Comprehensive Synthetic Satellite Forest Canopy Dataset for Height Estimation

Gil Isaac E. Gabriel

Department of Geography and Planning, West Chester University of Pennsylvania

Author Note

Gil Isaac E. Gabriel https://orcid.org/0000-0003-2414-8547

This project is funded by Grant 2244304 from the National Science Foundation. I thank Yang Tang, Dr. Rongjun Qin, and Dr. Jieun Hur for their guidance throughout the research process.

Correspondence regarding this article should be addressed to Gil Isaac E. Gabriel, 506 Business and Public Management Center, 50 Sharpless Street, West Chester, PA 19383. Email: isaac@gilisaacgabriel.com.

Abstract

Tree canopy height is an increasingly important measure due to its ability to inform climate change mitigation strategies. However, current research on tree canopy height estimation on a global scale has indicated limited predictive accuracy due to insufficient remote sensing data. For instance, current LiDAR data from the Global Ecosystems Dynamics Investigation (GEDI) have low spatiotemporal resolutions and only cover about 4% of the land surface. Machine Learning (ML) algorithms that predict tree canopy heights rely on LiDAR and satellite imagery data. Therefore, although satellite imagery is comprehensive and has high spatiotemporal resolutions, the lack of comprehensive quality LiDAR data reduces the ability of ML algorithms to produce accurate estimations of tree canopy height globally. Thus, the purpose of this research project is to leverage computer graphics technology to create synthetic environments with accurate tree canopy height measurements. 30,000 3D scenes will be generated using an automated workflow and will simulate diverse environments with tree type and height, leaf density, lighting condition, and seasonal variation as parameters. A multispectral (RGB) image, Digital Surface Model, and Digital Elevation Model are extracted from each synthetic scene. This could serve to train ML algorithms as ground truth and improve the accuracy of tree canopy height estimations. More precise tree canopy height measurements will improve ecological assessments such as global carbon stock assessments, informing climate change mitigation, land use management, and conservation efforts.

Comprehensive Synthetic Satellite Forest Canopy Dataset for Height Estimation

Literature Review

Canopy height estimation is crucial for sustainability initiatives and monitoring ecological changes. Accurate measurements of forest canopy height provide valuable insights into carbon stock and ecosystem services, which are essential for climate change mitigation and conservation efforts. Forests play a vital role in absorbing carbon dioxide from the atmosphere, making them important carbon sinks. Accurate tracking of carbon absorption by forests has significant implications for climate change mitigation, resilience, and global warming (**Lang2023**). Canopy height is a key measure that provides information about ecological changes and carbon stock (**Potapov2021**; **Tang2024**). **Potapov2021<empty citation>** demonstrate the ability of canopy height maps to monitor ecological changes, such as differences in tree heights between natural forests and suburban areas, secondary forests, and plantations.

Approaches to Canopy Height Mapping

Traditionally, canopy height mapping has been done through field observations, which are time-consuming and labor-intensive. This method is limited in scale and cannot be applied globally, compromising researchers' ability to assess carbon storage on a global scale and monitor progress towards climate goals and emission targets. To address this issue, researchers have turned to remote sensing data and Machine Learning (ML) techniques for global canopy height estimation. **Lang2023<empty citation>** produced a global canopy height model by combining Sentinel-2 optical satellite images with height data from the Global Ecosystem Dynamics Investigation (GEDI) LiDAR mission. Similarly, **Potapov2021<empty citation>** created a global canopy height map using Landsat imagery fused with GEDI RH95 metric, reserving 10% of GEDI data for validation and using Airborne Laser Scanning (ALS) as an independent reference.

Challenges and Limitations:

Despite the advancements in remote sensing technology, there are still challenges and limitations in canopy height estimation. GEDI provides sparse elevation data due to limited positioning and movement, and the mission only began in 2019, resulting in a limited collection of quality, analysis-ready data. In fact, only approximately 4% of land cover is captured by GEDI LiDAR (**Lang2023**). Furthermore, remote sensing data often contains noise from cloud and/or snow cover, requiring high spatiotemporal resolutions and image patching to produce comprehensive data products.

The sparse GEDI data has produced uncertainties that are more likely to affect certain regions over others. Regions in high latitudes such as Alaska, Yukon, and Tibet have high predictive uncertainty, with the latter two likely due to local characteristics that **Lang2023**'s **Lang2023** model has not encountered in training as they lie outside of GEDI coverage.

