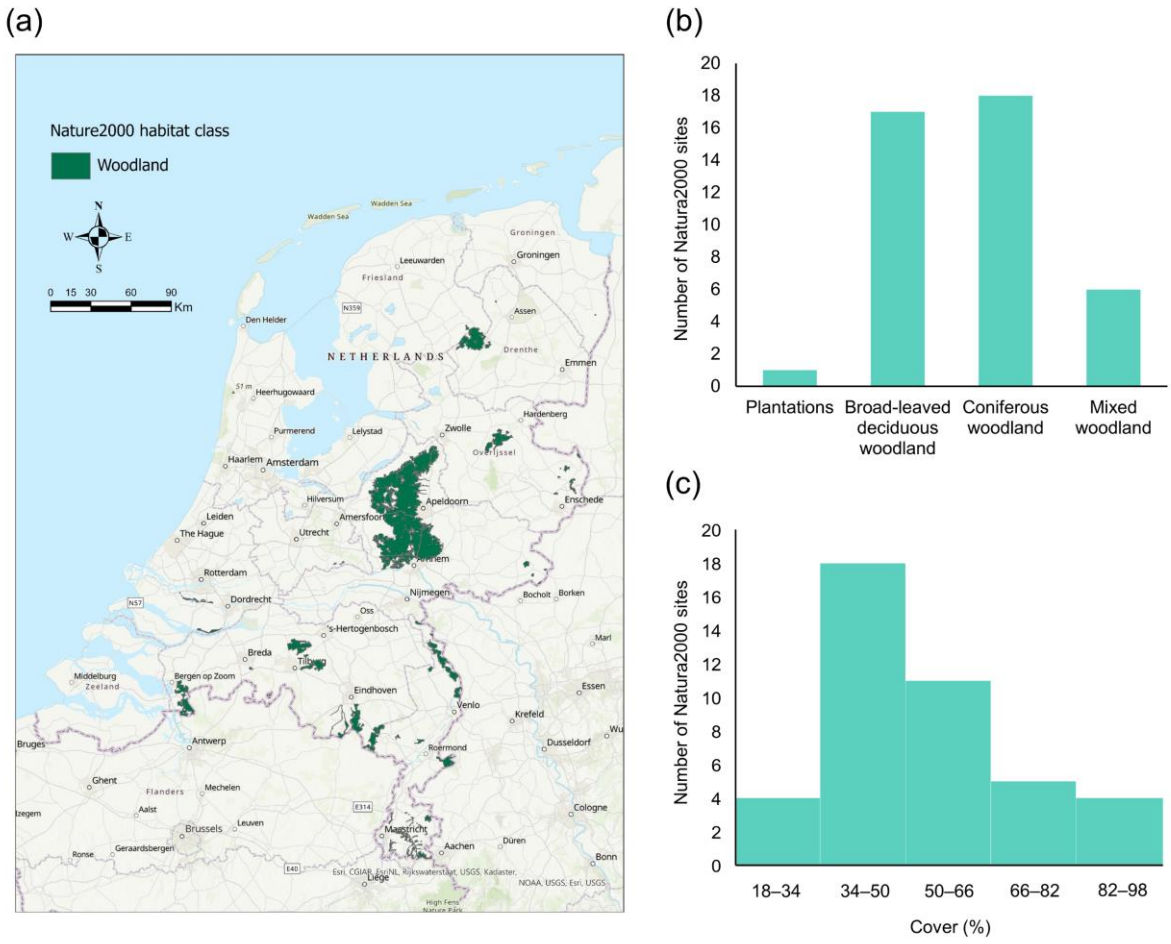

Figure A2: Dutch Natura 2000 sites with woodland habitats (*n* = 42). (a) Spatial distribution within the Netherlands. (b) Frequency of different woodland habitat classes. (c) Frequency distribution of the percentage cover of woodlands.

