Benchmarking Tree Instance Segmentation Algorithms: Operationalizing Algorithm Selection Across Forest Conditions and Platforms

Stefan Reder Koithahn[a], Josafat-Mattias Burmeister[b], Andreas Tockner[c], Markus Engel[d], Rico Richter[b], Jan-Peter Mund[a], Jürgen Döllner[b,e]

[a]*Eberswalde University of Sustainable Development, Department Forest-Wood-Environment, Eberswalde, Germany*
[b]*University of Potsdam, Digital Engineering Faculty, Potsdam, Germany*
[c]*BOKU University, Department of Ecosystem Management, Climate and Biodiversity, Institute of Forest Growth, Wien, Austria*
[d]*State Forestry Research Centre, Eberswalde, Germany*
[e]*Hasso Plattner Institute, Potsdam, Germany*

Keywords: 3D Point Clouds, Tree Instance Segmentation, LiDAR, Forest structure Stand descriptors, Decision Support

---

1. Objective

Tree instance segmentation (TIS) in LiDAR point clouds has advanced rapidly, driven in particular by deep learning. Despite these gains, operational use in forestry remains challenging because segmentation accuracy is highly sensitive to forest structural complexity, acquisition conditions, and LiDAR platform characteristics of UAV-borne (ULS), terrestrial (TLS) and SLAM-based (PLS) systems. This variability makes it difficult to select a method that performs reliably for a given task and site.

In this study, we benchmark 11 state-of-the-art TIS algorithms across 55 forest plots from publicly available LiDAR datasets. We organize plots using key stand and platform descriptors to

1. quantify dataset imbalances and potential benchmark bias and
2. analyze how algorithm performance changes across stand characteristics and acquisition settings.

The contribution addresses forest managers, service providers, and researchers by providing a decision-support framework for selecting robust TIS methods across realistic forest and sensor conditions instead of relying on a single global "best" algorithm.

2. Methods

We evaluated 11 TIS methods, comprising traditional algorithmic approaches and deep learning-based models, on 55 publicly available forest plots acquired with ULS, TLS, and PLS platforms. Plots were categorized using acquisition properties (platform type, season) and qualitative stand descriptors (e.g., forest type/species composition, canopy closure, understory density, vertical stratification, stand development stage).

Segmentation quality was quantified using both instance detection metrics and instance segmentation metrics. To test for statistically significant differences among algorithms while accounting for structured variability across datasets and plots, we applied parametric and non-parametric tests, including mixed-effects models. In addition, we trained a Random Forest model to identify the most influential descriptors driving performance variation. Feature importance analysis provides an interpretable link between stand/acquisition characteristics and expected segmentation quality, forming the basis for decision support recommendations.

3. Results

Benchmark bias and imbalance. The publicly available datasets are highly unbalanced, with a predominance of stands characterized by moderate canopy closure with gaps/openings and sparse or absent understory. Dense-canopy and dense-understory stands are underrepresented. Over 93% of stands fall within timber or mature development stages, and most, esp. ULS plots are smaller than 0.1 ha, which can bias global performance summaries toward conditions that are easier to segment. Further, algorithms which are sensitive to edge effects are in disadvantage due to small plot sizes.

Performance depends on platform, stand structure, and acquisition. Segmentation performance varies strongly with stand complexity, acquisition conditions, and algorithm choice. ULS in leafless, open forests with sparse understory achieved the best overall performance (mIoU = 0.75, F1 = 0.95). In denser and more occluded forests, TLS and PLS tended to perform better, reflecting their ability to capture lower canopy and understory structure in challenging conditions. Across structural categories, single-layered stands with absent understory achieved the highest performance (mIoU = 0.68, F1 = 0.91), whereas two-layered stands with dense understory performed substantially worse (mIoU = 0.50, F1 = 0.65).

Key predictors and decision-oriented recommendations. Descriptor importance differs by platform: stand development stage is the dominant predictor for TLS performance, while understory density and canopy closure are the main predictors for PLS and ULS, respectively. These relationships translate into actionable method guidance: for ULS, ForestFormer3D is recommended (mIoU = 0.80), except in deciduous stands with dense canopies where treeiso outperforms (mIoU = 0.84). For TLS, ForestFormer3D performs best in coniferous stands (mIoU = 0.64), while RayCloudTools is strongest in deciduous stands (mIoU = 0.62). For PLS, ForAINet is most effective for timber and older stands (mIoU = 0.69).

4. Operational Practicality and Significance

The decision support framework is intended for operational use relying on stand descriptors that can be field-assessed, taken from inventories, or inferred from point clouds. It enables practitioners to select the proper platform and algorithm according to forest structural conditions and risk-aware interpretation given benchmark imbalances, and targeted quality assurance for known failure cases (e.g., dense understory, multi-layer canopies).

Overall, the study delivers a cross-platform benchmark and exposes structural bias in public datasets. Combining statistical tests with descriptor importance yields interpretable, condition-dependent recommendations that improve method choice, anticipate failure modes, and strengthen the reliability of LiDAR-based tree-level products for forest inventory and monitoring.