Forest Science (2026) 72:305–327
https://doi.org/10.1007/s44391-026-00062-y

ORIGINAL ARTICLE

Anomaly Detection and Machine Learning for Stand-Level Growth and Yield Modeling in Hybrid Eucalypt Plantations

Gianmarco Goycochea Casas[1] · Zool Hilmi Ismail[2] · Pedro Henrique Fontes dos Santos[1] · Leonardo Ippolito Rodrigues[3] · Tassius Menezes Araújo[1] · Helio Garcia Leite[1]

Received: 11 September 2025 / Accepted: 11 February 2026 / Published online: 23 February 2026
© The Author(s) 2026

Abstract

Accurate prediction of forest growth and yield is essential for strategic planning in intensive plantation management. This study evaluates whether unsupervised anomaly detection can be used as a systematic data-quality layer in stand-level growth and yield modeling, after standard consistency checks have been applied. We used a multi-regional continuous forest inventory of hybrid *Eucalyptus urophylla* × *Eucalyptus grandis* plantations in Minas Gerais, Brazil (6,553 measurements from 1,749 permanent plots in three regions). Four unsupervised methods, Isolation Forest, One-Class SVM, Local Outlier Factor and a dense autoencoder, were applied within each region to identify multivariate anomalies in stand age, volume, basal area and dominant height. Using Isolation Forest with a contamination rate of 0.10, approximately 10% of records were flagged as anomalous, but this corresponded to the complete removal of 2–4% of plots, while most plots either retained all measurements or lost only a subset of them. We then calibrated nine nonlinear machine learning models to predict stand volume at the second measurement (V_2) under two scenarios: using the full dataset and using the anomaly-filtered dataset. Gradient Boosting and other tree ensembles achieved the best performance in both cases ($R^2 \approx 0.86$–0.88, relative RMSE ≈ 13–15%), and differences in accuracy between full and filtered datasets were small (changes in $R^2 < 0.03$ and in relative RMSE < 1 percentage point). A sensitivity analysis across contamination levels from 0.05 to 0.20 confirmed that a 10% threshold offers a practical compromise, removing a consistent set of multivariate extremes while preserving plot coverage and predictive performance.

Keywords Outlier analysis · Data preprocessing · Forest inventory · Forest management · Artificial Intelligence

Extended author information available on the last page of the article

Introduction

Forests are integral in addressing global challenges, including carbon sequestration, provision of renewable materials, energy resources, and ecosystem services. Consequently, robust models of forest stand dynamics, growth, and yield are imperative across various scales, from individual stands to entire landscapes. Such growth and yield models form the scientific basis for forest management planning and policy, underpinning decisions about silvicultural interventions, harvest scheduling, and sustainable yield regulation (Blanco and Lo 2023). Accurate forest growth predictions are crucial for planning in the forestry industry, including timber supply chains, financial valuations, and climate change mitigation. These frameworks support ecological stewardship and industrial use of forest resources globally.

Historically, forest growth and yield prediction has relied on statistical models, typically based on allometric equations or regressions linking tree metrics (e.g. diameter, height, age, density) to future growth. Classic approaches include linear and nonlinear regression, mixed-effects models, and other parametric methods, which operate under assumptions of independent, normally distributed, and homoscedastic data. In practice, these statistical assumptions are often violated by forestry datasets that are hierarchical (plots within stands) and repeatedly measured over time (Casas et al. 2022; Eitzel et al. 2013). Weiskittel et al. (2011) noted over a decade ago that conventional yield models tend to oversimplify relationships and may fail under changing external factors like climate or pest outbreaks. Studies echo these concerns, emphasizing that purely empirical models, usually based on fixed functional forms, cannot fully represent the biological complexity and variability of tree growth (Zhang et al. 2023).

Advances in computing have revolutionized forest modeling. Machine learning (ML) methods now offer powerful alternatives to traditional models by learning complex patterns directly from data. Modern AI enables ML to model forest growth with no a priori distributional requirements, deeper nonlinear relationship extraction, and higher predictive accuracy (Silva et al. 2024). Recent studies show machine learning's rising significance in forest management. (De Freitas et al. 2024) combined machine learning with cellular automata to simulate wood production with high accuracy (~4%), while (Taylor et al. 2023) enhanced the 3-PG process-based model using ensemble decision trees, achieving over 95% correlation in growth metrics prediction. (Chen et al. 2024) applied ensemble models like XGBoost and Random Forest to predict forest age using satellite data, showing superior performance across forest types. (Fu et al. 2024) introduced the SCARF algorithm, which leverages cascade random forests and Landsat data to improve biomass prediction accuracy in subtropical forests. (Casas et al. 2023) demonstrated that structuring growth data monthly and using deep learning improves yield prediction reliability for Eucalyptus stands. Another study applies ML models such as Artificial Neural Networks, Support Vector Machines, and Random Forests to estimate net tree growth in Brazil's Atlantic Forest using variables like soil acidity, basal area, abandonment age, and precipitation, finding that RF offers the highest predictive accuracy (Rocha et al. 2024).

The performance of any data-driven model, including ML approaches, hinges on the quality of the input data. In forestry, data quality issues are a persistent challenge. Field measurements and sensor observations can introduce errors or inconsistencies: for instance, mis-recorded tree diameters or heights, plot location errors, or missing readings from remote sensors. Large-scale forest databases often contain *anomalies* (outliers) arising from such measurement errors, atypical environmental conditions, or extreme disturbance events. For instance, when combining ground inventories with auxiliary data, misalignment in timing or spatial location can yield outlier observations that do not reflect true stand conditions (Knott et al. 2023). These anomalies, if left unaddressed, can *distort* model training and predictions by exerting disproportionate influence on the fit.

To mitigate the adverse effects of data anomalies, researchers have increasingly turned to anomaly detection techniques as a preprocessing step in the modeling pipeline. Anomaly (outlier) detection refers to the identification of observations that deviate markedly from the expected pattern of the majority of the data. By detecting and handling such aberrant points prior to model fitting, one can improve model robustness and accuracy (Davis and Clark 2011; Liu et al. 2021; Vajda et al. 2024). In essence, anomaly detection acts as a safeguard for data quality, it *filters* erroneous or extraordinary data points, thereby reducing their negative impact on model predictions. In the context of forest growth modeling, this step can enhance the reliability of yield forecasts by ensuring that the ML algorithms learn from representative, clean data reflective of underlying growth processes, rather than noise. Despite its clear benefits, anomaly detection is not yet standard practice in many forest modeling studies, and outliers are often handled in ad-hoc ways or overlooked entirely.

Forest inventory data serves as the empirical foundation for these models, and the reliability of predictions depends heavily on the quality and consistency of the data used. In operational forestry, it is standard practice to subject inventory data to a consistency analysis before applying any predictive modeling. This initial step aims to identify and address gross measurement errors, systematic biases, and logically inconsistent observations. These preprocessing protocols are essential to ensure that the data meets basic assumptions required for reliable modeling, including coherence between variables, absence of obvious outliers, and internal logic in plot-level data structures (Campos and Leite 2017).

However, while consistency analysis corrects major discrepancies, it does not always detect anomalies, subtle or multivariate outliers that may arise from rare events, atypical growth conditions, or residual data collection noise. These undetected anomalies can still exert disproportionate influence on data-driven models, particularly those using ML algorithms that are sensitive to the overall structure and distribution of training data (Nassif et al. 2021).

