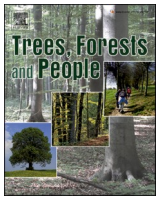
Trees, Forests and People 21 (2025) 100955

Contents lists available at ScienceDirect

Trees, Forests and People

journal homepage: www.sciencedirect.com/journal/trees-forests-and-people



Individual tree above-ground biomass estimation by integrating LiDAR and machine learning

Yan To Choi [a], Majid Nazeer [a,b], Man Sing Wong [a,b,c,*], Janet Elizabeth Nichol [d], Shao-Yuan Leu [e], Jin Wu [f,g], Amos P.K. Tai [h,i]

[a] Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Hong Kong, China
[b] Research Institute for Sustainable Urban Development, The Hong Kong Polytechnic University, Hong Kong, China
[c] Research Institute for Land and Space, The Hong Kong Polytechnic University, Kowloon, Hong Kong, China
[d] Department of Geography, School of the Environment and Life Sciences, University of Portsmouth, Hampshire, UK
[e] Department of Civil and Environmental Engineering, The Hong Kong Polytechnic University, Hong Kong, China
[f] School of Biological Sciences, The University of Hong Kong, Pok Fu Lam, Hong Kong, China
[g] Institute for Climate and Carbon Neutrality, The University of Hong Kong, Hong Kong, China
[h] Department of Earth and Environmental Sciences, The Chinese University of Hong Kong, Hong Kong, China
[i] State Key Laboratory of Agrobiotechnology, and Institute of Environment, Energy and Sustainability, The Chinese University of Hong Kong, Hong Kong, China

ARTICLE INFO

Keywords:
Allometric model
Tree felling
Tree biomass
Point-cloud

ABSTRACT

Global warming represents a critical challenge globally, while tree carbon sequestration is essential for achieving carbon neutrality. The existing global allometric models face challenges in accurately modelling local trees' biomass. To develop a localized allometric model using a small dataset, this study proposes an innovative framework for estimating tree above-ground biomass (AGB) that involves local tree felling data collection, Light Detection and Ranging (LiDAR) implementation, and the development of a machine learning-based allometric model. During the data collection period, 100 trees were felled in Hong Kong from March 2023 to April 2024, encompassing 31 tree species and 17 tree families. Point-cloud models of the felled trees were collected using a LiDAR backpack. Each felled tree's AGB was measured by integrating point-cloud technology and oven drying of samples. A data augmentation method was developed with a proposed tree point-cloud 'degrowth' algorithm to address the challenge of data limitation in allometric model development. The allometric models in this study were trained using advanced tree parameters measured by TreeQSM and tree family parameters. The best-performing allometric model developed by XGBoost, scored an accuracy of $R^2 = 0.82$, mean absolute percentage error (MAPE) = 40.70 %, and mean absolute error (MAE) = 214.37 kg. To summarize, this study enhanced AGB estimation in the local region by incorporating LiDAR, tree data augmentation, and machine learning for allometric model development.

1. Introduction

Global warming has reached a critical stage, with unprecedented record-breaking temperatures and catastrophic natural disasters occurring worldwide. The global average temperature in 2024 was estimated to be 1.54 ± 0.13 °C above the preindustrial average, marking the first year to surpass the 1.5 °C threshold (WMO, 2024). As a result, Hong Kong's Climate Action Plan 2050 was enacted, aiming to achieve carbon neutrality by 2050 (Carbon Neutrality and Sustainable Development, 2021). The main strategies in decarbonization include emission reduction and clean energy. For vegetation, tree carbon plays a pivotal role in achieving carbon neutrality by offsetting carbon emissions through carbon sequestration (Chen, 2021). Forests serve as substantial carbon sinks by absorbing and storing carbon in trees. Carbon sequestration

Abbreviations: MAE, Mean absolute error; MAPE, Mean absolute percentage error; R^2, Coefficient of determination; XGB, Extreme gradient boosting (XGBoost); AGB, Above-ground biomass; LiDAR, Light Detection and Ranging; DBH, Diameter at breast height; RF, Random forest; GBT, Gradient-boosted trees; LGBM, Light Gradient Boosting Machine (LightGBM); QSM, Quantitative structure model; DW, Dry weight; TH, Tree height; CL, Crown length; CA, Crown area; TL, Trunk length; BL, Branch length; TA, Trunk area; BA, Branch area; CBH, Crown base height.

* Corresponding author at: Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Hong Kong, China.
E-mail address: Ls.charles@polyu.edu.hk (M.S. Wong).

https://doi.org/10.1016/j.tfp.2025.100955

Available online 22 July 2025

involves photosynthesis, in which tree leaves absorb CO_2 from the atmosphere and store the absorbed carbon as tree biomass (Toochi, 2018). Therefore, estimating tree biomass becomes crucial in assessing the carbon stock of a tree. Tree biomass is a function that can be explained by wood volume, diameter, height, or other physical parameters. In tree above-ground biomass (AGB) estimation, allometric models are often employed to quantify biomass by inputs of tree physical parameters (Vieira et al., 2008). Various studies have developed allometric models for the local research areas (Chave et al., 2005; Chave et al., 2014; Djomo et al., 2010).

Existing methods in AGB estimation first conduct destructive tree harvesting and estimate the entire tree volume and dry weight by fresh weight to dry weight ratio or water displacement methods. Then these ground truth data are correlated with general tree parameters, including tree height, diameter at breast height (DBH), wood density by exponential model, and finally, constructed an allometric model (Chave et al., 2005, 2014; Djomo et al., 2010; Mugasha et al., 2016; Segura and Kanninen, 2005). Existing allometric models can mainly be stratified into mix-species and species-specific models, while the mix-species model has higher generalizability (Chave et al., 2014, 2005; Djomo et al., 2010; Sarker, 2010) and the species-specific model focuses on enhancing prediction accuracy for woodlands with specific species type (Sarker et al., 2013; Magalhães et al., 2021; Mulatu et al., 2024). A limitation in building allometric models is striking a balance between generalization and accuracy. Species-specific allometric models are built with higher accuracy for assessing homogeneous habitats, but they sacrifice applicability in high biodiversity regions (Van Wolputte, 2024). Tree species categories must inherit specific patterns in carbon stock levels (Kaul et al., 2010). Developing a model that incorporates the strengths of both mix-species and species-specific approaches can offer a more robust solution. To address this issue, an AGB estimation model that allows categorical input, including tree species or tree family, was required to enhance the applicability across diverse tree species.

Despite the high applicability of the mix-species allometric model, the reliance on simple exponential models introduces uncertainties in AGB estimation. The popular allometric model, developed by Chave et al. (2014), utilized over 4000 trees globally. This allometric model, built using the log-log regression approach, can effectively capture general and broad patterns across the large global dataset. Yet, the ability to capture homogeneous characteristics in small datasets remains uncertain. Most AGB prediction studies develop allometric model through regression or fitting on simple power law functions, linear or polynomial functions (Sileshi, 2014). When studying local habitats with mixed tree species, a more advanced approach should be explored to capture more complex characteristics of localized trees. Machine learning models have been employed in certain AGB estimation studies, offering an advantage over simple regression methods (Roy et al., 2024; Wongchai et al., 2022). Various machine learning algorithms are available for regression problems, and models can be stratified into simple models, K-Nearest Neighbors (KNN), tree-based models, and support vector machine (SVM). Simple models include linear, ridge, and lasso regressions that capture linear relationships between predictors and targets. These simple models were widely used in previous allometric model studies to predict AGB using a few simple variables, including DBH and tree height. However, these simple models failed to capture the nonlinear variations between tree parameters and AGB in a small localized tree dataset (Roy et al., 2024). The KNN model captures nonlinear relationships and creates predictions by averaging the output values in the feature space (Kohli et al., 2020). However, KNN, which is commonly used as an outlier detection algorithm, can be highly sensitive to outliers (Chen et al., 2010) if extremely large or small trees are collected in a small, localized tree dataset. SVM output predictions by using support vectors and kernels to develop a function that best fits the epsilon (Awad et al., 2015). However, SVM likely overfits data with a small number of samples and a large number of variables (Han and Jiang, 2014), and training of SVM requires careful choice of kernels and advanced hyperparameters tuning. The tree-based model includes decision trees, random forest (RF), gradient-boosted trees (GBT), XGBoost (XGB), and LightGBM (LGBM). Tree physical parameters can be presented in a tabular format, and tree ensemble models, such as XGB and RF, are highly recommended for regression and classification problems on tabular data (Grinsztajn et al., 2022). These tree-based machine learning algorithms require less computation and tuning time, and they even outperform various deep learning models (Shwartz-Ziv and Armon, 2022). For allometric model building, one of the challenges is the limited data size due to the difficulties associated with large-scale tree felling. Tree-based machine learning algorithms have been proven to work well with small datasets (Treboux et al., 2018). Based on these considerations, this will select ensemble tree models, including RF, GBT, XGB, LGBM, for the development of allometric model.

Another major challenge in developing allometric model is simple parameterization and limited tree data for training (Shi et al., 2013). Tree data acquisition involves manual tree measurement and destructive sampling, both of which are resource-demanding activities that lead to data scarcity. Existing allometric models estimate AGB by parameters such as DBH and tree height, which are easily retrievable by human measurement. For AGB estimation in local regions with high biodiversity, over-simplified parameters fail to accurately capture tree morphological features, which negatively impacts the performance of the allometric model. Given these limitations, an enhanced data collection method is required to generate sufficient data samples and acquire more complex tree parameters. Individual tree AGB estimations have been conducted in different scenarios, including estimation on the terrestrial level by manual measurement (Chave et al., 2005; Djomo et al., 2010), on terrestrial level using LiDAR technology (Calders et al., 2015; Chave et al., 2019). To address the problem of data scarcity, LiDAR technology can serve as a solution to replace manual measurements and extensive destructive sampling of trees. LiDAR point-cloud can capture the 3D geometry of an individual tree, while the 3D quantitative structure models (QSM) can reconstruct the 3D tree model, providing more complex tree parameters instead of only DBH or tree height (Lau et al., 2018). To address the data scarcity problem, the data collection approach should adopt LiDAR modelling to enable tree data augmentation and extraction of complex tree parameters.

Furthermore, the only local (Hong Kong) allometric model was developed by Sarker (2010), using 75 trees and 13 different species. Therefore, there is a pressing need to develop a new allometric model that incorporates a wider range of species and LiDAR technologies in order to provide a tool for tree AGB estimation in Hong Kong.

