A Multisource Framework for Reliable Open Urban Tree Species Dataset from Airborne LiDAR and Field Inventory Data

Nada Hamdani[1,4], Imane Abouhat[2], Kenza Ait El Kadi[1,2], Saloua Bensiali[1,3] & Imane Sebari[1,2]

[1]Research Unit of Geospatial Technologies for a Smart Decision, Hassan II Institute of Agronomy and Veterinary Medicine. 10101 Rabat, Morocco - n.hamdani@iav.ac.ma; i.abouhat@iav.ac.ma ; k.aitelkadi@iav.ac.ma; i.sebari@iav.ac.ma
[2]Department of Photogrammetry and Cartography, School of Geomatics and Surveying Engineering. Hassan II Institute of Agronomy and Veterinary Medicine. 10101 Rabat, Morocco
[3]Department of Applied Statistics and Computer Science - s.bensiali@iav.ac.ma
[4]Société Topographie Informatique, 91000 Evry Courcouronnes, France - nada.hamdani@sti-topographie.fr

Keywords: Airborne LiDAR, Urban trees, Field Inventory, Tree species classification, Open Dataset, Multi-source.

Abstract

Urban trees play a vital role in sustainable urban planning, biodiversity, conservation and climate resilience. The expanding availability of Light Detection and Ranging (LiDAR) data offers new opportunities for three-dimensional analysis of urban vegetation. However, the development of deep learning methods for tree species classification remains constrained by the lack of annotated datasets at the individual-tree level. This study proposes a robust and reproducible workflow for generating annotated urban tree point cloud datasets by integrating high density airborne LiDAR data with field inventory information, followed by field validation to ensure reliability. The methodology includes preprocessing of both point cloud and field inventory data: noise and outlier removal, vegetation filtering, elevation normalizing using a Digital Terrain Model, and extraction of acquisition dates for temporal consistency. Field inventory records were cleaned by removing duplicates, harmonizing species names, and eliminating incomplete or spatially inconsistent entries. Individual trees were then segmented from the point clouds using a combined DBSCAN and Watershed approach, and matched to inventory records through a Nearest Neighbor method. A field verification confirmed that the retained trees had not changed between the inventory and study dates. The workflow was applied to create a dataset of 152 individual urban trees across six species. The results demonstrate its effectiveness for integrating multi-sources geospatial data and producing high-quality annotated datasets, which will be made publicly available to support open science, reproducibility, and future applications in urban tree species classification, smart cities and urban digital twins.

1. Introduction

Trees are essential in urban environments. They regulate the climate, improve quality of life, purify the air and mitigate the urban heat effect. They refine human health and enhance the resilience of cities to climate change (Liang and Huang, 2023; O'Brien et al., 2022; Oliveira et al., 2026). Moreover, urban trees reduce stormwater runoff and other ecosystem function (Kasikam et al., 2026). However, these benefits strongly depend on tree species, recent advances further emphasize that tree species classification is crucial to assess ecosystem services and support sustainable city development (Li et al., 2026).

The increasing accessibility and coverage of airborne LiDAR datasets, such as the geoportal of the National Institute of Geographic and Forest Information (IGN) in France, or the Directorate General for Territory (DGT) in Portugal, provides high resolution point clouds covering entire territories. These points clouds have approximately 10 points per square meter, and include geometric and radiometric attributes such as intensity, RGB channels and near infrared (NIR), enabling a detailed characterization of the three-dimensional structure and spectral properties of vegetation.

Precise classification of tree species in a large urban environment requires airborne LiDAR point clouds combined with tree species inventories (Gong et al., 2023). Currently, such databases remain rare, which limited the development of automated methods based on deep learning models applied to point clouds. In this context, this study creates an innovative database by combining airborne LiDAR point clouds from the IGN portal with field inventories provided by Grand Paris Sud Urban Community. The aim is to provide a robust resource for training and validation classification trees species models.

2. Related Works

The use of airborne LiDAR data for tree species classification highlights a lack of resources in terms of individual tree segmentation and botanic annotations suitable for deep learning models. Several datasets have been proposed, but each presents significant limitations for tree-level species classification.

The PureForest dataset is one of the most extensive publicly datasets for identifying forest tree species in France using airborne LiDAR data and orthophotos. It covers 339 km² across 449 forests, but labels are assigned only to 50 x 50 m forest patches, without segmentation or labelling at the individual tree level, which limits its use for 3D models (Gaydon and Roche, 2024).