This study investigates whether combining machine learning with unsupervised anomaly detection allows the use of fewer effective inventory measurements in eucalyptus plantations without compromising the accuracy of growth and yield models. We focus on a realistic operational setting in which the forest inventory has already passed the standard consistency checks and then examine the added value of an anomaly detection step applied afterwards.

The main contributions of this study are threefold.

1. We evaluate unsupervised anomaly detection in a forest growth and yield context after conventional consistency checks, testing the added value of a statistical outlier filtering layer on an operational inventory database
2. We show that a modest removal of anomalous records can reduce the number of effective field measurements while maintaining the accuracy of machine learning based growth and yield models and preserving most of the permanent plot network
3. We present a scalable modeling pipeline that combines anomaly detection and machine learning for hybrid eucalyptus plantations and provides a transparent, reproducible framework for integrating intelligent data screening into forest inventory workflows

Material and Methods

The data come from a continuous forest inventory (CFI) of industrial hybrid *Eucalyptus urophylla*×*Eucalyptus grandis* plantations in Minas Gerais, southeastern Brazil. The operational area was stratified into three management regions (A, B and C), which differ in site quality and management history but belong to the same plantation estate.

The CFI consists of permanent plots remeasured over time. After applying the standard consistency checks and removing records with missing values in any predictor or response, the working dataset comprised 6,553 plot measurements from 1,749 plots: 4,174 records (726 plots) in Region A, 1,507 records (605 plots) in Region B, and 872 records (418 plots) in Region C. Plots had between 1 and 21 measurements (median = 3).

For each measurement, the following stand attributes were available: stand age at the first and second measurements (t_1, t_2 months), stand volume at the first measurement (V_1, $m^3 \cdot ha^{-1}$), basal area (G_1, $m^2 \cdot ha^{-1}$), dominant height (H_{dom}, m), and stand volume at the second measurement (V_2, $m^3 \cdot ha^{-1}$). The response variable in all models was V_2, while t_1, t_2, V_1, G_1, H_{dom} and Region (one-hot encoded) were used as predictors. Summary statistics by region show that Region A contains on average slightly older and more productive stands, whereas Region C tends to be younger and less productive (Table 1).

Table 1 Summary of the forest inventory dataset by region. t_1= stand age at first measurement (months); V_1 = stand volume at first measurement ($m^3 \cdot ha^{-1}$); H_{dom} = dominant height (m); V_2 = stand volume at second measurement ($m^3 \cdot ha^{-1}$)

Region	N records	N plots	Mean t_1 (months)	Mean V_1 ($m^3 \cdot ha^{-1}$)	Mean H_{dom} (m)	Mean V_2 ($m^3 \cdot ha^{-1}$)
A	4,174	726	37.3	110	18.8	225.9
B	1,507	605	35	112.4	18.2	214.3
C	872	418	28.3	82.9	17	211.3
Total	6,553	1,749	35.6	107	18.4	221.3

Anomaly Detection

To detect anomalies in forest inventory data, four unsupervised algorithms were applied: Isolation Forest (Liu et al. 2008), One-Class Support Vector Machine (OC-SVM) (Schölkopf et al. 2001), Local Outlier Factor (LOF) (Breunig et al. 2000), and Autoencoder (Goodfellow et al. 2016) based anomaly detection. These methods were implemented using Python, combining classical machine learning models and deep learning techniques. The input dataset consisted of four numerical variables: stand age (Months), volume (m^3/ha), basal area (m^2/ha), and dominant height (meters), along with the categorical variable "*Region*." Data were loaded from a spreadsheet and scaled using the *StandardScaler* from *scikit-learn* to standardize the variables to zero mean and unit variance.

Isolation Forest

The Isolation Forest algorithm was employed to detect anomalies in the forest inventory dataset due to its efficiency and robustness in high-dimensional settings. This method is based on the principle that anomalies are more susceptible to isolation than regular observations. Isolation Forest constructs an ensemble of binary trees (isolation trees), where each tree recursively partitions the data by randomly selecting a feature and a split value within its range. Each observation x traversed a tree until it became isolated, and the path length $h(x)$ denoted the number of edges traversed. The anomaly score $s(x, n)$ was calculated using the formula (Eq. 1):

$$s(x, n) = 2^{-\frac{E(h(x))}{c(n)}} \tag{1}$$

Note here that $E(h(x))$ is the average path length required to isolate point x, and $c(n)$ is the average path length of unsuccessful searches in a binary search tree, approximated by Eq. 2:

$$c(n) = 2H(n-1) - \frac{2(n-1)}{n}, with H(n) \approx ln(n) + \gamma \tag{2}$$

Note here that $\gamma \approx 0.5772$ being the Euler–Mascheroni constant. Anomaly scores close to 1 indicated that the observation was likely an outlier, while scores near 0 suggested that the observation was likely normal.

The implementation was carried out using the *IsolationForest* class from the *scikit-learn* library, with the following configuration: 100 isolation trees were used; the expected proportion of anomalies was set to 10%; the number of samples per tree was selected automatically as the minimum of 256 or the total number of observations; and the random seed was fixed of 42 to ensure reproducibility. The model was trained on standardized forest variables and predictions were generated, classifying instances as anomalies (−1) or normal observations (1), along with continuous anomaly scores obtained.

One-Class Support Vector Machine

The One-Class Support Vector Machine (OC-SVM) algorithm was employed to identify anomalies within the forest inventory dataset. OC-SVM is an unsupervised learning method that estimates the support of a high-dimensional distribution, effectively distinguishing the regions in feature space where the data is concentrated from the regions where it is sparse. This approach is particularly suitable for anomaly detection tasks, where the goal is to identify instances that deviate significantly from the norm.

The OC-SVM algorithm operates by mapping the input data into a high-dimensional feature space using a kernel function and then finding the hyperplane that best separates the data from the origin, maximizing the margin. The decision function $f(x)$ for a data point x is defined as Eq. 3:

$$f(x) = w \cdot \phi(x) - \rho \tag{3}$$

Note here that w is the weight vector, $\phi(x)$ is the feature mapping of x, and ρ is the offset.

The model was trained by solving the following convex optimization problem (Eq. 4 and Eq. 5):

$$min_{\{w,\xi,\rho\}}\left(\frac{1}{2}\|w\|^2\right) + \frac{1}{vn}\sum_{i=1}^{n}\xi_i - \rho \tag{4}$$

subject to:

$$(w * \phi(x)) \geq \rho - \xi_i, \xi_i \geq 0 \tag{5}$$

Note here that ξ_i are slack variables allowing for some data points to lie within the margin, $v \in (0,1]$ is a parameter that controls the trade-off between the fraction of outliers and the margin, and n is the number of training samples.

In this study, the OC-SVM was implemented using the *OneClassSVM* class from the *scikit-learn* library and it was configured using a radial basis function (RBF) kernel to capture non-linear relationships among the variables, and the value of v was set to 0.1, which allowed the model to tolerate up to 10% of the data as outliers. The kernel coefficient was determined automatically based on the data variance and number of features, and the optimization procedure was conducted with a tolerance of 0.001 and without limiting the number of iterations. After training, the model assigned a decision score to each observation, from which outliers were identified as those lying outside the estimated boundary of the distribution. The model was trained on the standardized dataset. After training, the predict method was used to classify each observation as either an inlier (1) or an outlier (−1).