In summary, the development of existing allometric models faces several challenges, including data scarcity, reliance on simple modelling approaches, limitations to either species-specific or mixed-species models, reliance on simple manually measured parameters, and the absence of a local model specifically for the study region. To address these issues, this study proposes the use of LiDAR technology for tree data augmentation and advanced measurement of tree parameters, thereby constructing an enhanced machine learning-based allometric model that enables the input of tree taxa variables for AGB estimation. The specific objectives are as follows:

- Developing an integrated framework that combines laboratory oven-drying of wood samples and tree point-cloud models to address the infeasibility of oven-drying entire trees, enabling accurate measurement for above-ground biomass (AGB) of felled trees.
- Developing a point-cloud processing algorithm for tree data augmentation to address the limited data challenge.
- Building machine learning-based allometric models that incorporate tree family parameters and advanced physical parameters for precise AGB estimation in Hong Kong's diverse tree species.

2. Study area and data collection

2.1. Study area

This research took place in Hong Kong, located in southeast Asia, spanning from approximately 22°08′ and 22°35′ North latitude and 113°49′ and 114°31′ East longitude (Lands department, 2024). Hong Kong is characterized by a humid subtropical climate, influenced heavily by the South Asian monsoon (Lam et al., 2022). The average temperature ranges from 16 °C in January, the coldest month, to about 29 °C in July, the warmest month (Cheung and Hart, 2014). Such climatic conditions are conducive to a rich diversity of vegetation and affect the growth patterns and biomass accumulation in trees. Hong Kong's vegetation is predominantly subtropical. The natural forests are largely secondary and have regrown after previous agricultural and urban development use (Zhang et al., 2024).

2.2. Tree felling and samples collection

Tree felling operations were conducted predominantly in urban areas of Hong Kong. Professional tree maintenance personnel felled the tree in parts, including twigs, branches, and trunks. Felled tree parts were further trimmed into sample slices by chainsaw and immediately sealed in sample bags upon felling to prevent water loss or contamination. All types of trees in the range of healthy to unhealthy, leaf-on to leaf-off, were harvested and collected for this study. Collections of trees include roadside trees, slope trees, estate trees, garden trees, and also forest trees from around Hong Kong (Fig. 1).

Tree sampling was conducted from March 2023 to April 2024, encompassing a total of 100 trees, which represented 31 tree species and 17 tree families. Three types of samples per tree were collected, including the trunks, branches, and leaves, as shown in Fig. 2. Slices of the trunk and branches were collected using a chainsaw on the felled tree. If available, we typically collected 3 slices of trunks (1 to 3 cm thick), 5 branch samples, and 1 bag of leaves. Generally, the trunk samples were retrieved in the wood part that was below the first tree fork.

According to the Hong Kong Biodiversity Information Hub, the dominant tree families in Hong Kong as of 2022 include *Euphorbiaceae, Sapotaceae, Moraceae, Sterculiaceae, Myrtaceae, Fagaceae, Lauraceae,* and *Theaceae*. Six of the listed families (75 %) are covered in the 17 families collected for this research. Table S1, S2 in the Supplementary Materials shows the species and families of the collected trees.

Fig. 2. Collection of tree samples.

2.3. Mobile laser scanning

LiDAR technology is considered an accurate solution for retrieving biophysical attributes of vegetation (Xu et al., 2021). Before the trees were felled, mobile laser scanning was performed by surrounding the tree in 360 degrees using a LiDAR backpack. As collecting the entire tree for AGB estimation was not available, we implemented mobile laser scanning to obtain the trees' 3D structure models. In this study, LiDAR surveys were conducted using the *GreenValley LiBackpack DGC 50 backpack laser scanning system*, which features Global Navigation Satellite System (GNSS) and Simultaneous Localization and Mapping (SLAM) technology, providing 1 cm + 1 ppm positioning accuracy and a point cloud with +/- 3 cm relative accuracy. Max ranging distance is up to 100 m with a scanning rate of 600,000 pts/s, enabling it to penetrate and capture tree top information (Figs. 3–7).

3. Methodology

This study proposes an innovative framework for developing an allometric model in Hong Kong. First, data collection was conducted using destructive sampling and backpack laser scanning of trees. Next, the aboveground biomass (AGB) of felled trees is measured by combining the methods of point-cloud processing and laboratory processing. The former involves calculating felled tree volume, while the latter includes oven-drying and measuring dry weight (biomass). Then, tree data augmentation and quantitative structure model (QSM) generation are conducted to enlarge the data size and extract detailed tree parameters, respectively. The measured tree AGB and extracted tree parameters were used to develop the allometric models. Finally, validation and analysis were performed to evaluate the performance of the

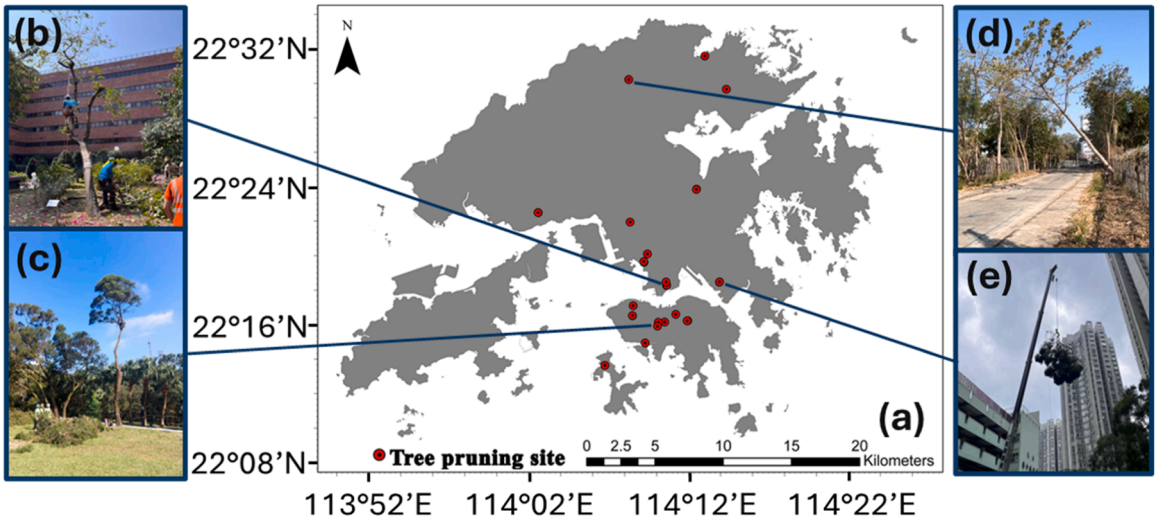


Fig. 1. (a) Map showing the tree felling locations around Hong Kong, (b, c, d, e) tree felling site photos.

Fig. 3. (a). LiDAR backpack; (b). LiDAR backpack survey; (c). Tree felling practice.

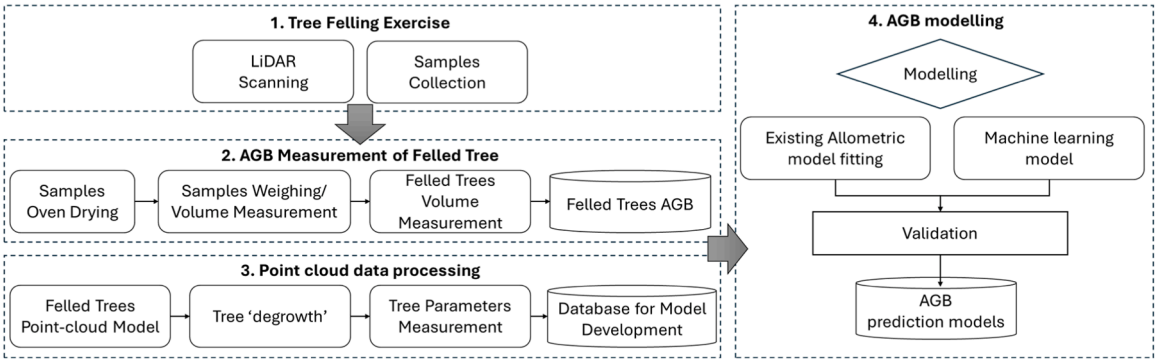


Fig. 4. Flowchart diagram of overall methodology.

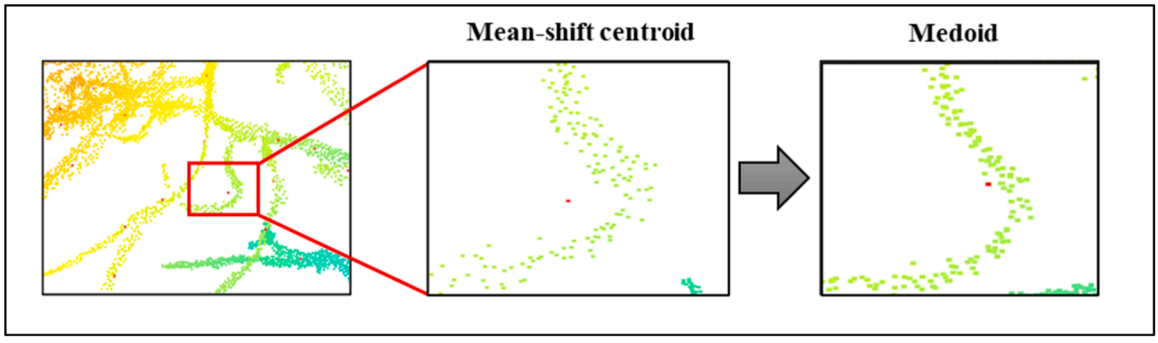


Fig. 5. Implementation of medoids.

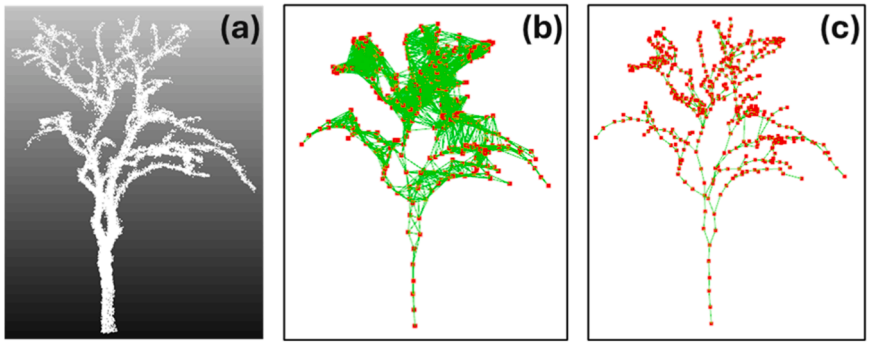


Fig. 6. (a) Raw point-cloud; (b) graph connection; (c) skeleton model.