The FOR-instance dataset focuses on the fine segmentation of individual trees using drone LiDAR data; with over 1 130 trees manually annotated according to their structural components (trunk, branches and crown). However, it misses taxonomic annotations, which prevents it being used for species classifications tasks (Puliti et al., 2023).

Other datasets, such FOR-species20K, compile a large number of trees (over 20 000) from various platforms: Terrestrial Laser Scanning (TLS), Mobile Laser Scanning (MLS) and Unmanned Aerial Vehicle (UAV). Although they are valuable or the comparative evolution of tree classifications algorithms, their large size and heterogeneous density of data points can pose challenges in terms of data processing and model training (Puliti et al., 2024).

The MS-ALS-SPECIES benchmark, acquired using multispectral airborne LiDAR, includes 6 326 trees across nine species. While valuable for testing classification algorithms, variations in point density and sensor-specific characteristics may affect model generalization (Taher et al., 2026).

More recently, StreetTree introduces a large-scale image dataset covering more than 12 million street trees in 133 countries, annotated according to a complete taxonomic hierarchy (order, family, genus, species) for species classification from images. Although this resource represents a significant advance for classification tasks in urban environments, it does not rely on 3D LiDAR point clouds, which limits its direct comparability with methods deep learning models based on point clouds (Li et al., 2026).

Overall, these datasets demonstrate the lack of resource that combine individual tree segmentation in airborne point clouds with precise species annotations, which are required for developing deep learning models for tree species classification using airborne LiDAR data.

This research aims to address this gap by proposing a robust and reproducible multi-source workflow for generating annotated urban tree point clouds. The proposed framework combines high-density airborne LiDAR data with municipal field inventories to enable individual tree point cloud segmentation and species labelling. It further incorporates a temporal consistency step and integrates spatial matching with on-site field surveys to ensure the reliability of the generated labels.

3. Materials and Methods

In this section, we present materials and methods used to build a reliable dataset for urban tree species analysis. The overall workflow includes data collection (airborne LiDAR data and a municipal tree inventory), preprocessing, tree segmentation, spatial matching and field validation. Figure 1 illustrates the overview of the proposed approach.

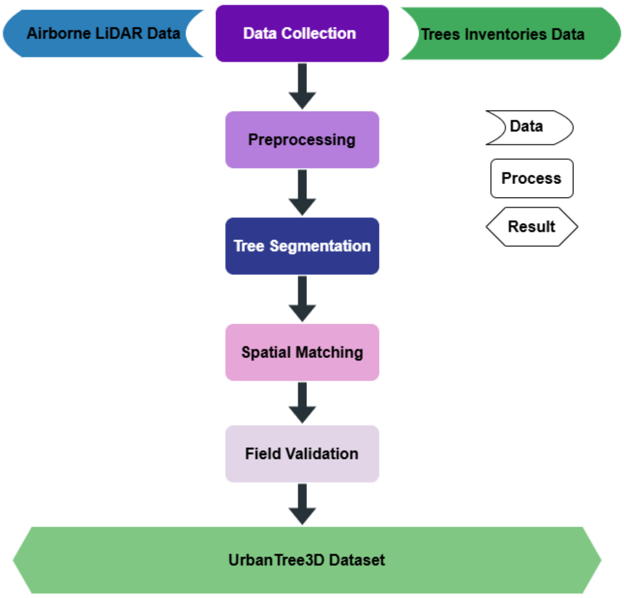

Figure 1. Workflow of the proposed approach

3.1 Study Area

The study was conducted in the municipality of Coudray-Montceaux in France. This urban area presents a varied vegetation structure with multiple tree species.

A smaller portion of the municipality, covering roughly 4 hectares, was selected for this work. The study area is illustrated in Figure 2.

Figure 2. Study area: part of Municipality of Coudray-Montceaux seen in Google Earth

3.2 Data Used

3.2.1 Airborne LiDAR Data

The airborne LiDAR data were obtained from the IGN Geoportal (French National Institute of Geographic and Forest Information). The dataset has a minimum point density of approximately 10 points per square meter, corresponding to the High Density (HD) product, in contrast to the older BD LiDAR products, which have a density around 2 points/m². This higher density enables a more detailed description of urban vegetation, this also leads to heavy data volumes, with tile sizes ranging between 500 MB to 2GB. Figure 3 presents a subset of the study area covered by the airborne LiDAR dataset.