Local Outlier Factor

The Local Outlier Factor (LOF) algorithm was employed to detect anomalies within the forest inventory dataset by assessing the local density deviation of each data point relative to its neighbors. This density-based method identified observations that exhibited significantly lower local densities compared to their surrounding points, thus flagging them as potential outliers. The process began by determining the *k*-distance for each data point, which is the distance to its *k*-th nearest neighbor. Subsequently, the reachability distance between two points, say A and B, was calculated as the maximum of the *k*-distance of B and the actual distance between A and B. This measure ensured a more stable estimation of local densities, especially in regions with varying densities. The local reachability density (LRD) of a point A was then computed as the inverse of the average reachability distance from A to its *k* nearest neighbors (Eq. 6):

$$LRD(A) = \left(\frac{1}{\left| N_{k(A)} \right|} \sum_{B \in N_{k(A)}} reach - dist_k(A, B) \right)^{-1} \tag{6}$$

Note here that where $N_{k(A)}$ denotes the set of *k* nearest neighbors of A. Finally, the LOF score for point A was determined by comparing its LRD to those of its neighbors (Eq. 7):

$$LOF_k(A) = \frac{1}{\left| N_{k(A)} \right|} \sum_{B \in N_{k(A)}} \frac{LRD(A)}{LRD(B)} \tag{7}$$

A LOF score approximately equal to 1 indicated that the point's local density was comparable to its neighbors, suggesting it was not an outlier. Conversely, a significantly higher LOF score signified that the point resided in a region of lower density relative to its neighbors, marking it as a potential anomaly. The LOF was implemented using the *LocalOutlierFactor* class from the *scikit-learn* library. In this study, the LOF algorithm was configured with 20 neighbors ($k = 20$) to balance sensitivity and robustness in detecting local anomalies. The contamination parameter was set to 0.1, reflecting the expectation that approximately 10% of the data could be anomalous. The Euclidean distance metric was utilized to compute distances between data points.

Autoencoder

An autoencoder is a type of artificial neural network designed for unsupervised learning, widely used for anomaly detection tasks in complex datasets. It consists of two main components: an encoder and a decoder. The encoder compresses the input data into a lower-dimensional latent representation, while the decoder attempts to reconstruct the original input from this representation. The core assumption is that normal data patterns can be reconstructed with minimal error,

whereas anomalous observations will exhibit significantly higher reconstruction errors.

Given an input vector $x \in \mathbb{R}^n$, the encoder x to a latent representation $x \in \mathbb{R}^m (m < n)$ using a nonlinear transformation function $f_{\theta(x)}$ (Eq. 8):

$$z = f_{\theta(x)} = \sigma\left(W_e x + b_e\right) \tag{8}$$

Note here that W_e and b_e are the weights and biases of the encoder, and σ denotes the activation function, here the Rectified Linear Unit (ReLU).

The decoder reconstructs the input from the latent representation via Eq. 9:

$$\hat{x} = g_{\phi(z)} = \sigma\left(W_d z + b_d\right) \tag{9}$$

Note here that W_d and b_d are the weights and biases of the decoder. The output layer uses the sigmoid activation function to normalize the reconstruction output in the range [0, 1].

The autoencoder is trained by minimizing the mean squared error (MSE) between the input and its reconstruction. After training, the reconstruction error is computed for each input. Anomalies are identified as those instances whose reconstruction error exceeds a predefined threshold. In this study, the threshold was set at the 90th percentile of the reconstruction error distribution, corresponding to a contamination rate of 10%.

The implementation was performed using TensorFlow and Keras. The autoencoder architecture included: An input layer with four neurons (representing stand age, volume, basal area, and dominant height), two encoding layers with ReLU activation functions, two symmetric decoding layers, and a sigmoid output layer. The model was trained for 50 epochs using a batch size of 32 and the Adam optimizer. No validation split was applied to ensure full utilization of the dataset in this unsupervised learning context.

Evaluation Strategy and Anomaly-Based Plot Reduction

To quantify the effect of anomaly detection on growth and yield model performance and on the number of plots required for calibration, we adopted a two-stage evaluation strategy based on (i) alternative datasets (with and without anomalies) and (ii) multiple machine learning models.

First, two main datasets were defined:

1. Full Data, containing all inventory records that passed conventional consistency checks and had complete information for the predictor and response variables;

2. Filtered Data, obtained by removing all records classified as anomalous (label $= -1$) by the Isolation Forest model with a contamination rate of 0.10, applied independently within each region.

In all modeling scenarios, the response variable was stand volume at the second measurement (V_2, $m^3 \cdot ha^{-1}$). Predictor variables comprised stand age at the first and second measurements (t_1 and t_2 months), stand volume, basal area and dominant height at the first measurement (V_1, $m^3 \cdot ha^{-1}$; G_1, $m^2 \cdot ha^{-1}$; H_{dom}, m), and the categorical region indicator, which was included via one-hot encoding. Thus, both Full Data and Filtered Data share the same variable set, differing in the presence or absence of records flagged as anomalous.

For each dataset (Full Data and Filtered Data), a set of nonlinear machine learning models was calibrated: Random Forest, Extra-Trees, Gradient Boosting, Support Vector Regression (SVR), k-Nearest Neighbours (KNN), Decision Tree, multilayer perceptron (MLP), AdaBoost and Bagging; XGBoost was also included when available. Model evaluation was based on fivefold cross-validation, selected to balance bias and variance in error estimation given the sample size, using the entire dataset in each scenario rather than a single arbitrary training–test split. In each fold, four folds were used for training and one for validation, ensuring that all observations served as validation data exactly once.

To avoid reliance on arbitrary default settings, each model was tuned via *GridSearchCV* (scikit-learn) within the cross-validation framework. For each algorithm, a modest but targeted hyperparameter grid was defined: for Random Forest and Extra-Trees, combinations of number of trees (200, 400), maximum depth (none, 20), minimum samples per split (2, 5) and minimum samples per leaf (1, 2); for Gradient Boosting and AdaBoost, number of estimators (300, 500) and learning rate (0.05, 0.10); for KNN, number of neighbours (5, 10, 20); for SVR, penalty parameter C (10, 100), epsilon (0.1, 0.2) and kernel scale (gamma set to "scale" or "auto"); for the MLP, size of the hidden layers ((50, 50) or (100, 50)) and L2 regularization strength (0.0001, 0.001); and for XGBoost, number of trees (300, 500), learning rate (0.05, 0.10) and maximum depth (3, 6). The grid was searched using negative mean squared error as the objective function. For each model and dataset, the best hyperparameter combination was selected, and cross-validated predictions were obtained using *cross_val_predict* to derive performance metrics.

Four metrics were computed for each model and scenario: (i) the Pearson correlation coefficient between observed and predicted V_2 (R), (ii) the coefficient of determination (R^2), (iii) the mean prediction error or bias (bias, defined as the mean of predicted minus observed V_2, $m^3 \cdot ha^{-1}$), and (iv) the relative root mean squared error (RRMSE, %), calculated as the root mean squared error divided by the mean observed V_2 and multiplied by 100. These metrics were computed using all cross-validated predictions, providing an unbiased estimate of model performance under both Full Data and Filtered Data.