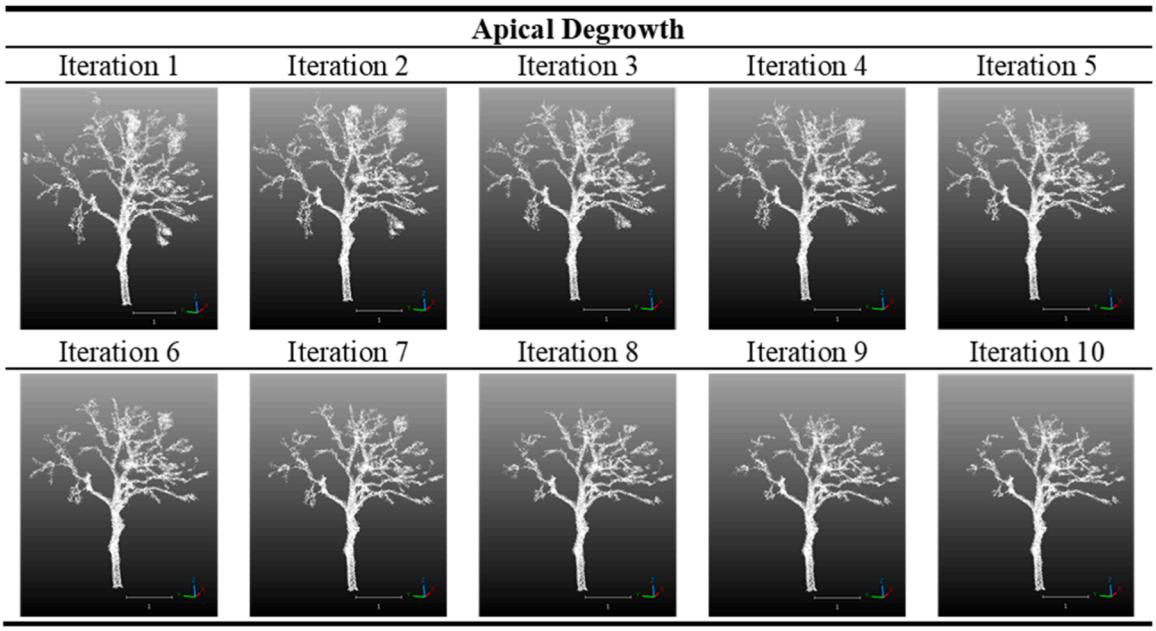


Fig. 7. Apical 'degrowth' with 10 iterations.

developed allometric model.

3.1. AGB measurement of felled trees

In this section, the AGB of each felled tree is measured. After the tree samples were collected, they were oven-dried at 50 degrees Celsius for two weeks to remove all moisture. The weight of the samples after oven drying refers to dry weight (DW), and the sample's DW is equal to the sample's biomass (Parresol, 1999). Apart from measuring the samples' biomass, the samples' volume is calculated by measuring the samples' thickness and diameters.

Next, the tree point-cloud model of the felled tree is input into the software TreeQSM, which calculates the entire tree volume. TreeQSM reconstructs the quantitative structure model (QSM) of the felled trees from the point-cloud data. A QSM is composed of cylinders to capture the topology, geometry, and volume of tree's wood structure (Raumonen et al., 2013). The TreeQSM is open-source software with constant updates in version, and the performance of TreeQSM in metrics extraction was explained in Raumonen et al. (2013). With TreeQSM, the entire tree volume, including trunk volume and branch volume can be measured.

Once the sample's volume, sample's biomass, and entire tree volume are calculated, the entire tree above ground biomass (AGB) can be measured. The entire tree AGB is calculated as the sum of trunk and branch biomass using Eq. (2), where density is derived from the ratio of sample biomass to sample volume (Eq. (1)).

$$trunk\ density = \frac{sample_{trunk}\ biomass}{sample_{trunk}\ volume};$$

$$branch\ density = \frac{sample_{branch}\ biomass}{sample_{branch}\ volume} \quad (1)$$

$$AGB = trunk\ volume \times trunk\ density + branch\ volume \times branch\ density \quad (2)$$

3.2. Tree data augmentation by 'degrowth'

In a local region with high tree species diversity, trees' volume can be different even though both trees have the same height and DBH due to local tree species richness and structural variability (Kunz et al., 2019). Addressing this issue is hindered by the fact that there is a limited amount of felled-tree data. Given that real-world data is insufficient, we propose a tree 'degrowth' method to generate a substantial training dataset by synthesizing tree data to create a robust model.

Tree growth can be mainly stratified into apical growth and radial growth (Wilson, 2000), and apical growth is focused on data augmentation in this study. We propose geometrically shrinking the point-cloud model of the felled tree apically, and virtual trees can be extracted at certain intervals during the shrinking process. Tree allometric model describes the relationship between tree AGB and dendrometric parameters. Existing allometric models are highly data-dependent and require substantial datasets for accurate predictions (Roxburgh et al., 2015). Augmented trees enable better generalization to unseen tree data and enhance the diversity of training dendrometric parameters. Regarding the apical degrowth method, we first assume apical growth of tree biomass occurs at the tips of branches and trunk, contributing to longitudinal growth.

3.2.1. Mean shift segmentation

The initial step involves transforming scattered point-clouds into meaningful 3D structures by generating tree nodes that emulate physical tree structures. The mean-shift clustering algorithm is applied and this iterative clustering algorithm does not require a pre-set number of clusters, and the level of clustering is defined by the bandwidth parameter σ (Daghigh et al., 2022). To initialize, assume the tree point-cloud = $\{x_1, x_2, \ldots, x_n\}$, a Gaussian function G was selected for the kernel as shown in (Eq. (3)) (Wang and Wong, 2023). To enhance the mean-shift clustering performance, a decaying bandwidth σ_h was used. For every tree point x_i, $m(x_i)$, which is the mean of all points weighted by Gaussian kernel within the bandwidth radius was calculated. Next, x_i was translated to computed $m(x_i)$, and this process iterates until convergence. To avoid over-clustering in the trunk region, the bandwidth should be set to a size similar to the DBH. Conversely, a smaller bandwidth should be set to preserve finer details in a small branch (tree-top) region. Hence, the bandwidth should vary along the height to establish a suitable searching range, and the bandwidth setting is shown in (Eq. (4)).

$$m(x_i) = \frac{\sum_{x_j \in N(x_i)} x_j G\left(\| \frac{x_j - x_i}{\sigma} \|^2 \right)}{\sum_{x_j \in N(x_i)} G\left(\| \frac{x_j - x_i}{\sigma} \|^2 \right)} - x \quad (3)$$

$$\sigma_h \begin{cases} \sigma_{max} \approx DBH \\ \sigma_{min} = 0.01 \ (default) \end{cases} \quad (4)$$

3.2.2. Wood leaf separation

Next, leaf points are removed to extract only the trunk and branch for precise AGB estimation. We adopted the graph-based wood-leaf separation (GBS) algorithm proposed by Tian and Li (2022). After performing mean-shift clustering in the previous step, we adopted the GBS method for initial and final wood points extraction through the source codes provided by (Tian and Li, 2022). Finally, the extracted wood point-cloud is input into the TreeQSM for parameter extraction.

3.2.3. Tree node generation

Mean-shift clustering is a centroid-based clustering algorithm, and this enables the estimation of local maxima where density is highest (Carreira-Perpinán, 2015). However, when dealing with large datasets, the computed mean-shift centroid can fall outside the classified cluster. This occurs when kernel bandwidth is incompatible with scenarios such as noisy or irregularly shaped clusters. Out-of-boundary centroids could be found in tree forks or turning regions, where non-convex-shaped clusters typically dominated. To address this, out-of-boundary centroids were first detected and identified; then, a medoid is implemented to replace these out-of-boundary centroids. In the mean-shift-generated cluster, the cluster's medoid is the point that has minimal average dissimilarity to other data points within the cluster (Kaur et al., 2014). This ensures that the generated tree node accurately reflects the tree morphology, as the medoid is always the actual point within the cluster's point set.

3.2.4. Tree skeleton construction

Following the tree node generation, nodes were interconnected by edges to construct tree edge graph. For every node, neighbouring node points are identified and connected by edge. A varying searching radius is applied across different node points (Eq. (5)). The searching radius was defined to be slightly higher than the bandwidth parameter in previous mean-shift clustering. This adjustment was made to ensure the formation of meaningful connections by linking consecutive nodes, while simultaneously preventing over-connections that could compromise the tree's morphological structure. σ is the bandwidth parameter and c is the searching buffer. The buffer c will adjust the complexity of the graph connection, while c was defaulting as 0.5 m for most tree data.

$$searching\ radius_{node\ i} = \sigma_h + c \quad (5)$$

After constructing the tree graph by connecting nodes with edges, the graph is filtered by pruning excess edges to form a tree skeleton model. The skeleton construction process was initiated by registering the root node (Eq. (6)), which was identified as the node point with the minimum z value. For every node point, the shortest path distance along the connected edges to the root node was computed. Next, delete the edges that the computed shortest paths had never passed. Most of the redundant edges were filtered in this step, generating a preliminary tree skeleton model.

$$node_{root} = node[\min(z)] \quad (6)$$

At this stage, certain edges may still be incorrectly connected, contradicting the original tree structure. Numerous studies have proposed sophisticated algorithms and solutions for constructing tree skeleton models (Li et al., 2022; Cárdenas et al., 2022; Livny et al., 2010). Yet it remains challenging for a single algorithm to handle diverse tree types. Rather than further refining the skeleton construction algorithm, it is proposed to reinput the original tree point-cloud for final tree skeleton refinement. The preliminary skeleton is overlaid on the original point cloud; this allows for a straightforward way to detect falsely connected edges. To remove such edges, a searching radius that is perpendicular to the edge is established for every edge, with a radius value that varies among edges (Eq. (8)). For $edge_i$, a group of points within the searching radius are defined as $point\text{-}set_i$ (Eq. (9)). Next, $edge_i$ is divided into n segments *(usually $n = 5$)* of equal length (Eq. (10)). Then project all the points in $pointset_i$ to $edge_i$, if the middle segment of $edge_i$ record points' projection, $edge_i$ is retained, otherwise it is deleted (Eq. (11)). The tree skeleton is then constructed for further 3D tree modelling.

$$shortest\ path\ distance\ d_i = d(node_i, node_{root}) \quad (7)$$

$$searching\ radius\ of\ edge\ e_i = r_i \quad (8)$$

$$pointset_i = \{p_i | distance(p_i, e_i) \le r_i\} \quad (9)$$

$$e_i = \{S_{i(1)}, S_{i(2)}, S_{i(3)}, S_{i(4)}, S_{i(5)}\} \quad (10)$$

$$retain\ e_i = \begin{cases} true, \ if\ any\ points\ from\ pointset_i\ project\ onto\ S_{i(3)} \\ false, \ otherwise \end{cases} \quad (11)$$

The following figure demonstrated the skeleton generation:

3.2.5. Tree augmentation

With the tree skeleton model, boundary nodes are defined as nodes that are connected to only one edge, except the base root node. Next, the degrowth is activated by deleting the first boundary nodes, resulting in a 1st updated skeleton model. Then, the detection and removal of boundary nodes on the newly updated skeleton model iterate. Define iteration number = n, and n updated skeleton models are generated. Augmented tree point-cloud models and QSMs can be generated with these updated skeleton models. The following figures demonstrate only the apical degrowth process, with 10 iterations. Note that in every iteration, all the boundary nodes will be removed.