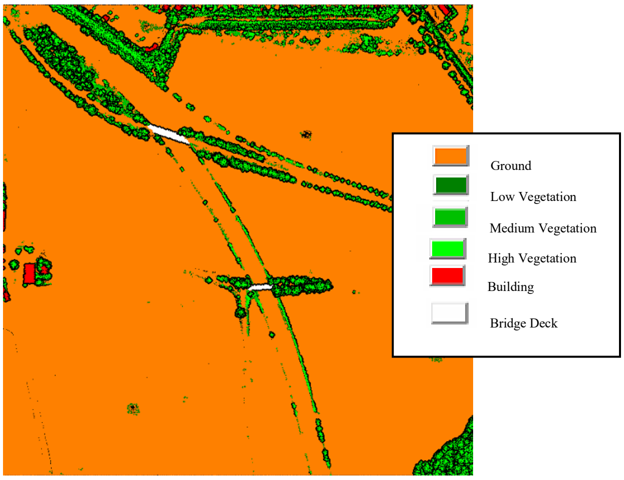

Figure 3. Subset of the study area covered by airborne LiDAR data downloaded from the IGN Geoportal.

Each LiDAR point contains a "gps_time" field following ASPRS standards, referenced to a fixed epoch starting on September 14, 2011. In this study, the dataset used was acquired on January 27, 2024.

3.2.2 Field Inventory Data

The reference field data were collected by botanical experts and provided by the Grand Paris Sud Urban Community, through its GIS department.
For each recorded tree, the dataset includes an identifier, location, species name in French and Latin, and the date of acquisition. An excerpt of the dataset is presented in Table 1.

Tree ID	Municipality	Position X(m) Y(m) Lambert93	Species (Latin)	Creation Date
13555	Le Coudray-Montceaux	661635,242 6829998,438	*Tilia platyphyllos*	3/9/2024
13556	Le Coudray-Montceaux	661653,292 6830007,691	*Tilia platyphyllos*	3/9/2024
13557	Le Coudray-Montceaux	661660,007 6829990,869	*Tilia platyphyllos*	3/9/2024

Table 1. Some attributes extracted from field inventory data

The inventory was carried out on September 3, 2024. We used this information to check the temporal consistency between the LiDAR point clouds and the inventory, ensuring that only structures existing at the time of LiDAR acquisition were considered.

3.3 Data Preprocessing

3.3.1 Airborne LiDAR data:

The airborne LiDAR point clouds were acquired from the IGN geoportal and provided with a semantic segmentation generated using the Myria3D deep learning framework (Gaydon, 2022). As with any deep learning-based segmentation approach, residual classification uncertainties may occur, particularly at the boundaries between spatially adjacent objects, such as trees located near buildings. However, this effect is limited in the present study, as the surveyed trees mainly correspond to isolated or aligned trees along streets and avenues.
A Statistical Outlier Removal (SOR) filter was applied to reduce the influence of isolated points (Rusu and Cousins, 2011). The mean distance to the k nearest neighbours was computed for each point cloud, and the point was classified as an outlier if this mean distance exceeded $\mu_d + \alpha . \sigma_d$ of the global distance distribution. Where μ_d and σ_d are the mean and standard deviation, respectively, of the local mean neighbour distances computed across the entire point cloud. Meanwhile, α is a tolerance parameter that controls the outlier detection threshold.
Vegetation points were then extracted and classified into three classes according to ASPRS standard, low vegetation (0.1-0.5m), medium vegetation (0.5-2m) and high vegetation (points above 2m).
To better represent tree structure, point elevations were normalized using a Digital Terrain Model (DTM), generated by triangulation-based interpolation. Figure 4 illustrates the preprocessing steps applied to the airborne LiDAR data.