In addition, for each model and dataset, we generated residual diagnostic plots based on the cross-validated predictions. We generated Q–Q plots of residuals against the normal distribution and residual-versus-fitted plots. These diagnostics were produced for all models and both scenarios but are reported in detail for the best-performing tree-based models, to illustrate how anomaly removal affects the distribution and spread of residuals.

Finally, to evaluate the sensitivity of the results to the choice of contamination level, we repeated the anomaly detection and modeling workflow using Isolation

Forest with contamination values of 0.05, 0.10, 0.15 and 0.20. For each contamination level, a filtered dataset was constructed (keeping records labeled as normal), and all models were recalibrated with the same fivefold cross-validation and hyperparameter tuning procedure. For every model and contamination level, we recorded R, R^2, bias, RRMSE, the number of records retained, and the number of plots retained. This sensitivity analysis provided a systematic assessment of how increasingly aggressive anomaly filtering affects both predictive performance and the amount of plot data available for calibration.

Results

Unsupervised Anomaly Detection

Across the full dataset (6,553 records), each method flagged close to 10% of records as anomalous. Isolation Forest, LOF, and the Autoencoder identified 656–657 anomalous records (10.0–10.1%), while One-Class SVM identified 661 anomalous records (10.1%). When summarized by region, all methods consistently flagged around 10% of records as anomalous within each region (Table 2). Figure 1 illustrates the age–volume distribution of anomalous and not anomalous records by region and method, and the corresponding filtered subsets. For Region A (4,174 records), the number of anomalies per method ranged from 418 to 422 (10.0–10.1%). For Region B (1,507 records), 149–151 records (9.9–10.0%) were flagged, and for Region C (872 records), 87–90 records (10.0–10.3%) were identified as anomalies.

Pairwise Jaccard similarity indices were computed to quantify overlap among the sets of records flagged as anomalous by each method (Table 3). Isolation Forest showed the greatest concordance with One-Class SVM (0.44) and the Autoencoder (0.37), whereas LOF exhibited comparatively low concordance with the other

Table 2 Number and percentage of anomalous records by region and method for the main contamination level (0.10)

Region	Method	N records	N anomalies	Pct anomalies (%)
A	Isolation Forest	4,174	418	10.01
A	OneClass SVM	4,174	422	10.11
A	LOF	4,174	418	10.01
A	Autoencoder	4,174	418	10.01
B	Isolation Forest	1,507	150	9.95
B	OneClass SVM	1,507	149	9.89
B	LOF	1,507	151	10.02
B	Autoencoder	1,507	151	10.02
C	Isolation Forest	872	88	10.09
C	OneClass SVM	872	90	10.32
C	LOF	872	87	9.98
C	Autoencoder	872	88	10.09

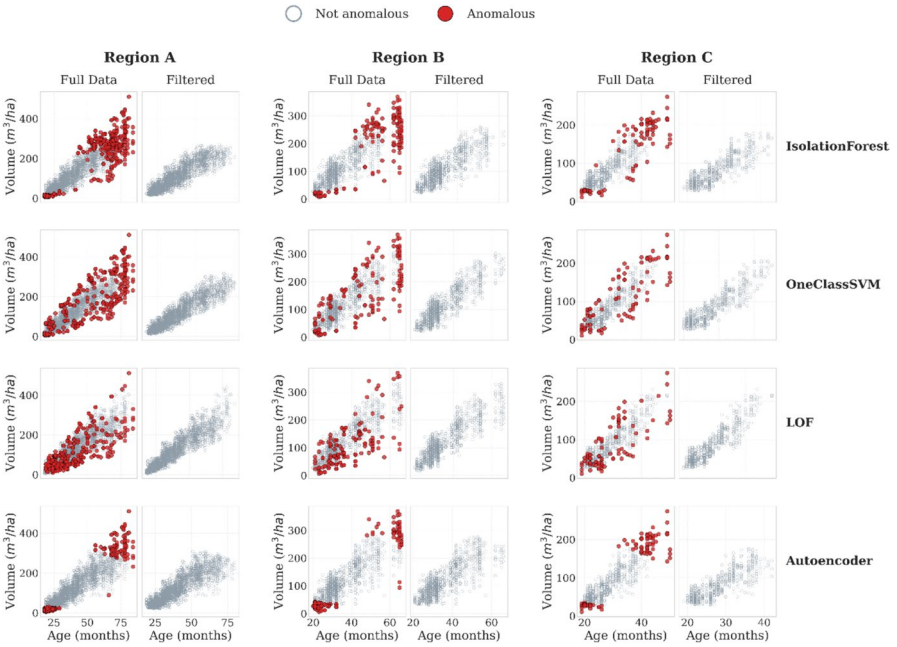


Fig. 1 Age–volume relationship (t_1 vs. V_1) by region and unsupervised method. Grey points are not anomalous records and red points are anomalous records flagged by each method. "Filtered" panels show only not anomalous records. Regions A–C correspond to the three management regions of the eucalypt plantation estate

Table 3 Pairwise Jaccard similarity indices measuring the overlap of records classified as anomalous by each unsupervised method

	OneClass SVM	LOF	Autoencoder
Isolation Forest	0.44	0.15	0.37
OneClass SVM	—	0.30	0.24
LOF	—	—	0.07

The matrix is symmetric; redundant entries are omitted and diagonal elements equal to 1 are suppressed

methods (0.07–0.30). The strongest pairwise agreement was observed between Isolation Forest and One-Class SVM, and the weakest between LOF and the Autoencoder (0.07).

For the main method used in the subsequent modeling (Isolation Forest), summary statistics of stand attributes were computed for records classified as normal (label = 1) and anomalous (label = −1) (Table 4). Anomalous records had higher mean age (t_1), volume at the first measurement (V_1), basal area (G_1), dominant height (H_{dom}), and volume at the second measurement (V_2) than normal records. Across the full dataset, 656 records were flagged as anomalous by Isolation Forest and 5,897 as not anomalous.

Table 4 Mean stand attributes for records classified as anomalous (− 1) and normal (1) by Isolation Forest (main contamination level 0.10)

Anomaly label (Isolation Forest)	t_1(months)	V_1 (m^3/ha)	G_1 (m^2/ha)	H_{dom} (m)	V_2 (m^3/ha)
− 1 (anomalous)	50.72	189.55	17.28	22.41	266.11
1 (not anomalous)	33.87	97.85	12.42	17.94	216.30

Loss of Plots after Anomaly Filtering

Anomaly filtering for the main analysis was performed using Isolation Forest with a contamination parameter of 0.10, applied separately within each region. Records classified as anomalous (label = − 1) were removed, and records classified as normal (label = 1) were retained for subsequent modeling.

At the plot (parcel) level, the impact of anomaly filtering was much smaller (Table 5). In the full dataset, there were 1,749 plots: 726 in Region A, 605 in Region B, and 418 in Region C. After removing anomalous records, 1,708 plots remained in the filtered dataset, corresponding to a loss of 41 plots in total (2.34% of all plots). Region-specific reductions were 3 plots in Region A (0.41% of plots), 18 plots in Region B (2.98%), and 20 plots in Region C (4.78%).