Apart from the "degrowth" method, a random pruning of structures was carried out by randomly removing some branch structures. In a real-world scenario, after moving a segment of a branch, existing allometric models are unable to detect biomass loss since tree parameters, including height, DBH, and crown length, remain unchanged.

3.3. Allometric model development

The 100 felled trees were shuffled and divided into train and test sets, in an 8:2 ratio. The felled tree data were then augmented to increase data size. Each felled tree will generate 10 augmented trees; hence, the entire dataset contains 1100 trees. The training phase of all machine learning models utilized a 5-fold cross-validation to maintain robustness and reliability. The testing dataset was unseen data for performing the out-of-sample test to prevent overfitting.

For the parameters input in allometric model development, the TreeQSM-extracted parameters were utilized. Table 1 summarises the tree parameters that were used as predictors for trees' AGB estimation,

Table 1
Tree parameters.

Parameter Table		
Family Parameter		family
Advanced parameter	Basic parameter	DBH Tree Height (TH) Crown Length (CL) Crown Area (CA) Trunk Length (TL) Branch Length (BL) Trunk Area (TA) Branch Area (BA) Crown Base Height (CBH)

and they can be classified into family parameters, basic parameters and advanced parameters.

This study adopts two approaches to develop the AGB estimation model: (1) fitting the existing allometric models, and (2) developing machine learning based allometric models. The first approach aims to at assess the performance of the existing allometric models on local tree data, while the second approach aims to enhance the AGB prediction accuracy by leveraging machine learning algorithms.

3.3.1. Existing allometric model

Three published allometric models (Chave et al., 2005; Djomo et al., 2010) were selected as candidate models.

Model 1:

$$ln(AGB) = a + b\ ln\left(H * \rho * DBH^{\,2}\right) \tag{12}$$

Model 2:

$$ln(AGB) = a + b[ln(DBH)]^2 + c\ ln\,(H) + dln\,(\rho) \tag{13}$$

Model 3:

$$ln(AGB) = a + b[ln(DBH)]^2 + c\ ln\left(H * DBH^{\,2}\right) + d\ ln\,(\rho) \tag{14}$$

These selected models applied logarithmic transformation and relied on the same basic parameters listed in Table 1, including tree height (H), diameter at breast height (DBH), and wood density (ρ). To evaluate the performance of these models, all three models were fitted to the local dataset using the ordinary least squares regression method, and the goodness-of-fit and error will be assessed.

3.3.2. Machine learning model

In machine learning training, basic, advanced and family parameters were employed. For the tree family, the one-hot encoding technique was applied to these taxonomic parameters before machine learning training. Furthermore, to evaluate the impact of different parameters on prediction accuracy, three configurations with varying parameter combinations were designed for training: (1) basic parameters, (2) advanced parameters, and (3) advanced and family parameters. We selected various machine learning algorithms, and the best-performing algorithms were chosen as the final allometric model. Considering the model inference stage in AGB estimation, it is challenging to obtain wood density data, as the wood density value of the same species varies with wood age, tree height, growth rate, etc. (Chave et al., 2009). Given the main goal of this study is to use parameters extracted from LiDAR model to predict tree AGB, wood density is excluded from the AGB predictor variables.

Next, the selected machine learning algorithms are introduced and explained.

a) Random Forest (RF)

Random Forest operates by growing numerous trees to increase accuracy while reducing overfitting. Random Forest searches for the best feature in the random data groups (Breiman, 2001). In random forest training, bootstrap sampling is employed and each tree is trained on a random subset of data. RF is effective in capturing non-linear relationships in tabular data, while it features bootstrap mechanism and max_depth function to prevent overfitting. RF is a simple yet robust model that establishes a resilient baseline for the study.

b) Gradient Boosting Tree (GBT)

Unlike Random Forest, Gradient Boosting Tree (GBT) grows trees iteratively, and each new decision tree is trained to reduce the error of the previous tree through optimizing the loss function (Friedman, 2001). Apart from bagging models, including RF or decision trees, GBT is a baseline algorithm in boosting models, where trees are grown sequentially. GBT was employed to test the general performance of the boosting tree model in AGB prediction.

c) XGBoost (XGB)

XGBoost (Extreme Gradient Boosting) is an optimized machine learning model based on gradient boosting. Similar to GBT, XGB builds trees to correct the error of the previous tree. It also enhances the traditional GBT by adding Lasso and Ridge regularization to prevent overfitting (Chen and Guestrin, 2016), avoiding the error induced by significant outliers due to extremely large tree data.

d) LightGBM (LGBM)

LightGBM (light gradient boosting machine) is also an enhanced GBT machine learning algorithm. LightGBM features histogram-based learning, which involves data binning to increase training efficiency (Ke et al., 2017). LGBM supports categorical features with integer input, eliminating the need for one-hot encoding, which enables more efficient training when considering the family parameters in this study.

Finally, Table 2 summarizes all six models to be developed in this study.

3.4. Accuracy metrics

The model accuracy is evaluated by mean absolute error (MAE), mean absolute percentage error (MAPE) and R^2 for goodness-of-fit.

Mean absolute error MAE measures the average magnitude of errors in a prediction, while MAPE expresses the error relative to the true value.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \widehat{y}_i| \tag{15}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \widehat{y}_i}{y_i}\right| * 100 \tag{16}$$

Where, n = number of predictions, y_i = measured AGB for the i^{th} prediction, and $\widehat{y}_i$ = predicted AGB for the i^{th} observation

R^2 value (coefficient of determination), evaluates the goodness of fit of a regression model.

$$R^2 = 1 - \frac{SS_{residuals}}{SS_{total}} \tag{17}$$

Where, $SS_{residuals}$ = sum of squared residuals and SS_{total} = total sum of squares

4. Results and discussion

4.1. Tree parameters measurement by QSM

Table 3 summarizes the statistics of the extracted tree parameters of all 100 trees. Basic parameters including tree height and DBH, displayed a skewness value of around 0.5, indicating a slightly symmetrical distribution. On the contrary, the advanced parameters extracted by QSM showed a larger positive skewness, exhibiting an asymmetrical distribution. This suggests that on the same dataset, there are differences in basic and advanced parameters in terms of distribution. The advanced parameters extracted by QSM can inherit heterogeneous information across the mixed-species dataset.

Next, 50 felled trees were randomly selected for evaluation by assessing tree parameters extracted from QSM, including tree height, DBH, and crown length. The QSM-measured metrics were compared

Table 2
Summary of models.

Final Models	
Model 1	$ln(AGB) = a + b\ ln\left(H * \rho * DBH^{\,2}\right)$
Model 2	$ln(AGB) = a + b[ln(DBH)]^2 + c\ ln\,(H) + dln\,(\rho)$
Model 3	$ln(AGB) = a + b[ln(DBH)]^2 + c\ ln\left(H * DBH^{\,2}\right) + d\ ln\,(\rho)$
Model 4	ML_{best} (basic)
Model 5	ML_{best} (advanced)
Model 6	ML_{best} (advanced, family)

Table 3
Statistics of tree parameters.

Metrics	Tree parameters								
	Tree Height (m)	Trunk Length (m)	Branch Length (m)	Trunk Area (m^2)	Branch Area (m^2)	DBH (m)	Crown Area (m^2)	Crown Base Height (m)	Crown Length (m)
Min	3.05	0.29	0.00	0.02	0.00	0.02	0.05	0.51	0.00
Max	20.62	17.71	391.00	15.70	77.25	0.61	111.70	10.17	18.49
Mean	10.61	8.27	125.58	5.23	24.05	0.26	35.59	3.70	6.90
Median	9.99	7.92	101.54	4.82	21.04	0.25	29.40	3.34	6.54
Std	3.51	3.76	88.05	2.89	17.42	0.12	28.63	2.09	3.31
Q1	8.63	5.91	58.18	3.32	10.44	0.17	12.14	2.01	5.42
Q3	12.56	10.37	172.70	6.98	35.35	0.32	45.98	4.88	8.24
Skewness	0.56	0.67	1.03	0.82	1.04	0.49	1.06	0.88	1.13

against manual measurements from raw point-cloud tree model. Manual measurement of tree metrics was performed by human annotation using the point-picking tools in CloudCompare, where tree dimensions (tree height, DBH, and crown length) were measured directly on the point-cloud data. The scatter plots in Fig. 8 compare the QSM-extracted and manual-measured tree parameters. The highest correlation was achieved with DBH ($R^2 = 0.96$), followed by the crown length ($R^2 = 0.89$) and tree height ($R^2 = 0.89$). For measurement of height and crown length in raw point-cloud model, leaf and noises were taken into account, while QSM-extracted height and crown excluded the effect of leaf points. Overall QSM-measured results showed a strong alignment with the CloudCompare-manual measurement, providing promising and accurate results for tree metrics measurement.

4.2. AGB measurement of felled trees

The AGB measurement results are presented in Table S3 of the Supplementary Materials. The performance of AGB measurement is determined by the quality of destructive sampling processing and the accuracy of QSM. Since it is impossible to oven-dry the entire trees, we compared the measured AGB of felled trees (obtained in this study) to the estimated AGB of felled tree calculated by widely cited existing allometric model. We select the two published allometric models, identified as model A and model B, from Chave et al. (2014) to compare the measured AGB derived in this study. This aims to validate our measured trees' AGB by checking the alignment with the well-established model, ensuring that the results in this study were consistent with existing methodologies. Eqs. (18) and 19 demonstrate the selected model A and model B, and Fig. 9 compared the measured felled tree AGB and model A, B predicted felled tree AGB.

Model A:

$$AGB = 0.0673 \times \rho D^2 H^{0.976} \tag{18}$$

Model B:

$$AGB = \exp\left(-1.803 - 0.976E + 0.976\ln(\rho) + 2.673\ln(D) - 0.0299[\ln(D)]^2\right) \tag{19}$$

Model A and B predicted AGB values indicated a moderate correlation to the measured AGB, with $R^2 = 0.61$ and $R^2 = 0.57$ respectively. Despite moderate correlations, typically high MAE values of 434.9 (kg) and 549.43 (kg), and MAPE of 48.73 % and 67.79 % were observed in Models A and B. The moderate correlation validated the credibility of the measured AGB, while the high MAE value suggested that the existing allometric models were not sufficiently accurate for the local dataset.