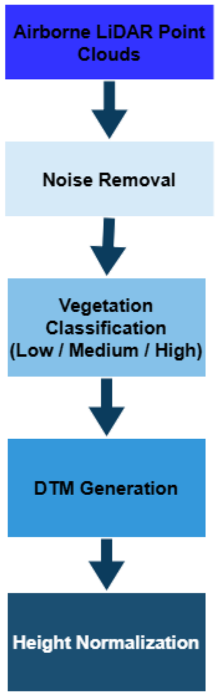

Figure 4. Airborne LiDAR Data pre-processing

3.3.2 Inventory data:

The field inventory was pre-processed to ensure data quality. Duplicates records were removed based on identical tree identifiers and spatial coordinates. Species names were standardized to ensure taxonomic consistency, and entries with incomplete or inconsistent information (invalid dates or erroneous coordinates) were discarded. A spatial quality control was performed to ensure that all positions were correctly projected in the Lambert-93 reference system. Species represented by fewer than ten individuals were excluded to improve class representativeness. The main steps of the field inventory pre-processing are illustrated in Figure 5.

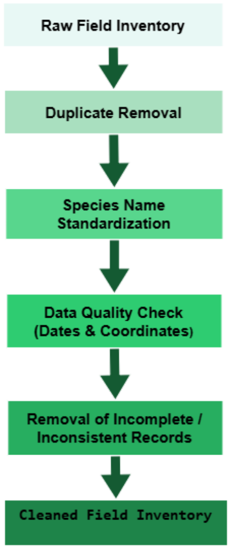

Figure 5. Workflow of field inventory data pre-processing

3.4 Tree Segmentation, Tree Matching & Field Validation:

3.4.1 Tree Segmentation:

Tree-level objects were delineated from the LiDAR point cloud using a hybrid approach. First, DBSCAN clustering (Density Based Spatial Clustering of Applications with Noise) was applied to identify dense point groupings corresponding to vegetation while filtering out noise (Ester et al., 1996). The DBSCAN parameters were empirically selected according to the density of the HD LiDAR data, with a neighborhood distance (ε) of 1.5 m and a minimum cluster size (min_samples) of 10 points. In a second step, Watershed segmentation algorithm was applied to a 1 m resolution Canopy Height Model (CHM). Local maxima were detected using a 3 x 3 pixel neighborhood window to initialize the Watershed algorithm, which refined crown boundaries and improved the separation of neighboring trees (Popescu and Wynne, 2004).

3.4.2 Tree Matching:

Each extracted tree was associated with the corresponding field inventory record to assign a species label and build the reference dataset.
The LiDAR trees were spatially matched with the inventory data by overlaying both datasets. Trees present in the inventory but not detected in the LiDAR data were excluded.
Matching was performed using a nearest neighbor search based on planimetric coordinates (X, Y). A distance threshold of 2 m was applied to minimize incorrect associations (Chirici et al., 2016; Räty et al., 2018). This distance was chosen to exceed the combined positional uncertainty of the field inventory and LiDAR-derived tree positions while remaining below the typical spacing between adjacent street trees.

3.4.3 Field Validation:

Due to the temporal gap between the LiDAR acquisition (January 2024), the field inventory (September 2024), and the study period (March 2026), an additional validation step was conducted.
The matched trees were exported in KML format and inspected in the field to verify their presence and detect possible changes such as removal or structural alteration. This verification step helped ensure the reliability of the final dataset.

4. Results

4.1 Tree Segmentation Results

Individual trees were segmented using a combined DBSCAN and Watershed approach. The DBSCAN algorithm grouped points into dense clusters, while filtering out noise. The Watershed algorithm was subsequently applied to refine the segmentation by delineating individual tree crowns and identifying tree locations. This hybrid segmentation produced 302 individual tree instances (Figure 6). The quantitative and qualitative evaluation of the hybrid DBSCAN-Watershed segmentation method, are described in detail in our methodological study (Hamdani et al., 2026a). In the present work, the resulting individual tree instances served as the input for the subsequent matching with field inventory.