To better characterize how filtering affected the measurement history of individual plots, plots were classified into three categories according to the Isolation Forest labels of their records. *All removed*: all records of the plot were labeled as anomalous (− 1), so the plot disappeared entirely from the filtered dataset. *Some removed*: the plot has at least one not anomalous record (1) and at least one anomalous record (− 1); the plot remains in the filtered dataset but with a reduced number of measurements. *None removed*: all records of the plot were labeled as not anomalous (1); the plot and all its measurements are retained unchanged in the filtered dataset.

Using this classification, 41 plots (2.34%) were fully removed, 390 plots (22.3%) retained at least one record but lost one or more measurements, and 1,318 plots (75.4%) were completely unaffected by anomaly filtering (Table 6). By region, 3 plots in Region A, 18 in Region B and 20 in Region C fell into the "all removed" category. The majority of plots in each region (502 in Region A, 476 in Region B, and 340 in Region C) were in the "none removed" category, while 221, 111 and 58 plots in Regions A, B and C, respectively, had at least one record removed but remained in the dataset.

Table 5 Number of plots before and after anomaly filtering with Isolation Forest (contamination = 0.10), and relative reduction by region

Region	N plots full	N plots filtered	Plots removed	Plot reduction (%)
A	726	723	3	0.41
B	605	587	18	2.98
C	418	398	20	4.78
Total	1,749	1,708	41	2.34

Table 6 Plot-level impact of anomaly filtering with Isolation Forest (contamination = 0.10): number of plots with all records removed, with partial record removal, and with no records removed, by region and overall

Region	Plots all removed	Plots some removed	Plots none removed	Total plots
A	3	221	502	726
B	18	111	476	605
C	20	58	340	418
Total	41	390	1,318	1,749

Model Performance with Full and Filtered Datasets

Cross-Validated Performance with Full Data

When all records (6,553) were used, the best overall performance was obtained with Gradient Boosting, followed by the ensemble tree models and the multilayer perceptron (Table 7). Gradient Boosting achieved R = 0.938, R^2 = 0.880, RRMSE = 13.15%, and a nearly unbiased prediction (bias = 0.06 m^3/ha). Random Forest, Extra Trees, Bagging, and the MLP showed similar behavior, with R between 0.921 and 0.925, R^2 between 0.849 and 0.855, and RRMSE between 14.43% and 14.73%.

The remaining models had noticeably lower performance: KNN and SVR reached $R^2 \approx 0.81$–0.83 with RRMSE ≈ 15.7–16.4%, the single Decision Tree had R^2 = 0.787 and RRMSE = 17.49%, and AdaBoost showed the lowest accuracy with R^2 = 0.685 and RRMSE = 21.27%.

Across all models, the mean RRMSE in the Full Data scenario was 15.80%, with a range from 13.15% (Gradient Boosting) to 21.27% (AdaBoost).

Cross-Validated Performance after Anomaly Filtering

After removing records classified as anomalous by Isolation Forest (656 records), the filtered dataset included 5,897 records. Under this Filtered Data scenario, Gradient Boosting remained the best-performing model (Table 8), with R = 0.929, R^2 = 0.863, RRMSE = 13.43%, and bias = 0.11 m^3/ha. For the other ensemble

Table 7 Cross-validated performance of all models using the full dataset (Full Data scenario)

Model	R	R^2	bias (m^3/ha)	RRMSE (%)
Gradient Boosting	0.938	0.880	0.06	13.15
Random Forest	0.925	0.855	0.18	14.43
MLP	0.924	0.854	−0.88	14.46
Bagging	0.924	0.853	0.26	14.51
Extra Trees	0.921	0.849	0.27	14.73
KNN	0.910	0.828	0.55	15.70
SVR	0.905	0.812	−3.17	16.43
Decision Tree	0.888	0.787	−0.24	17.49
AdaBoost	0.840	0.685	7.89	21.27

Table 8 Cross-validated performance of all models using the filtered dataset (Filtered Data scenario; records flagged by Isolation Forest were removed)

Model	R	R^2	bias (m^3/ha)	RRMSE (%)
Gradient Boosting	0.929	0.863	0.11	13.43
MLP	0.912	0.832	0.65	14.86
Random Forest	0.910	0.829	0.23	14.99
Bagging	0.909	0.827	0.37	15.07
Extra Trees	0.909	0.827	0.38	15.08
KNN	0.896	0.803	0.72	16.08
SVR	0.895	0.796	− 2.60	16.37
Decision Tree	0.870	0.750	0.23	18.10
AdaBoost	0.829	0.671	7.99	20.77

models and the MLP, performance remained high: Random Forest, Extra Trees, Bagging, and MLP exhibited R between 0.909 and 0.912, R^2 between 0.827 and 0.832, and RRMSE between 14.86% and 15.08%.

KNN and SVR showed intermediate performance ($R^2 \approx 0.80$, RRMSE ≈ 16.1–16.4%), while the single Decision Tree and AdaBoost again showed the lowest accuracies, with $R^2 = 0.750$ and 0.671, and RRMSE = 18.10% and 20.77%, respectively.

The mean RRMSE across models in the Filtered Data scenario was 16.08%, with values ranging from 13.43% (Gradient Boosting) to 20.77% (AdaBoost).

For completeness, residual diagnostic plots for all models under both the full and anomaly-filtered datasets are reported in Supplementary Fig. S1.

Differences between Full and Filtered Scenarios

The direct comparison between Full Data and Filtered Data for each model (Table 9, Fig. 2) shows that, for most models, the relative changes in performance metrics were small. For the tree-based ensembles and the MLP, RRMSE increased by 0.28–0.57 percentage points and R^2 decreased by 0.017–0.027 after anomaly

Table 9 Change in performance metrics between Full Data and Filtered Data for each model (Filtered – Full)

Model	ΔR (Filtered – Full)	ΔR^2	Δbias (m^3/ha)	ΔRRMSE (p.p.)
Gradient Boosting	− 0.009	− 0.017	0.06	0.28
Random Forest	− 0.014	− 0.027	0.05	0.57
MLP	− 0.012	− 0.023	1.53	0.40
Bagging	− 0.014	− 0.026	0.11	0.56
Extra Trees	− 0.012	− 0.022	0.11	0.34
KNN	− 0.014	− 0.026	0.17	0.38
SVR	− 0.010	− 0.016	0.57	− 0.06
Decision Tree	− 0.018	− 0.037	0.46	0.60
AdaBoost	− 0.011	− 0.014	0.10	− 0.51