Current models, such as the Canopy Height model by **Lang2023<empty citation>**, struggle with height estimation due to factors like the limited amount of available data for model validation. Improving these models will require more comprehensive datasets that are inclusive of diverse environmental features and conditions. Models are likely to inaccurately estimate canopy heights in regions that are underrepresented in the training data.

Researchers are working to address the limitations in canopy height estimation by creating synthetic datasets that can be used for model calibration. For example, TreeNet3D is a recent dataset generated through a fully automated process, containing 1,300 instances each of 10 different tree species (**Tang2024**). Such datasets can help improve the accuracy and generalizability of canopy height estimation models.

Canopy height estimation is a critical tool for monitoring ecological changes and assessing carbon stock, making it essential for sustainability initiatives and climate change mitigation efforts. While traditional field observations are time-consuming and limited in scale, modern approaches utilizing remote sensing data and machine learning techniques have enabled global canopy height mapping. However, challenges such as sparse elevation data, noise in remote sensing imagery, and limited training data representing diverse environments still persist. Ongoing efforts to create comprehensive synthetic datasets offer promise for improving the accuracy and applicability of canopy height estimation models in the future.

Methods

General framework.

I utilized Unreal Engine 5.4 to generate a forest canopy scene, simulating diverse conditions and environments. The scene include varying tree species, densities, and distributions to represent realistic forest structures. Next, I captured images in a 512×512 pixel square at a spatial resolution of 30 m by positioning a virtual camera in the scene at a constant height of {~10,000} m above ground. From this position, three simulated image types are taken: (a) optical image in natural RGB color to simulate satellite imagery; (b) a Depth Map to represent a Digital Terrain Model (DTM); and (c) another Depth Map simulating a Digital Surface Model (DSM), which adds above ground elevation data to the DTM. I created a total of 30,000 samples, amounting to 90,000 images—three images taken for each forest canopy scene. A compiled set of these images serve as the dataset which could be used for training and validating Machine Learning (ML) models and estimated canopy height products.

3D Scene Generation

Procedural Content Generation (PCG) framework. Unreal Engine version 5.4 provides a number of features that allow for an automatic placement of 3D assets such as trees onto scenes. UE5.4 presents an opportunity to batch-create scenes with varying configurations and parameters. The Foliage tool in Unreal Engine allows for the placement of Static Meshes, which are 3D models representing individual trees. These Static Meshes can be automatically scattered across a landscape based on customizable rules such as density, scale, and rotation. Additionally, the Landscape tool enables the creation of realistic terrain features like hills, valleys, and slopes, providing a diverse environment for the procedurally generated forest scenes

provides artists and designers the ability to build fast, iterative tools and content of any complexity ranging from asset utilities, such as buildings or ***biome generation***_, up to entire worlds._

Parameters

Defining the landscape

The chart outlines a detailed methodology for generating procedurally created forest scenes and rendering their outputs in a 3D scene. The process begins with defining the landscape using ground texture assets and elevation data. This landscape foundation is enhanced using the Procedural Content Generation (PCG) framework, which allows for the specification of various parameters such as tree types, biomes, leaf and branch density, time of day, light conditions, and seasonal changes.

Once the parameters are set, 3D tree assets are scattered across the landscape based on the defined criteria like density, scale, and rotation. The landscape tool further refines the scene by incorporating realistic terrain features such as hills, valleys, and slopes, ensuring the forest scenes are procedurally generated to mirror natural environments.

Next, the scene is positioned using a virtual camera, preparing it for image rendering and output. At this stage, the process allows for updates to the 3D scene by changing parameter values or modifying the landscape, ensuring flexibility and iterative refinement. The rendered outputs can be viewed in different modes, including scene depth view mode, regular lit view mode, and optical "satellite" natural RGB color images. These outputs are organized to correspond to each other, ensuring consistency across the digital elevation/terrain model, digital surface model, and optical images.

- Height range
- Per biome
- Tree Types
- Biomes
  - Biome-level analysis of canopy heights showing **distribution of heights** (Lang et al., 2023). Will use these height specifications to parameterize tree heights in the PCG.
- Leaf/Branch Density

- Time
- Light Conditions
- Seasonal Changes
- Spatial Resolution: 30 m
  - Grid size (area)
  - Goal: 512*512 pixels

Image Rendering and Output

3D Scene Modification

Automation and Batch Processing