4.3. Accuracy assessment of models

The following Fig. 10 shows the results of model 1 to model 6. All results of R^2, MAE, and MAPE were calculated in the out-of-sample testing. The left-most bar chart in Fig. 10 shows the results of model 1 – 3, which model 1 refers to Chave_1 (equation 12); model 2 refers to Chave_2 (equation 13); model 3 refers to Djomo (Eq. (14)). Model 4 – 6 were machine learning training, and the results of all selected machine learning algorithms were presented in the bar charts.

Table 4 summarizes the final results of the six models and presents the best-performing machine learning algorithms for models 4 – 6. Model 5 – 6, which utilized machine learning algorithms and advanced tree parameters, had an overall better accuracy compared to model 1 – 3, which relied on existing allometric equations. Model 2 scored the best-performing model among the three existing allometric models, with $R^2 = 0.52$; MAPE = 53.97 %, while the remaining model 1 and 3 scored $R^2 = 0.52$; MAPE = 60.83 %, and $R^2 = 0.53$; MAPE = 56.40 %, respectively. Best-performing machine learning algorithms for model 4, model 5 and model 6 are Random Forest ($R^2 = 0.41$; MAPE = 79.20 %), LightGBM ($R^2 = 0.79$; MAPE = 42.59 %) and XGBoost ($R^2 = 0.82$; MAPE = 40.70 %), respectively. The best-performing machine learning algorithms were selected based on R^2 value and MAPE. MAE calculate the

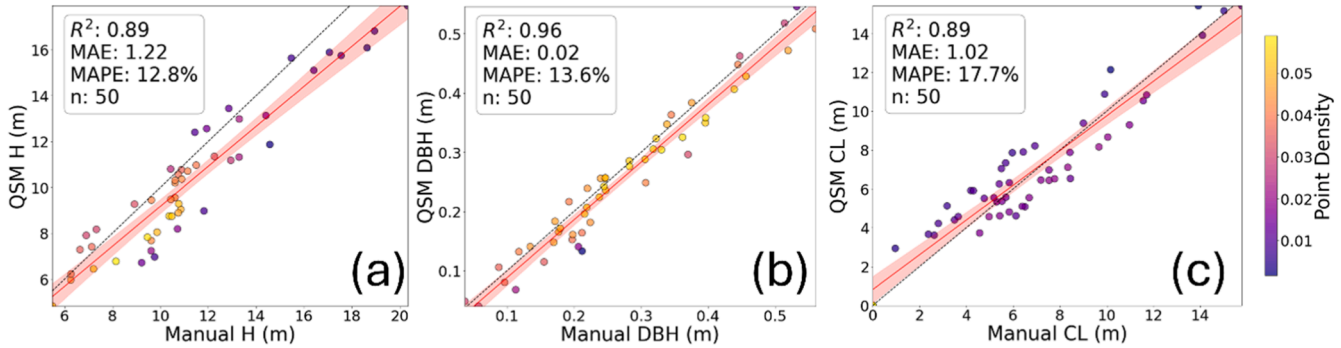


Fig. 8. QSM vs manual measurement. (a) Tree Height; (b) DBH; (c) Crown Length.

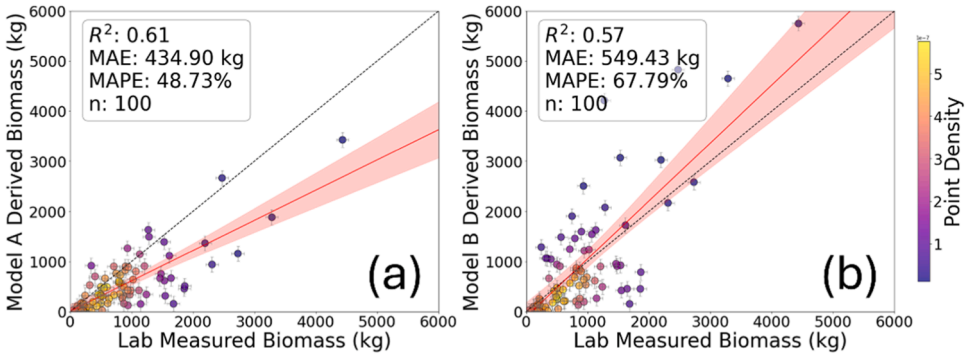


Fig. 9. Measured AGB vs model predicted AGB. (a) Model A, (b) Model B.

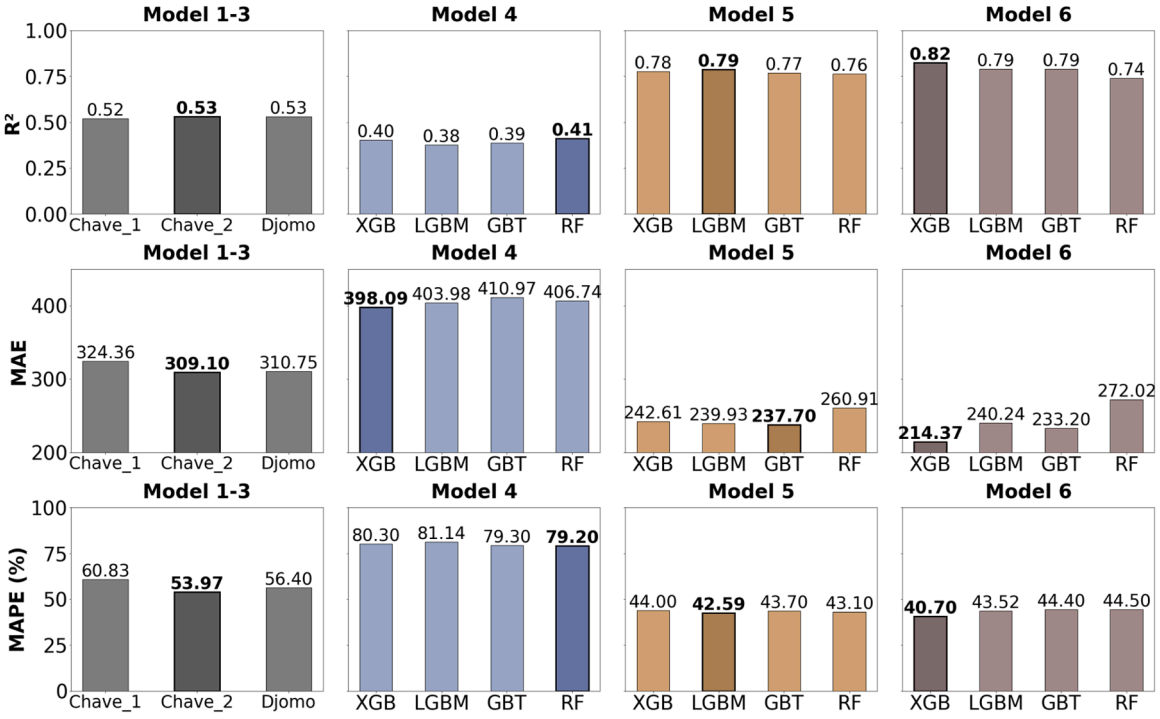


Fig. 10. Results comparison of model 1 – 6 using R^2, MAE, MAPE.

Table 4
Allometric models result.

Allometric Models	Best ML algorithms	R^2	MAE (kg)	MAPE (%)
Model 1	–	0.52	324.36	60.83
Model 2	–	0.52	309.10	53.97
Model 3	–	0.53	310.75	56.40
Model 4	RF	0.41	398.09	79.20
Model 5	LGBM	0.79	237.70	42.59
Model 6	XGB	0.82	214.37	40.70

absolute differences between measured and predicted values, while MAPE calculate the absolute percentage differences. In selecting the best-performing models, MAPE weighs more than MAE since MAE disproportionately penalizes errors given the wide ranges of AGB values. MAPE normalizes error relative to the measured value, making this more suitable for interpreting models' performance on AGB prediction.

Fig. 11 compares the model-predicted AGB and measured AGB, Fig. 11 (a – c) are scatter plots of existing allometric models (model 1 – 3) and Fig. 11 (d – f) are scatter plots of the best-performing machine learning models (model 4 – 6). By assessing the scatter plots, an underestimation of AGB is observed in high AGB trees, and a slight overestimation is exhibited in low AGB trees. Most data points are concentrated in the AGB range of 0 – 1000 kg, indicating that high AGB tree data is insufficient for the development of AGB prediction model.

4.3.1. Parameterization of existing allometric model (Model 1 – 3)

The following displays the results of the existing allometric models (models 1–3), parameterized using a local dataset.

Model 1:

$$ln(AGB) = 2.364 + 0.471\ ln\left(H * \rho * DBH^{\,2}\right) \tag{20}$$

Model 2:

$$ln(AGB) = 2.398 + 0.912[ln(DBH)]^2 + 0.456\ ln\,(H) + 1.06ln\,(\rho) \tag{21}$$

Model 3:

$$ln(AGB) = 2.574 + 0.894[ln(DBH)]^2 + 0.419\ ln\left(H * DBH^{\,2}\right) + 0.182\ ln\,(\rho) \tag{22}$$

Models 1, 2, and 3 displayed similar R^2 and MAPE values, as shown in Table 4. A R^2 value of 0.52–0.53 represents a relatively weak