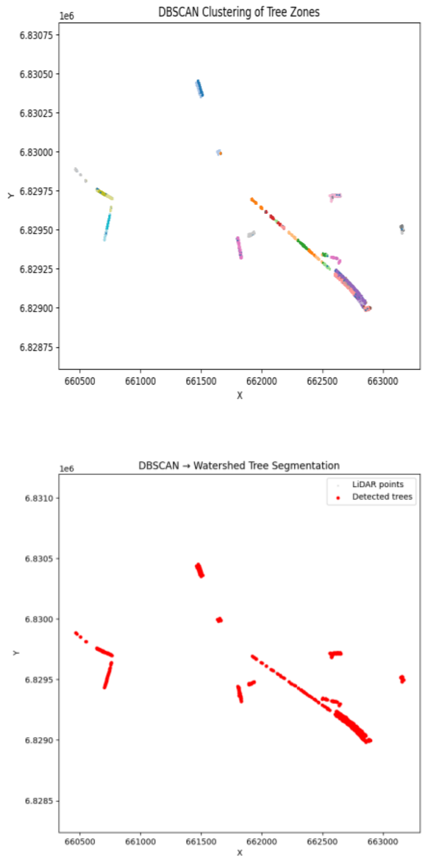

Figure 6. Tree segmentation results using a DBSCAN-Watershed approach: (left) DBSCAN clustering of LiDAR point clouds highlighting tree zones; (right) refined segmentation showing detected individual trees after Watershed processing.

4.2 Trees Matching Results

We observed strong correspondence between LiDAR detected trees and inventory trees. This confirms the accuracy of tree segmentation point clouds and validate also the field inventory as reliable reference data. Figure 7 shows the overlay of LiDAR point clouds with the extracted tree centroids (green circles) and the corresponding locations from the field inventory (red circles). Figure 8 illustrates a detailed view of tree 13397.

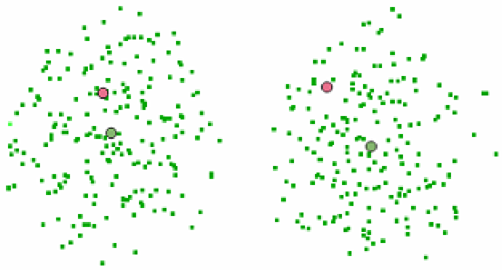

Figure 7. Spatial overlay of LiDAR point clouds, extracted tree centroids (green circles), and field inventory tree positions (red circles).

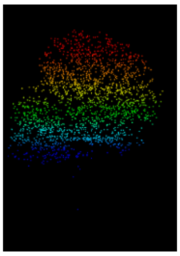

Figure 8. Visualization of tree 13397 (*Aesculus hippocastanum)* displayed by elevation.

The hybrid segmentation approach generated 302 trees from the LiDAR point cloud. These objects were matched to field inventory records using a nearest-neighbour strategy with a 2 m distance threshold. After removing unmatched objects, rare species, the final dataset consisted of 152 individual trees representing six species. Table 1 summarizes the matching performance and positional agreement between LiDAR-derived tree centroids and field inventory locations.

Metric	Value
LiDAR-detected tree objects	302
Matched tree objects (< 2 m)	166
Unmatched tree objects (> 2 m or no correspondence)	136
Mean centroid-to-inventory distance	0.83 m
Standard deviation	0.46 m
Median distance	0.77 m
Maximum distance to inventory distance	1.96 m
Matches with distance < 1 m	68%
Final validated trees	152 trees
Number of species	6

Table 1. LiDAR-field inventory matching statistics.

4.3 Field Validation Results:

A field verification of 152 detected trees was conducted on one day. Each tree was located using GPS coordinates. Observations confirmed that all trees detected were indeed present in the ground. Figure 9 shows an example of tree observed in the field, alongside its representation and position in the LiDAR data and field inventory.

Figure 9. Example of a tree observed in the field and its corresponding entries in LiDAR and field inventory data.

4.4 Dataset Overview:

The produced dataset consists of 152 individual urban trees derived from a combination of airborne LiDAR, a field inventory and field verification. After the preprocessing and cleaning steps, a total of six tree species is represented, providing sufficient diversity for further deep learning processing.

We observe that the distribution of species in resulting dataset is unbalanced. *Aesculus hippocastanum* is the most abundant species with 62 individuals, while *Prunus serrulata* is the least represented with 12 individual trees. This imbalance between classes may influence the performance of classification models. Hence the importance of data preparation steps, particularly data augmentation, before applying the deep learning model.
The use of this dataset for tree species classification, including the evaluation of a deep learning-based baseline model, as well as the contribution of data augmentation strategies to improving the model's generalization and classification performance, are presented separately in the article dedicated to the dataset (Hamdani et al., 2026b)

5. Discussion

The strong alignment observed between the LiDAR data and the field inventory highlights two key aspects. First, it emphasizes the important role of the preprocessing step, which enabled the removal of a significant number of outliers and contributed to improve the overall data quality. Second, it confirms the reliability of the methods used for tree extraction and tree matching.