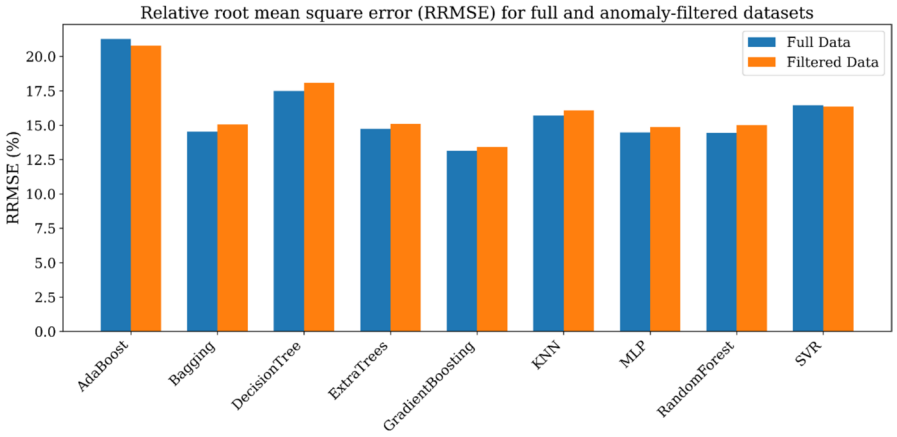


Fig. 2 Relative root mean square error (RRMSE) for all models under the full and anomaly-filtered datasets

removal. For example, Gradient Boosting changed from RRMSE = 13.15% and $R^2 = 0.880$ (Full Data) to RRMSE = 13.43% and $R^2 = 0.863$ (Filtered Data), a difference of + 0.28 percentage points in RRMSE and − 0.017 in R^2. Random Forest changed from RRMSE = 14.43% and $R^2 = 0.855$ to RRMSE = 14.99% and $R^2 = 0.829$.

KNN, SVR, and Decision Tree showed similar patterns, with RRMSE differences of + 0.38 to + 0.60 percentage points and decreases in R^2 between − 0.016 and − 0.037. AdaBoost and SVR were the models for which RRMSE slightly decreased after filtering (− 0.51 and − 0.06 percentage points, respectively), although both remained clearly less accurate than the best-performing models.

The mean RRMSE across all models increased from 15.80% (Full Data) to 16.08% (Filtered Data), i.e. by 0.28 percentage points, while the mean R^2 decreased from 0.820 to 0.807.

All computational analyses, including data preprocessing and model training, were performed on a high-performance workstation equipped with an Intel Core i7-13700H processor, 32 GB of RAM, and an NVIDIA GeForce RTX 4060 GPU. The computational time required for the complete training, hyperparameter tuning, and validation process varied among the methods, ranging from less than 0.2 s for simpler algorithms (e.g., KNN, Decision Trees) to approximately 15s for complex ensemble methods and neural networks.

Sensitivity Analysis of the Contamination Level

Effect on the Number of Records and Plots

At the record level, increasing the contamination parameter from 0.05 to 0.20 progressively reduced the number of observations used for model calibration (Table 10). Starting from 6,553 records in the full dataset, 4.97% of records were

Table 10 Number of records and plots retained after filtering with Isolation Forest at different contamination levels, and percentage of records and plots removed relative to the full dataset (6,553 records and 1,749 plots)

Variable	Contamination			
	0.05	0.10	0.15	0.20
N records filtered	6,227	5,897	5,570	5,241
Records removed	326	656	983	1,312
Pct records removed (%)	4.97	10.01	15	20.02
N plots full	1,749	1,749	1,749	1,749
N plots filtered	1,729	1,708	1,677	1,631
Plots removed	20	41	72	118
Pct plots removed (%)	1.14	2.34	4.12	6.75

N denotes the number of records (for record-level variables) or plots (for plot-level variables), and Pct denotes the percentage removed relative to the full dataset

removed at contamination 0.05 (6,227 records retained), 10.01% at 0.10 (5,897 records retained), 15.00% at 0.15 (5,570 records retained), and 20.02% at 0.20 (5,241 records retained).

At the plot level, the impact was smaller. From a total of 1,749 plots, the number of plots retained decreased from 1,729 plots at contamination 0.05 to 1,631 plots at contamination 0.20 (Table 10). This corresponds to 1.14% of plots removed at 0.05 (20 plots), 2.34% at 0.10 (41 plots), 4.12% at 0.15 (72 plots), and 6.75% at 0.20 (118 plots).

Effect on Model Performance across Contamination Levels

For each contamination level, all nine models were recalibrated on the filtered dataset and evaluated with fivefold cross-validation. Table 11 summarizes the distribution of RRMSE and R^2 values across models for each contamination level (mean, minimum, and maximum).

Mean RRMSE across models increased from 15.86% at contamination 0.05 to 16.37% at contamination 0.20, while mean R^2 decreased from 0.808 to 0.788. For all contamination levels, the best-performing model was Gradient Boosting, with RRMSE between 13.16% and 13.56% and R^2 between 0.870 and 0.857. The weakest model remained AdaBoost, with RRMSE between 20.26% and 20.83% and R^2 between 0.692 and 0.662.

For the Gradient Boosting model, RRMSE and R^2 values varied slightly as contamination increased (Table 12). RRMSE changed from 13.16% at contamination 0.05 to

Table 11 Summary of model performance across contamination levels: mean, minimum and maximum RRMSE (%) and R^2 across all models

Contamination	mean RRMSE (%)	min RRMSE (%)	max RRMSE (%)	mean R^2	min R^2	max R^2
0.05	15.86	13.16	20.26	0.808	0.692	0.870
0.10	16.08	13.43	20.77	0.800	0.671	0.863
0.15	16.32	13.53	20.70	0.792	0.671	0.859
0.20	16.37	13.56	20.83	0.788	0.662	0.857

Table 12 Performance of selected models (Gradient Boosting and Random Forest) across contamination levels

Model	Contamination	R	R^2	bias (m^3/ha)	RRMSE (%)
Gradient Boosting	0.05	0.933	0.870	0.02	13.16
Gradient Boosting	0.10	0.929	0.863	0.11	13.43
Gradient Boosting	0.15	0.927	0.859	0.12	13.53
Gradient Boosting	0.20	0.926	0.857	0.15	13.56
Random Forest	0.05	0.914	0.836	0.32	14.79
Random Forest	0.10	0.910	0.829	0.23	14.99
Random Forest	0.15	0.908	0.825	0.02	15.11
Random Forest	0.20	0.905	0.817	0.27	15.31

13.43%, 13.53% and 13.56% at 0.10, 0.15 and 0.20, respectively, while R^2 decreased from 0.870 to 0.863, 0.859 and 0.857. Similar trends were observed for the other tree-based ensemble models (Random Forest, Extra Trees, Bagging) and the MLP, with RRMSE increasing by approximately 0.3–0.7 percentage points and R^2 decreasing by 0.01–0.02 between contamination 0.05 and 0.20.

As an example, for Random Forest, RRMSE increased from 14.79% at contamination 0.05 to 14.99%, 15.11% and 15.31% at 0.10, 0.15 and 0.20, respectively, while R^2 decreased from 0.836 to 0.829, 0.825 and 0.817. For Extra Trees, RRMSE increased from 14.92% to 15.08%, 15.34% and 15.59%, and R^2 decreased from 0.833 to 0.827, 0.819 and 0.811. The MLP model showed RRMSE values of 14.67%, 14.86%, 15.15% and 15.06%, and R^2 values of 0.838, 0.832, 0.824 and 0.823 for contamination levels 0.05, 0.10, 0.15 and 0.20, respectively.

Across all contamination levels, the ranking of models remained stable: Gradient Boosting consistently provided the lowest RRMSE and the highest R^2, followed by Random Forest, Extra Trees, Bagging, and MLP. KNN and SVR showed intermediate performance, while Decision Tree and AdaBoost remained the least accurate, with higher RRMSE and lower R^2 at all contamination levels. Prediction bias remained small (close to zero) for the ensemble models and the MLP, and substantially larger for AdaBoost.