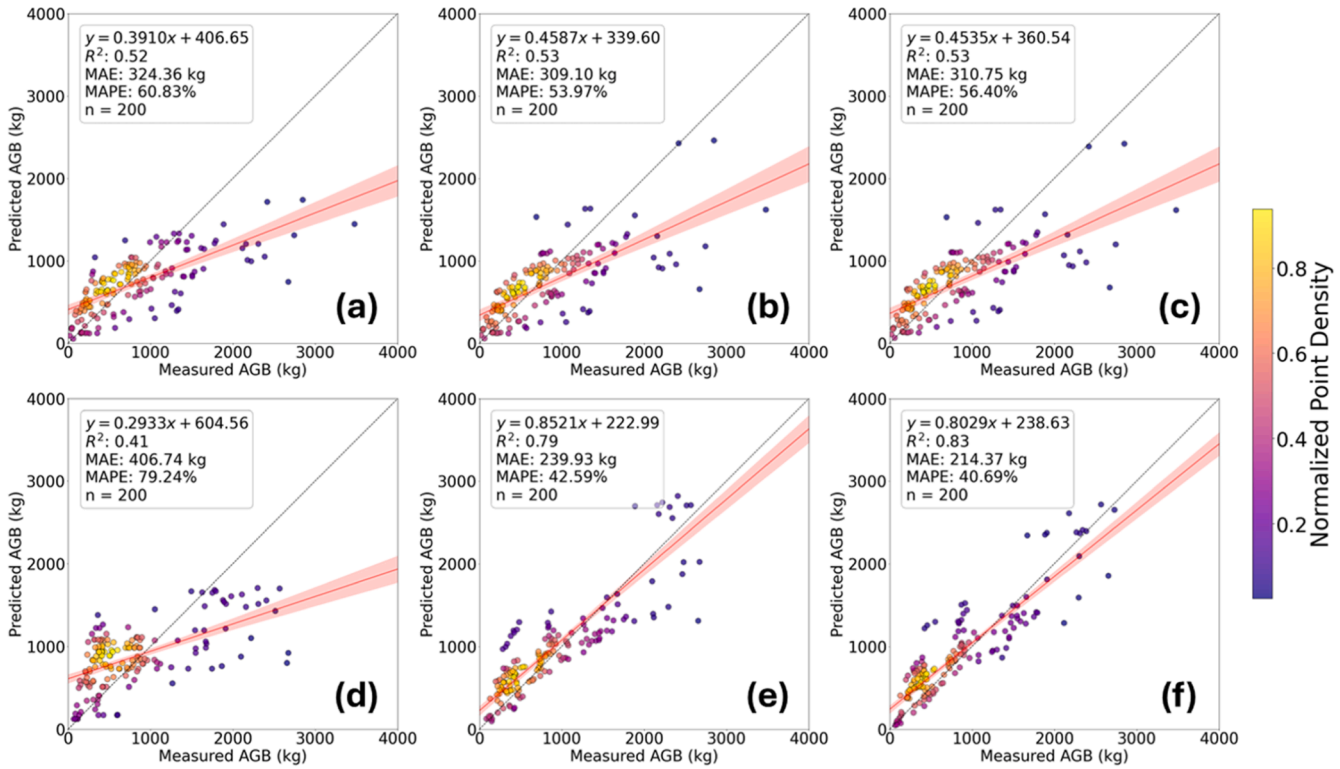


Fig. 11. Scatter plots comparing model-predicted AGB and measured AGB, (a) model 1; (b) model 2; (c) model 3; (d) model 4 [Random Forest]; (e) model 5 [LightGBM]; (f) model 6 [XGBoost]. The black line, red line, and red shaded area represent the 1:1 reference line, regression line, and 95 % confidence interval for the regression line, respectively.

correlation, indicating that the existing allometric model predicts approximately 50 % of the allometric relationship. Our results indicate that applying existing global allometric models to a specific local dataset led to poor performance. Developed using large and global datasets, these existing allometric models perform well across universal datasets (Chave et al., 2005). However, the specific local variability exhibited in our local dataset cannot be well captured by the existing models, as trees in localized regions exhibit complex variabilities in AGB, wood density, and health conditions (Temesgen et al., 2015). Our finding supports the need for site-specific allometric models (Yuen et al., 2016; Yang et al., 2022) to enhance the accuracy of local AGB estimation.

Overall, the accuracy results suggested that the existing allometric models failed to predict AGB accurately, and they are insufficient to capture the complexity of tree structure in the local tree data. To summarize, the existing allometric models are limited by their non-localized application and the simplicity that results in poor performance on local datasets. In order to capture non-linear relationships between tree physical parameters and AGB in Hong Kong, enhanced methods of machine learning and advanced parameters implementation are required to handle the complexity of local trees.

4.3.2. *Machine learning based allometric model (Model 4 – 6)*

Model 4 – 6 employed machine learning algorithms to predict AGB using advanced parameters. Model 4 used only basic parameters (DBH and tree height), in order to mimic the existing allometric model. Eventually, model 4 performs worse than model 1 – 3 (utilized simple linear regression), in terms of R^2 (0.41) and MAPE (79.20 %).

When only basic parameters are available, existing allometric models remain a more robust choice for AGB prediction. Basic parameters, including tree height and DBH, exhibit a strong linear relationship with AGB (Liu et al., 2018). Hence simple linear regression is more suitable for AGB prediction when using basic parameters, as it generalizes better in small and limited features datasets, while machine learning models are more prone to overfitting as the depth of trees increase (Van der Putten and Van Someren, 2004; Roelofs et al., 2019).

Given the poor accuracy in existing allometric models, more advanced tree parameters are required to support the training of machine learning algorithms. Comparing model 4 – 6, the performance of model 4 is limited by the insufficient input predictors. The superior performance of models 5–6 over model 4 suggested that machine learning models require richer predictors to fully leverage the non-linear modelling capabilities for AGB prediction. This finding also reflects the contribution of advanced parameters to enhance AGB prediction accuracy and highlights the limitations of basic parameters, which existing allometric models commonly rely on. By comparing model 5 – 6 and model 1 – 3, a huge enhancement in accuracy of model 5 and 6 is observed in terms of R^2 and MAPE value, with R^2 value reaching over 0.8 and MAPE reduced from range of 53 – 79 % in existing allometric models to 40 – 43 % in model 5 and 6. Model 5 – 6 are machine learning algorithms trained with advanced parameters listed in Table 1. This suggests that AGB variations in local trees can be well-captured by integrating advanced parameters and machine learning algorithms.

By comparing model 5 and 6, model 6 demonstrated a superior accuracy of $R^2 = 0.82$ and MAPE = 40.70 % compared to model 5's $R^2 = 0.79$ and MAPE = 45.59 %. This indicated that the inclusion of "tree family" parameters enhanced the modelling accuracy. Tree family parameters could conceal trees' taxonomic characteristics that exhibit family-specific growth dynamics, wood densities, or intrinsic structure that contribute to the accuracy of AGB estimation (Mensah et al., 2016). To summarize, the improvement on model 6 accuracy supports a data-driven finding that tree family parameter embodies meaningful value in enhancing AGB prediction accuracy. As most existing allometric models are stratified as either mix-species or species-specific models, the superior performance in model 6 provide a strong foundation for developing allometric model that enables taxa-inputs.

4.4. Application of advanced parameters

Correlation analysis was conducted on advanced parameters, and the results are presented in Fig. 12. The correlation was computed using the Pearson correlation coefficient, where 1 represents a perfect positive correlation, 0 indicates no relationship, and −1 refers to a perfect negative correlation (Sedgwick, 2012). By assessing the bottom row in the heatmap, all parameters displayed a positive correlation to AGB, except crown base height (CBH). This is reasonable as the increase of CBH is followed by the decrease of crown volume, which results in the decrease of AGB. For the basic parameter, tree height (TH), scored a relatively low correlation (0.19) among all parameters. Among all parameters, branch area (BA) calculated by QSM scored the highest correlation to AGB, followed by DBH, trunk area (TA), branch length (BA) and crown area (CA). The calculation of these important parameters, except DBH are enabled by QSM reconstruction, and this underscores the significance of QSM application in allometric model development.

The importance of each advanced parameter was assessed by permutation importance and the result is displayed in Fig. 13. Across all models, DBH, or a feature comprised of DBH scored the highest importance, this indicated that DBH is the most important parameter in AGB estimation. The aforementioned advanced parameters including trunk area (TA), branch area (BA), and branch length (BL) contribute a vast proportion of importance among all parameters in models 5 and 6.

Surprisingly, tree height (TH) achieved negative importance in models 2, 3, and 5, indicating that the tree height parameter could harm the model's performance. A possible reason for this could be the multicollinearity between tree height and other parameters. According to Fig. 12, tree height is highly correlated to trunk length (TL) and crown length (CL), this might reduce the importance of tree height as tree height information is hidden in trunk length and crown length. That said, the importance of trunk length and crown length was not significantly high according to Fig. 13, and the negative importance of tree height was also achieved in model 2. The tree condition in the local study area is also one of the factors that reduces the importance of tree height. In Hong Kong, Arecaceae and Moraceae are two common tree families, the former can grow very tall with no branches (Edelman and Richards, 2019) and the latter can grow dense branches (Primack et al., 1985). Moreover, other research discovered that the inclusion of tree height in the allometric model led to an overestimation of AGB, while the exclusion of tree height led to an underestimation (Goodman et al., 2014). Overall, these findings challenge the parameter tree height as a fundamental predictor in existing allometric models. When more advanced parameters are available, tree height is not the most reliable or impactful parameter for predicting AGB. Furthermore, the tree family

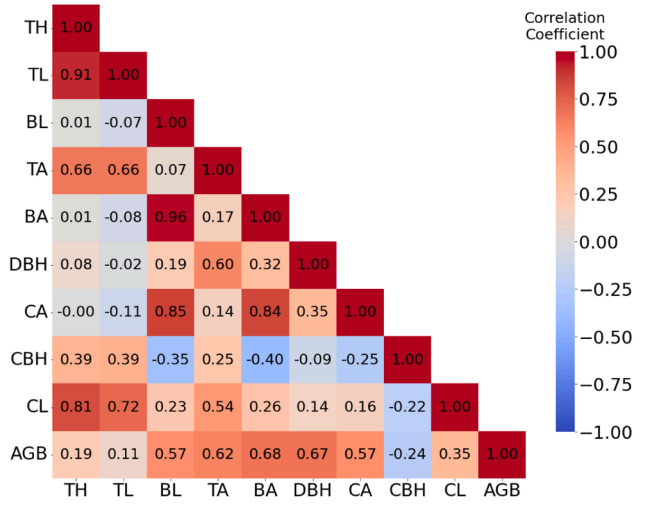


Fig. 12. Correlation heatmap of tree parameters and AGB.

contributed a relatively small amount of importance compared to other predictors. However, according to Fig. 10, the XGBoost model from model 5 to model 6 recorded an increase in R^2 value from 0.78 to 0.82. This data-driven result supports the importance of family parameters. The reason for the low importance of family is that the family-inherited ecological traits of trees could be partially explained by other features such as DBH.

Overall, while DBH still remains the most important AGB predictor, the importance and correlation analysis revealed that the most advanced parameters displayed relatively comparable importance, indicating that both basic and advanced parameters exhibit complementary roles in developing accurate allometric models for local region. To conclude, this highlights one of the study's key contributions: integrating LiDAR and QSM-measured advance parameters, which enables the capture of detailed structural variability of trees and hence enhances AGB estimation accuracy.