10.6 Figure A3: Effect of point density variation on LiDAR vegetation metrics

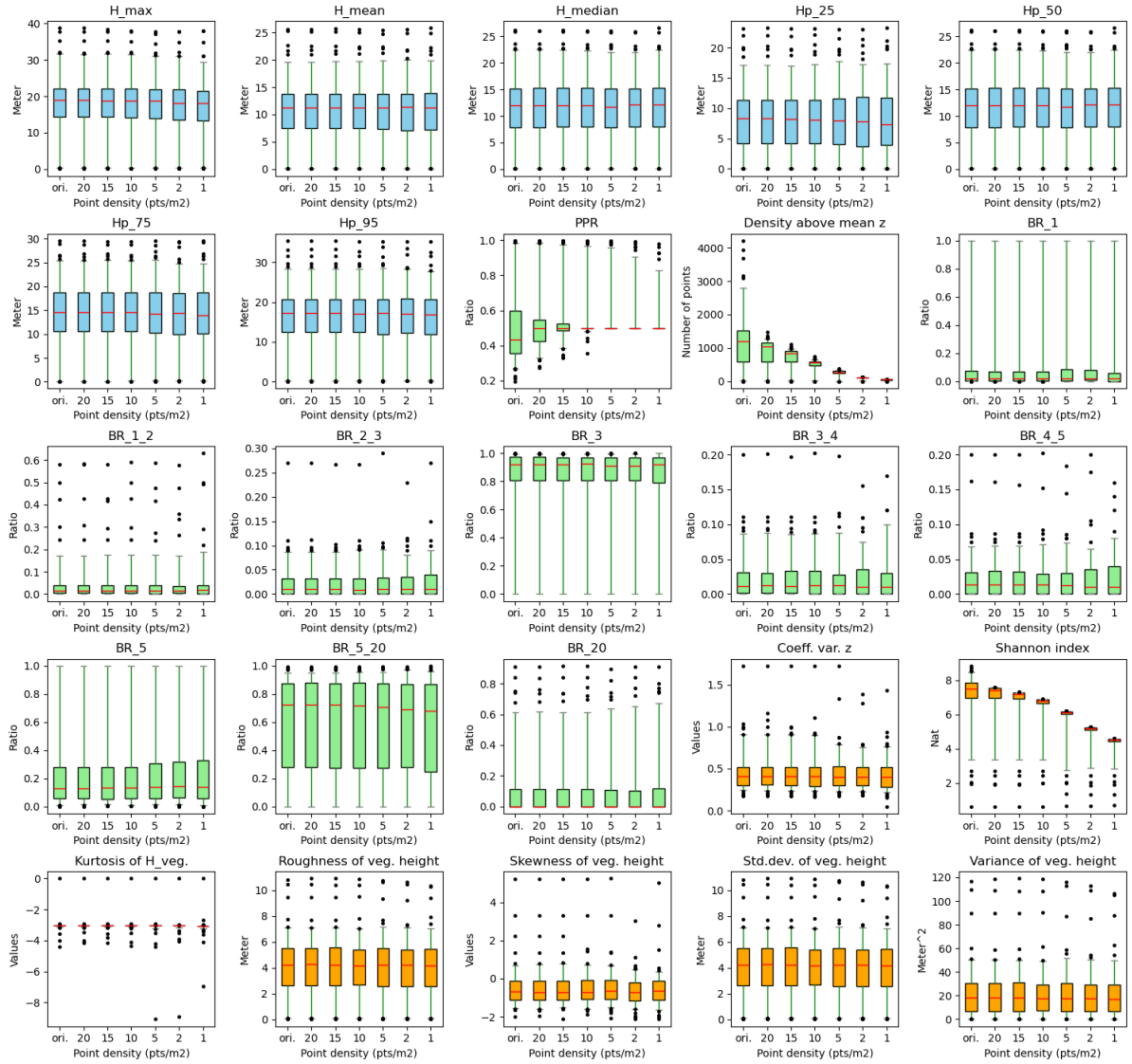

Figure A3: Twenty-five LiDAR vegetation metrics (blue: height metrics, green: cover metrics, orange: structural complexity metrics) calculated with different point densities across plots of 10 × 10 m resolution (*n* = 96) in woodland habitats of the Netherlands. Point densities were down-sampled from the original Dutch AHN4 dataset ('ori.') to six lower point densities (note the inverted x-axes). Boxes represent the interquartile range, horizontal red lines the medians, whiskers extend to the 5th and 95th percentiles, and outliers are plotted as dots. See Annex Text A1 for methodological details and Annex Table A2 for metric explanations.