Discussion

Operational Implications and Use-Case Dependent Removal of Anomalies

From an operational standpoint, our results indicate that unsupervised anomaly detection is best interpreted as a flexible tool to flag and manage atypical observations, rather than as a rigid rule to always remove them. In our case study, applying Isolation Forest with a contamination rate of 10% after standard consistency checks led to the removal of about 10% of plot records, but only 2.3% of plots disappeared completely, while many plots simply lost a subset of their measurements. The fact that model performance remained very similar with and

without these anomalous records, particularly for the best-performing tree ensembles, suggests that growth and yield models can be robust to modest reductions in training data that mainly affect multivariate extremes, especially when the underlying database has already undergone thorough operational quality control.

Analyses of random forest regression in ecology emphasize that addressing major data issues (e.g. gross errors, strong leverage points) leads to more stable predictions, but that the models themselves are relatively robust to moderate irregularities (Allen Akselrud 2024). In our case, the small performance differences between full and filtered datasets indicate that most of the influential measurement or coding problems had already been addressed by standard consistency checks, and that the anomalies detected here behave more like multivariate extremes than like pervasive noise.

The way anomaly flags are used should therefore be explicitly linked to management objectives. If the primary goal is to develop a planning model that accurately reflects typical stand performance along the main productivity gradient, it can be both acceptable and useful to exclude part of the extremes. This is analogous to long-standing practices in growth and yield modeling where obviously influential plots or highly atypical management histories are down-weighted or excluded so that the model focuses on the conditions most relevant for "business-as-usual" planning. In this context, the anomaly-filtered dataset provides a well-defined core of conditions for which the model is well supported and easier to explain to practitioners.

By contrast, if objectives focus on estimating maximum yields, evaluating intensive management scenarios or studying exceptional sites, systematic exclusion of extreme observations is not appropriate. Studies on forest biomass and carbon estimation have shown that rare but extreme stands can critically affect estimates for high-productivity classes and for risk analysis (Knott et al. 2023). In such cases, flagged records should be reviewed to distinguish likely measurement or processing errors from genuinely high- or low-yield stands. Confirmed extremes should then be documented as special cases (e.g. fertilization trials, pest outbreaks, local damage) and, when valid, used in dedicated analyses, for example, by fitting separate high-productivity models, stress-testing planning scenarios, or quantifying risk under extreme conditions.

Under this view, the same anomaly detection pipeline can support several operational workflows. Plots with all measurements flagged as anomalous are natural candidates for targeted field verification or exclusion from the base calibration dataset, similar to the quality-control approaches proposed by Knott et al. (2023). Plots with a mixture of normal and anomalous measurements can remain in the monitoring network, using their non-anomalous records for standard growth and yield modeling while reserving anomalous records for diagnostic or scenario analyses. Plots with not anomalous measurements form the backbone of the inventory and calibration dataset. Thus, rather than replacing expert judgement or conventional consistency checks, anomaly detection provides a structured way to prioritize attention, field effort and modeling decisions, in line with recent calls to couple statistical tools with explicit decisions about how field data are used in inventory and planning.

Discussion in a Broader Context

This study examined whether, after a standard consistency analysis of forest inventory data, the additional application of anomaly detection techniques could reduce the number of plot measurements used for growth and yield prediction without compromising model performance. This question is particularly relevant to operational inventory and planning, where fieldwork is costly and where studies have demonstrated that data quality can be more limiting than model choice for predictive performance (Rota et al. 2024). Consistent with common forestry practice, our dataset had already undergone routine company checks to remove logically inconsistent or clearly erroneous measurements (e.g. negative values, impossible diameter–height combinations, missing key fields). Even so, using four unsupervised methods applied by region, approximately 10% of records were flagged as anomalous despite having passed all rule-based checks, confirming that multivariate extremes can remain hidden to classical consistency procedures.

When these anomalous records were removed using Isolation Forest with a 10% contamination rate, predictive performance of the growth and yield models changed very little. For the best-performing algorithms (Gradient Boosting and other tree ensembles), differences in R^2 between the full and filtered datasets were generally below 0.03, and changes in relative RMSE were typically below one percentage point; bias remained small in both scenarios and did not exhibit a systematic reduction after filtering. In contrast, Knott et al. (2023) reported reductions in estimator variance of up to 28% after removing outliers associated with GPS errors and zero-biomass artefacts in model-assisted biomass estimation using national inventory and remote sensing data. The much smaller performance gains observed in our study likely reflect both the different modeling objectives (stand-level prediction versus design-based biomass estimation) and the fact that we started from an already curated operational database. In other words, anomaly detection here is refining an already cleaned dataset, whereas in other contexts outliers may represent more severe data problems.

Mohammed et al. (2024) has highlighted that data curation is often more critical than model choice for achieving reliable predictions. Likewise, a study used quality control charts to detect deviations in annual inventory data, reinforcing the idea that systematic monitoring and correction of atypical values can stabilize long-term trends and model behavior (Roesch and Van Deusen 2010).

The largest benchmark of unsupervised anomaly detection algorithms on real-world tabular data, show that many algorithms tend to identify overlapping sets of anomalies and that a small subset of well-chosen methods is sufficient for most applied problems (Bouman et al. 2024). Similarly, our four methods (Isolation Forest, One-Class SVM, LOF and a dense autoencoder) agreed on a core subset of multivariate extremes and using that subset for filtering was enough to demonstrate the operational implications without major gains in accuracy. At the same time, Bouman and Heskes (2025) noted that autoencoders for anomaly detection can be sensitive to architecture and training choices, which reinforces the need for cautious use and benchmarking rather than blind trust in a single method.

Taylor et al. (2023) used random forest, extremely randomized trees, gradient boosting and XGBoost to predict site fertility and improve 3-PG-based growth projections for *Eucalyptus nitens*, and highlighted the importance of high-quality input data and interpretable ensemble models. In our study, similar ensemble decision-tree algorithms provided the highest predictive performance for stand volume and maintained their accuracy even when calibrated on anomaly-filtered datasets. Together, these studies suggest that combining ensemble ML models with explicit data-screening procedures is a promising direction for operational growth and yield modeling, but that the magnitude of the benefit depends strongly on the initial data quality and on whether outliers mainly represent gross errors or genuine extremes.

Several limitations and avenues for future work remain. We treated all records flagged as anomalous as candidates for exclusion, even though many of them likely represent genuine biophysical extremes (e.g. exceptionally productive stands, local damage, or atypical management histories) rather than measurement errors. In contrast to studies where outlier removal clearly improves estimator precision, our modest changes in R^2 and RRMSE underscore the risk of over-filtering informative but rare conditions when starting from a well-curated database. Future research should therefore distinguish more explicitly between "error-like" anomalies and biologically meaningful outliers and compare alternative treatments such as down-weighting, retaining anomalies for scenario analyses, or restricting their removal to particular modeling tasks (e.g. standard vs. high-yield scenarios). We used fixed contamination levels applied uniformly across regions; adaptive or model-driven thresholds informed by site class, management regime, residual patterns or risk tolerance may better align anomaly filtering with operational objectives, particularly in heterogeneous plantation estates. At the same time, results in ecology emphasize not only prediction but also the need to propagate uncertainty and account for data issues explicitly in model outputs (Allen Akselrud 2024). Integrating anomaly detection with formal uncertainty analysis, for example, by quantifying how filtering affects prediction intervals and not only point estimates, would help express the benefits and trade-offs of data screening directly in terms of risk for planning and decision making, consolidating anomaly detection as a standard, but not exclusive, component of growth and yield modeling workflows.