4.5. Significance of tree data augmentation

To assess the effectiveness of the tree data augmentation method, we compare the best-performing machine learning based allometric model (model 6) and the existing allometric model (model 2). Model 6 is trained with the augmented tree data while model 2 is the existing allometric model developed by Chave et al. (2005). Both model-predicted AGB values are plotted against QSM-generated volume and DBH, while the results are displayed in the 3D scatter plot in Fig. 14. In the 3D scatter plot, distinct colours were assigned to the tree data to represent each tree family. A 3D linear regression surface was fitted onto the data point for interpretation and comparison. In Fig. 14b, for trees with same DBH, the predicted AGB remains unchanged as QSM-derived volume increase. Given the volume and weight of the tree, a positive correlation is expected. However, it is found that the existing allometric model approach fails to handle complex tree shapes. Assume a scenario that multiple tree species with the same DBH, allometric model that trained by conventional sampling technique will fail to model AGB accurately. The proposed tree data augmentation method in this study enhanced the limited dataset of felled trees. In Fig. 14a, for trees with same DBH, QSM-derived volume and predicted AGB displayed a positive correlation observed in the 3D regression plane. In the augmentation stage, synthetic trees were generated by trimming branches, thus simulating as many possible tree shapes of a single tree at the age of felling day. Consider a scenario in which a tree branch was shortened or trimmed off due to general tree maintenance, existing allometric models were unable to detect such AGB reduction as the basic dendrometric parameters of a tree before and after trimming remained unchanged. Existing allometric models were developed by collecting felled trees, and directly relied on measured DBH, tree height and AGB (Chave et al., 2005, 2014). The direct assumption of a linear relationship between basic parameters and AGB limits the performance of existing allometric models in local datasets, as trees with similar DBH exhibit different AGB due to various factors, including tree morphology or species variation. The comparison in Fig. 14 underscores the significance of data augmentation for AGB prediction. Without being developed by the augmented data, the existing allometric model is unable to detect subtle AGB changes resulting from minor tree structural changes.

The tree data augmentation method is a key innovation of this study. By systematically trimming branches from a tree's point cloud models, synthetic trees were generated, simulating a wide range of possible tree shapes for the same DBH, effectively enhancing the structural diversity of the training data. The augmentation method addresses the existing allometric model's limitation, which is the inability to capture AGB change outside the dimension of DBH and height. The augmented tree data resolves this limitation by systematically simulating branch pruning scenarios, enabling the model to predict minor AGB changes. Moreover, the tree augmentation enhanced the structural diversity of the tree data, which mitigates the challenge of limited tree data. By

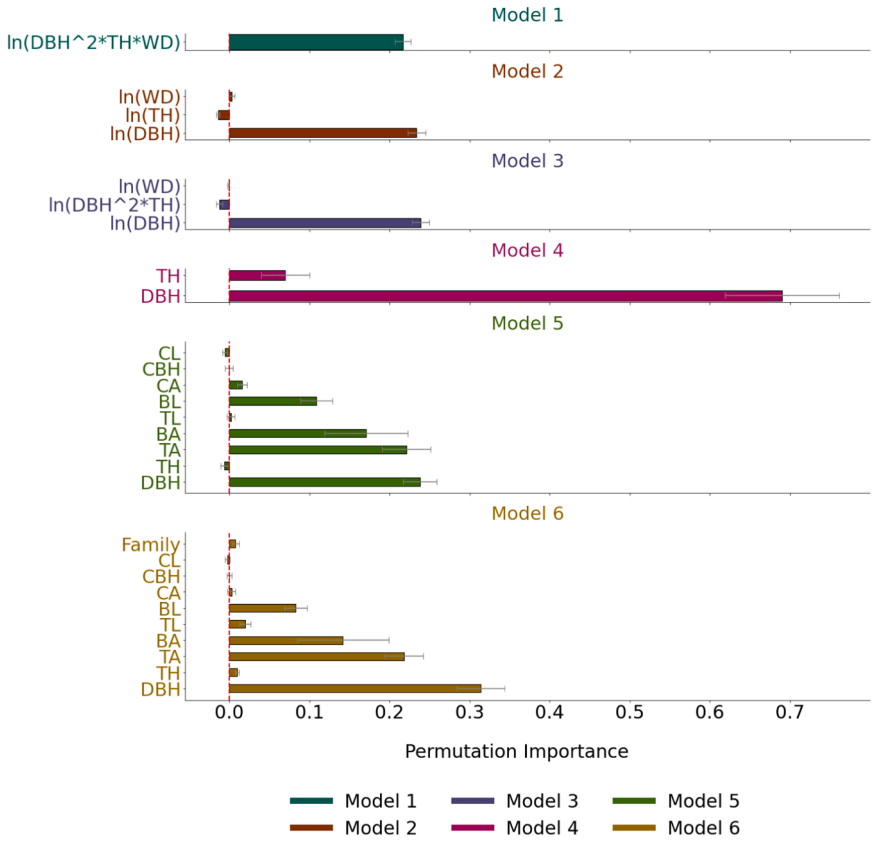


Fig. 13. Feature importance of AGB predictors.

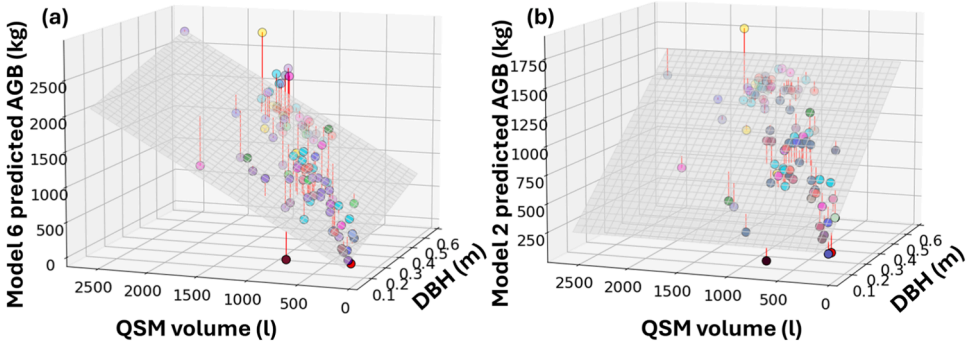


Fig. 14. (a) 3D scatter plot derived by model 6 predicted AGB; (b) 3D scatter plot derived by model 2 predicted AGB.

introducing enhanced structural variation to the tree data, the tree data augmentation method enables the trained model to reflect better the complexity of the local real-world trees' AGB.

5. Conclusion

The study presented an innovative workflow for estimating tree AGB by leveraging LiDAR technology with machine learning algorithms. Given the scarcity of available data for AGB modelling, the study proposed a tree data augmentation method to enhance the model's generalization and accuracy. Then, tree reconstruction by TreeQSM was carried out to retrieve advanced tree dendrometric parameters. In model development, it was found that the existing allometric equations did not fit the local dataset well. The best AGB estimation model was found to be the XGBoost, with the input of the advanced QSM-derived parameters and tree family parameters, scoring $R^2 = 0.82$ and MAPE = 40.70 %. Compared to the previous allometric model developed by Sarker (2010) in Hong Kong, this study encompassed 100 trees (17 species), whereas Sarker (2010) collected 75 trees (14 species). Instead of relying on manual tree measurement, this study further exploits LiDAR technology by applying TreeQSM and proposing a tree data augmentation method to address the limitations of traditional destructive sampling techniques. For the development of the allometric model, Sarker (2010) and other related research relied on simple linear regression. This study proposes a local AGB prediction model using machine learning algorithms that utilize augmented data and detailed QSM parameters. Finally, the proposed model incorporates the tree taxa category, which allows broader ecological applicability with higher generalization. Overall, the study combines city-scale tree data collection with LiDAR and machine learning algorithms, surpassing the capabilities of the existing allometric models to achieve accurate AGB estimation in the local region.

CRediT authorship contribution statement

Yan To Choi: Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Conceptualization. **Majid Nazeer:** Writing – review & editing, Visualization, Methodology, Funding acquisition, Formal analysis. **Man Sing Wong:** Writing – review & editing, Supervision, Project administration, Methodology, Funding acquisition, Formal analysis, Conceptualization. **Janet Elizabeth Nichol:** Writing – review & editing, Investigation. **Shao-Yuan Leu:** Writing – review & editing, Resources. **Jin Wu:** Writing – review & editing, Resources. **Amos P.K. Tai:** Writing – review & editing, Resources.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This project is substantially funded by the General Research Fund (Grant No 15603923 and 15609421), and the Collaborative Research Fund (Grant No C5062–21GF) and Young Collaborative Research Fund (Grant No C6003–22Y) from the Research Grants Council, Hong Kong, China. The authors acknowledge the funding support (Grant No BBG2 and CD81) from the Research Institute for Sustainable Urban Development, Research Institute for Land and Space, The Hong Kong Polytechnic University, Kowloon, Hong Kong, China. Majid Nazeer was substantially supported through the General Research Fund from the Research Grants Council of the Hong Kong China (Project No PolyU-15306224).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.tfp.2025.100955.

Data availability

The authors do not have permission to share data.