This initial database can be considered as a pilot study. Future work will focus on expanding the study area in order to increase the number of samples. In addition, in areas where airborne LiDAR data are unavailable but field inventory data exist, drone LiDAR acquisitions will be conducted in order to increase the number of trees recorded and thus build a larger, more representative database.

However, an imbalance in species distribution is observed, with certain species being significantly overrepresented, such as *Aesculus hippocastanum* and *Tilia platyphyllos.* If this database is to be used as training deep learning models, this imbalance could lead to biased predictions. Therefore, the application of data augmentation techniques will be necessary to mitigate this issue to compensate. Furthermore, the planned expansion of the study area should reduce differences between species and build a more balanced database that meets the requirements of deep learning models.

6. Conclusion

Urban trees play a crucial role in improving environmental quality of cities. In this study, we combined airborne LiDAR

https://doi.org/10.5194/isprs-annals-XII-4-W1-2026-195-2026

data with field inventory data in order to create a reliable database for urban tree species classification.

The results show a high degree of agreement between the two data sources, confirming the effectiveness of proposed workflow, including preprocessing, tree segmentation, tree matching, and field verification. This work thus addresses the current lack of high-quality datasets needed to develop methods for tree species classification, particularly using deep learning techniques.

By providing a structured, a validated and open-source database, this study constitutes a solid foundation for future work in urban forestry and remote sensing. It also opens new avenues for the automated classification of tree species in urban environments, as well as for the development of applications within the context of smart cities and urban digital twins, where precise knowledge of vegetation is essential.

Future prospects include expanding the study area, improving species balance, and using this database to train advanced species classification model in urban environments.

Acknowledgements

The authors would like to express their sincere gratitude to the "Communauté d'Agglomération Grand Paris Sud (CA GPS)", particularly the GIS division of the Department of Foresight, Territorial Observation and GIS (Direction de la Prospective, de l'Observation Territoriale et du SIG), for their valuable support and active involvement throughout this research.

References

Chirici, G., Mura, M., McInerney, D., Py, N., Tomppo, E.O., Waser, L.T., Travaglini, D., McRoberts, R.E., 2016. A meta-analysis and review of the literature on the k-Nearest Neighbors technique for forestry applications that use remotely sensed data. Remote Sensing of Environment 176, 282–294. https://doi.org/10.1016/j.rse.2016.02.001

Direção-Geral do Território (DGT), LiDAR Portal (DGT), https://cdd.dgterritorio.gov.pt/dgt-fe , accessed 24 March 2026.

Ester, M., Kriegel, H.-P., Xu, X., 1996. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. 1996.

Gaydon, C., 2022. Myria3D. https://github.com/IGNF/myria3d (accessed 13 July 2026).

Gaydon, C., Roche, F., 2024. PureForest: A Large-Scale Aerial Lidar and Aerial Imagery Dataset for Tree Species Classification in Monospecific Forests. https://doi.org/10.48550/arXiv.2404.12064

Gong, Y., Li, X., Du, H., Zhou, G., Mao, F., Zhou, L., Zhang, B., Xuan, J., Zhu, D., 2023. Tree Species Classifications of Urban Forests Using UAV-LiDAR Intensity Frequency Data. Remote Sensing 15, 110. https://doi.org/10.3390/rs15010110

Hamdani, N., Abouhat, I., Ait El Kadi, K., Bensiali, S., Sebari, I., 2026a. Hybrid Segmentation of Urban Trees from Airborne LiDAR Data Using DBSCAN and Watershed Algorithms. E3S Web Conf. 708, 03009. https://doi.org/10.1051/e3sconf/202670803009

Hamdani, N., Abouhat, I., Ait El Kadi, K., Bensiali, S., Sebari, I., 2026b. UrbanTree3D: An Open Dataset for Urban Tree Species Classification Using Airborne LiDAR and Field Inventory Data. Data 11. https://doi.org/10.3390/data11060147

IGN, Geoportal of the National Institute of Geographic and Forest Information, https://cartes.gouv.fr/telechargement/IGNF_NUAGES-DE-POINTS-LIDAR-HD, accessed 24 March 2026.

Kasikam, N., Yarnvudhi, A., Leksungnoen, N., Näsholm, T., Tor-ngern, P., 2026. Effects of long term canopy change on regulating ecosystem services in a tropical urban park. Sci Rep 16, 5077. https://doi.org/10.1038/s41598-026-36098-w

Li, J., Huang, Y., Zhang, F., Liu, Y., 2026. StreetTree: A Large-Scale Global Benchmark for Fine-Grained Tree Species Classification. https://doi.org/10.48550/arXiv.2602.19123

Liang, D., Huang, G., 2023. Influence of Urban Tree Traits on Their Ecosystem Services: A Literature Review. Land 12, 1699. https://doi.org/10.3390/land12091699

O'Brien, L.E., Urbanek, R.E., Gregory, J.D., 2022. Ecological functions and human benefits of urban forests. Urban Forestry & Urban Greening 75, 127707. https://doi.org/10.1016/j.ufug.2022.127707

Oliveira, W., Silva, J.L.S., Tabarelli, M., Lopes, A.V., 2026. Benefits of urban trees to people and their potential contribution to all the 17 sustainable development goals. Trees, Forests and People 24, 101198. https://doi.org/10.1016/j.tfp.2026.101198

Popescu, S.C., Wynne, R.H., 2004. Seeing the Trees in the Forest: Using Lidar and Multispectral Data Fusion with Local Filtering and Variable Window Size for Estimating Tree Height. photogrammetric engineering and remote sensing 70, 589–604.

Puliti, S., Pearse, G., Surový, P., Wallace, L., Hollaus, M., Wielgosz, M., Astrup, R., 2023. FOR-instance: a UAV laser scanning benchmark dataset for semantic and instance segmentation of individual trees. https://doi.org/10.48550/arXiv.2309.01279

Puliti, S., Lines, E.R., Müllerová, J., Frey, J., Schindler, Z., Straker, A., Allen, M.J., Winiwarter, L., Rehush, N., Hristova, H., Murray, B., Calders, K., Terryn, L., Coops, N., Höfle, B., Junttila, S., Krůček, M., Krok, G., Král, K., Levick, S.R., Luck, L., Missarov, A., Mokroš, M., Owen, H.J.F., Stereńczak, K., Pitkänen, T.P., Puletti, N., Saarinen, N., Hopkinson, C., Torresan, C., Tomelleri, E., Weiser, H., Astrup, R., 2024. Benchmarking tree species classification from proximally-sensed laser scanning data: introducing the FOR-species20K dataset. https://doi.org/10.48550/arXiv.2408.06507

Räty, J., Packalen, P., Maltamo, M., 2018. Comparing nearest neighbor configurations in the prediction of species-specific diameter distributions. Annals of Forest Science 75, 26. https://doi.org/10.1007/s13595-018-0711-0

Rusu, R.B., Cousins, S., 2011. 3D is here: Point Cloud Library (PCL), in: 2011 IEEE International Conference on Robotics and Automation. Presented at the 2011 IEEE International Conference on Robotics and Automation (ICRA), IEEE, Shanghai, China, pp. 1–4. https://doi.org/10.1109/ICRA.2011.5980567

Taher, J., Hyyppä, E., Hyyppä, M., Salolahti, K., Yu, X., Matikainen, L., Kukko, A., Lehtomäki, M., Kaartinen, H., Thurachen, S., Litkey, P., Luoma, V., Holopainen, M., Kong, G., Fan, H., Rönnholm, P., Vaaja, M., Polvivaara, A., Junttila, S., Vastaranta, M., Puliti, S., Astrup, R., Kostensalo, J., Myllymäki, M., Kulicki, M., Stereńczak, K., Pires, R. de P., Valbuena, R., Carbonell-Rivera, J.P., Torralba, J., Chen, Y.-C., Winiwarter, L., Hollaus, M., Mandlburger, G., Takhtkeshha, N., Remondino, F., Lisiewicz, M., Kraszewski, B., Liang, X., Chen, J., Ahokas, E., Karila, K., Vezeteu, E., Manninen, P., Näsi, R., Hyyti, H., Pyykkönen, S., Hu, P., Hyyppä, J., 2026. Multispectral airborne laser scanning for tree species classification: A benchmark of machine learning and deep learning algorithms. ISPRS Journal of Photogrammetry and Remote Sensing 233, 278–309. https://doi.org/10.1016/j.isprsjprs.2026.01.031