References

Awad, M., Khanna, R., Awad, M., Khanna, R., 2015. Support vector regression. Effic. Learn. Mach.: Theor. Concepts Appl. Eng. Syst. Des. 67–80.
Breiman, L., 2001. Random forests. Machine learning 45 (1), 5–32.
Calders, K., Newnham, G., Burt, A., Murphy, S., Raumonen, P., Herold, M., Kaasalainen, M., 2015. Nondestructive estimates of above-ground biomass using terrestrial laser scanning. Methods Ecol. Evol. 6 (2), 198–208.
Carbon Neutrality and Sustainable Development, 2021. *Hong Kong's climate action plan 2050: Carbon neutrality and sustainable development*. In: Government of the Hong Kong Special Administrative Region. https://www.eeb.gov.hk/sites/default/files/pdf/cap_2050_en.pdf.
Cárdenas, J.L., Ogayar, C.J., Feito, F.R., Jurado, J.M., 2022. Modeling of the 3d tree skeleton using real-world data: a survey. IEEe Trans. Vis. Comput. Graph. 29 (12), 4920–4935.
Carreira-Perpinán, M.A. (2015). A review of mean-shift algorithms for clustering. arXiv preprint arXiv:1503.00687.
Chave, J., Andalo, C., Brown, S., Cairns, M.A., Chambers, J.Q., Eamus, D., Yamakura, T., 2005. Tree allometry and improved estimation of carbon stocks and balance in tropical forests. Oecologia 145, 87–99.
Chave, J., Coomes, D., Jansen, S., Lewis, S.L., Swenson, N.G., Zanne, A.E., 2009. Towards a worldwide wood economics spectrum. Ecol. Lett. 12 (4), 351–366.
Chave, J., Davies, S.J., Phillips, O.L., Lewis, S.L., Sist, P., Schepaschenko, D., Saatchi, S., 2019. Ground data are essential for biomass remote sensing missions. Surv. Geophys. 40, 863–880.
Chave, J., Réjou-Méchain, M., Búrquez, A., Chidumayo, E., Colgan, M.S., Delitti, W.B., Vieilledent, G., 2014. Improved allometric models to estimate the aboveground biomass of tropical trees. Glob. Chang. Biol. 20 (10), 3177–3190.
Chen, J.M., 2021. Carbon neutrality: toward a sustainable future. Innovation 2 (3).
Chen, T., Guestrin, C., 2016. Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp. 785–794.
Chen, Y., Miao, D., Zhang, H., 2010. Neighborhood outlier detection. Expert. Syst. Appl. 37 (12), 8745–8749.
Cheung, C.S.C., Hart, M.A., 2014. Climate change and thermal comfort in Hong Kong. Int. J. Biometeorol. 58, 137–148.
Daghigh, H., Tannant, D.D., Daghigh, V., Lichti, D.D., Lindenbergh, R., 2022. A critical review of discontinuity plane extraction from 3D point cloud data of rock mass surfaces. Comput. Geosci. 169, 105241.
Djomo, A.N., Ibrahima, A., Saborowski, J., Gravenhorst, G., 2010. Allometric equations for biomass estimations in Cameroon and pan moist tropical equations including biomass data from Africa. For. Ecol. Manage. 260 (10), 1873–1885.
Edelman, S.M., Richards, J.H., 2019. Review of vegetative branching in the palms (Arecaceae). Bot. Rev. 85, 40–77.
Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat. 1189–1232.
Goodman, R.C., Phillips, O.L., Baker, T.R., 2014. The importance of crown dimensions to improve tropical tree biomass estimates. Ecol. Appl. 24 (4), 680–698.
Grinsztajn, L., Oyallon, E., Varoquaux, G., 2022. Why do tree-based models still outperform deep learning on typical tabular data? Adv. Neural Inf. Process. Syst. 35, 507–520.
Han, H., Jiang, X., 2014. Overcome support vector machine diagnosis overfitting. Cancer Inf. 13. CIN-S13875.
Kaul, M., Mohren, G.M.J., Dadhwal, V.K., 2010. Carbon storage and sequestration potential of selected tree species in India. Mitig. Adapt. Strateg. Glob. Change 15, 489–510.
Kaur, N.K., Kaur, U., Singh, D., 2014. K-medoid clustering algorithm-a review. Int. J. Comput. Appl. Technol. 1 (1), 42–45.
Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Liu, T.Y., 2017. Lightgbm: a highly efficient gradient boosting decision tree. Adv. Neural Inf. Process. Syst. 30.
Kohli, S., Godwin, G.T., Urolagin, S., 2020. Sales prediction using linear and KNN regression. In: Advances in Machine Learning and Computational Intelligence: Proceedings of ICMLCI 2019. Springer Singapore, Singapore, pp. 321–329.
Kunz, M., Fichtner, A., Härdtle, W., Raumonen, P., Bruelheide, H., von Oheimb, G., 2019. Neighbour species richness and local structural variability modulate aboveground allocation patterns and crown morphology of individual trees. Ecol. Lett. 22 (12), 2130–2140.
Lam, C.K.C., He, Q., Cheng, K.L., Fan, P.Y., Chun, K.P., Choi, B., Yetemen, O., 2022. Impact of climate change and socioeconomic factors on domestic energy consumption: the case of Hong Kong and Singapore. Energy Rep. 8, 12886–12904.
Lands Department, 2024. The Government of the Hong Kong Special Administrative Region. Hong Kong Geographic Data. Lands Department. January 10. https://www.landsd.gov.hk/en/resources/mapping-information/hk-geographic-data.html.
Lau, A., Bentley, L.P., Martius, C., Shenkin, A., Bartholomeus, H., Raumonen, P., Herold, M., 2018. Quantifying branch architecture of tropical trees using terrestrial LiDAR and 3D modelling. Trees 32, 1219–1231.
Li, J., Wu, H., Xiao, Z., Lu, H., 2022. 3D modeling of laser-scanned trees based on skeleton refined extraction. Int. J. Appl. Earth Obs. Geoinf. 112, 102943.
Liu, G., Wang, J., Dong, P., Chen, Y., Liu, Z., 2018. Estimating individual tree height and diameter at breast height (DBH) from terrestrial laser scanning (TLS) data at plot level. Forests. 9 (7), 398.
Livny, Y., Yan, F., Olson, M., Chen, B., Zhang, H., & El-Sana, J. (2010). Automatic reconstruction of tree skeletal structures from point clouds. In ACM SIGGRAPH Asia 2010 papers (pp. 1–8).
Magalhães, T.M., Cossa, V.N., Guedes, B.S., Fanheiro, A.S.M., 2021. Species-specific biomass allometric models and expansion factors for indigenous and planted forests of the Mozambique highlands. J. For. Res. 32 (3), 1047–1065.
Mensah, S., Veldtman, R., Du Toit, B., Glèlè Kakaï, R., Seifert, T., 2016. Aboveground biomass and carbon in a South African mistbelt forest and the relationships with tree species diversity and forest structures. Forests. 7 (4), 79.
Mugasha, W.A., Mwakalukwa, E.E., Luoga, E., Malimbwi, R.E., Zahabu, E., Silayo, D.S., Kashindye, A., 2016. Allometric models for estimating tree volume and aboveground biomass in lowland forests of Tanzania. Int. J. For. Res. 2016 (1), 8076271.
Mulatu, A., Negash, M., Asrat, Z., 2024. Species-specific allometric models for reducing uncertainty in estimating above ground biomass at Moist Evergreen Afromontane Forest of Ethiopia. Sci. Rep. 14 (1), 1147.
Parresol, B.R., 1999. Assessing tree and stand biomass: a review with examples and critical comparisons. For. sci. 45 (4), 573–593.
Primack, R.B., Ashton, P.S., Chai, P., Lee, H.S., 1985. Growth rates and population structure of Moraceae trees in Sarawak, East Malaysia. Ecology. 66 (2), 577–588.
Raumonen, P., Kaasalainen, M., Åkerblom, M., Kaasalainen, S., Kaartinen, H., Vastaranta, M., Lewis, P., 2013. Fast automatic precision tree models from terrestrial laser scanner data. Remote Sens. (Basel) 5 (2), 491–520.
Roelofs, R., Shankar, V., Recht, B., Fridovich-Keil, S., Hardt, M., Miller, J., Schmidt, L., 2019. A meta-analysis of overfitting in machine learning. Adv. Neural Inf. Process. Syst. 32.
Roxburgh, S.H., Paul, K.I., Clifford, D., England, J.R., Raison, R.J., 2015. Guidelines for constructing allometric models for the prediction of woody biomass: how many individuals to harvest? Ecosphere 6 (3), 1–27.
Roy, A.D., Debbarma, S., 2024. Comparing the allometric model to machine learning algorithms for aboveground biomass estimation in tropical forests. Ecol. Front. 44 (5), 1069–1078.
Sarker, M.L.R., 2010. Estimation of Forest Biomass Using Remote Sensing. The Hong Kong Polytechnic University. Doctoral dissertationPolyU Electronic Theses. https://theses.lib.polyu.edu.hk/handle/200/6079.

Sarker, S.K., Das, N., Chowdhury, M.Q., Haque, M.M., 2013. Developing allometric equations for estimating leaf area and leaf biomass of Artocarpus chaplasha in Raghunandan Hill Reserve, Bangladesh. South. For.: J. For. Sci. 75 (1), 51–57.
Sedgwick, P., 2012. Pearson's correlation coefficient. BMJ 345.
Segura, M., Kanninen, M., 2005. Allometric models for tree volume and total aboveground biomass in a tropical humid forest in Costa Rica 1. Biotropica: J. Biol. Conserv. 37 (1), 2–8.
Shi, Y., Choi, S., Ni, X., Ganguly, S., Zhang, G., Duong, H.V., Myneni, R.B., 2013. Allometric scaling and resource limitations model of tree heights: part 1. Model optimization and testing over continental USA. Remote Sens. (Basel) 5 (1), 284–306.
Shwartz-Ziv, R., Armon, A., 2022. Tabular data: deep learning is not all you need. Inf. Fusion 81, 84–90.
Sileshi, G.W., 2014. A critical review of forest biomass estimation models, common mistakes and corrective measures. For, Ecol, Manage 329, 237–254.
Temesgen, H., Affleck, D., Poudel, K., Gray, A., Sessions, J., 2015. A review of the challenges and opportunities in estimating above ground forest biomass using tree-level models. Scand. J. For. Res. 30 (4), 326–335.
Tian, Z., Li, S., 2022. Graph-based leaf–Wood separation method for individual trees using terrestrial lidar point clouds. IEEE Trans. Geosci. Remote Sens. 60, 1–11.
Toochi, E.C., 2018. Carbon sequestration: how much can forestry sequester CO2. For. Res. Eng.: Int. J. 2 (3), 148–150.
Treboux, J., Genoud, D., Ingold, R., 2018. Decision tree ensemble vs. nn deep learning: efficiency comparison for a small image dataset. In: 2018 International Workshop on Big Data and Information Security (IWBIS). IEEE, pp. 25–30.
Van Der Putten, P., Van Someren, M., 2004. A bias-variance analysis of a real world learning problem: the CoIL challenge 2000. Mach. Learn. 57, 177–195.
Van Wolputte, A. (2024). Mobile LiDAR technology for forest biomass inventories and carbon certification monitoring.
Vieira, S.A., Alves, L.F., Aidar, M., Araújo, L.S., Baker, T., Batista, J.L.F., Trumbore, S.E., 2008. Estimation of biomass and carbon stocks: the case of the Atlantic Forest. Biota Neotrop. 8, 21–29.
Wang, M., Wong, M.S., 2023. A novel geometric feature-based wood-leaf separation method for large and crown-heavy tropical trees using handheld laser scanning point cloud. Int. J. Remote Sens. 44 (10), 3227–3258.
Wilson, B.F., 2000. Apical control of branch growth and angle in woody plants. Am, J. Bot 87 (5), 601–607.
Wongchai, W., Onsree, T., Sukkam, N., Promwungkwa, A., Tippayawong, N., 2022. Machine learning models for estimating above ground biomass of fast growing trees. Expert. Syst. Appl. 199, 117186.
World Meteorological Organization, 2024. *2024 is on track to be hottest year on record as warming temporarily hits 1.5°C.* https://wmo.int/news/media-centre/2024-track-be-hottest-year-record-warming-temporarily-hits-15degc.
Xu, D., Wang, H., Xu, W., Luan, Z., Xu, X., 2021. LiDAR applications to estimate forest biomass at individual tree scale: opportunities, challenges and future perspectives. Forests. 12 (5), 550.
Yang, M., Zhou, X., Liu, Z., Li, P., Tang, J., Xie, B., Peng, C., 2022. A review of general methods for quantifying and estimating urban trees and biomass. Forests. 13 (4), 616.
Yuen, J.Q., Fung, T., Ziegler, A.D., 2016. Review of allometric equations for major land covers in SE Asia: uncertainty and implications for above-and below-ground carbon estimates. For, Ecol, Manage 360, 323–340.
Zhang, H., Lee, C.K., Law, Y.K., Chan, A.H., Zhang, J., Gale, S.W., Wu, J., 2024. Integrating both restoration and regeneration potentials into real-world forest restoration planning: a case study of Hong Kong. J. Environ., Manage 369, 122306.