Finally, in this study anomalous records were removed only to quantify their effect on model performance and plot requirements. We do not advocate automatic deletion of all flagged records in all applications; instead, we recommend using anomaly flags to support case-dependent decisions about exclusion, down-weighting or dedicated scenario analysis, depending on whether the focus is on typical conditions or on extreme stands.

Conclusion

This study shows that integrating unsupervised anomaly detection with machine-learning growth and yield models is a practical data-quality control step, rather than a way to radically boost predictive accuracy. Using four unsupervised methods

applied by region, about 10% of records were flagged as anomalies, but this translated into the loss of only a small fraction of plots (≈2–5%), with most plots either unaffected or losing only a few measurements.

When models were calibrated with and without these anomalous records, the best algorithms (especially Gradient Boosting and other tree ensembles) maintained high predictive performance ($R^2 \approx 0.85$–0.88, RRMSE ≈ 13–15%) and changed very little between full and filtered datasets. Sensitivity analysis across contamination levels (5–20%) confirmed that a 10% threshold is a reasonable compromise, removing a clear set of multivariate extremes while preserving both plot coverage and model accuracy. In practical terms, the proposed workflow offers a transparent, reproducible way to clean inventory data and control outliers before fitting growth and yield models, with minimal cost in terms of usable plots and predictive power. Therefore, anomaly detection should be viewed as a complementary data-quality layer in stand-level growth and yield modelling, rather than as a substitute for careful inventory design and consistency checks.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s44391-026-00062-y.

Funding The Article Processing Charge (APC) for the publication of this research was funded by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) (ROR identifier: 00x0ma614). The authors did not receive support from any organization for the submitted work.

Data Availability The datasets used in this study are available from the corresponding author upon reasonable request.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Allen Akselrud, C.I. 2024. Random forest regression models in ecology: Accounting for messy biological data and producing predictions with uncertainty. *Fisheries Research* 280 (December): 107161. https://doi.org/10.1016/j.fishres.2024.107161.

Blanco, J.A., and Y.-H. Lo. 2023. Latest trends in modelling forest ecosystems: New approaches or just new methods? *Current Forestry Reports* 9 (4): 219–229. https://doi.org/10.1007/s40725-023-00189-y.

Bouman, R., Z. Bukhsh, and T. Heskes. 2024. Unsupervised anomaly detection algorithms on real-world data: How many do we need? *Journal of Machine Learning Research* 25:1–34.

Bouman, R., and T. Heskes. 2025. "Autoencoders for Anomaly Detection Are Unreliable." ArXiv:2501.13864v1.

Breunig, M. M., H.-P. Kriegel, R. T. Ng, and J. Sander. 2000. "LOF." In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, 93–104. New York, NY, USA: ACM. https://doi.org/10.1145/342009.335388.

Campos, J.C.C., and H.G. Leite. 2017. *Forest measurement: Questions and answers*. 5th Ed. Viçosa: UFV.

Casas, G.G., C.P. Boechat Soares, M.L. de Romarco Oliveira, D.H.B. Binoti, L.P. Fardin, M.M. Coimbra Limeira, Z.H. Ismail, A.A. da Lopes Silva, and H.G. Leite. 2023. Assessment of a monthly data structure for growth and yield projections from early to harvest age in hybrid eucalypt stands. *Pertanika Journal of Tropical Agricultural Science* 46 (4): 1127–1150. https://doi.org/10.47836/pjtas.46.4.04.

Casas, G.G., L.P. Fardin, S. Silva, R.R. de Oliveira Neto, D.H.B. Binoti, R.V. Leite, C.A. Ramos Domiciano, et al. 2022. Improving yield projections from early ages in eucalypt plantations with the clutter model and artificial neural networks. *Pertanika Journal of Science and Technology* 30 (2): 1257–1272. https://doi.org/10.47836/pjst.30.2.22.

Chen, J., H. Du, F. Mao, Z. Huang, C. Chen, M. Hu, and X. Li. 2024. Improving forest age prediction performance using ensemble learning algorithms base on satellite remote sensing data. *Ecological Indicators* 166 (September): 112327. https://doi.org/10.1016/j.ecolind.2024.112327.

Davis, J.J., and A.J. Clark. 2011. Data preprocessing for anomaly based network intrusion detection: A review. *Computers & Security* 30 (6–7): 353–375. https://doi.org/10.1016/j.cose.2011.05.008.

Eitzel, M., J. Battles, R. York, J. Knape, and P. de Valpine. 2013. Estimating tree growth from complex forest monitoring data. *Ecological Applications* 23 (6): 1288–1296. https://doi.org/10.1890/12-0504.1.

Freitas, P. H. De, M. G. Carneiro, T. P. Protasio, D. A. Gonçalves, R. O. V. Miranda, A. A. V. Soares, and L. G. A. Martins. 2024. "Prediction of Managed Forest Growth Based on Machine Learning and Cellular Automata." In *2024 IEEE Congress on Evolutionary Computation (CEC)*, 1–8. IEEE. https://doi.org/10.1109/CEC60901.2024.10611880.

Fu, Y., R. Li, Z. Zhu, Y. Xue, H. Ding, X. Wang, J. Na, and W. Xia. 2024. SCARF: A new algorithm for continuous prediction of biomass dynamics using machine learning and Landsat time series. *Remote Sensing of Environment* 314 (December): 114348. https://doi.org/10.1016/j.rse.2024.114348.

Goodfellow, I., Y. Bengio, and A. Courville. 2016. *Deep Learning*. MIT Press.

Knott, J. A., G. C. Liknes, C. L. Giebink, S. Oh, G. M. Domke, R. E. McRoberts, V. F. Quirino, and B. F. Walters. 2023. Effects of outliers on remote sensing-assisted forest biomass estimation: A case study from the United States National Forest Inventory. *Methods in Ecology and Evolution* 14 (7): 1587–1602. https://doi.org/10.1111/2041-210X.14084.

Liu, B., J. Guo, T. Koike-Akino, Y. Wang, K. J. Kim, K. Parsons, P. Orlik, and J. Yuan. 2021. "Anomaly Detection and Diagnosis Using Pre-Processing and Time-Delay Autoencoder." In *2021 26th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)*, 1–8. IEEE. https://doi.org/10.1109/ETFA45728.2021.9613574.

Liu, F. T., K. M. Ting, and Z.-H. Zhou. 2008. "Isolation Forest." In *2008 Eighth IEEE International Conference on Data Mining*, 413–22. IEEE. https://doi.org/10.1109/ICDM.2008.17.

Mohammed, S., L. Budach, M. Feuerpfeil, N. Ihde, A. Nathansen, N. Noack, H. Patzlaff, F. Naumann, and H. Harmouch. 2024. "The Effects of Data Quality on Machine Learning Performance." ArXiv:2207.14529v5, December. https://doi.org/10.48550/arXiv.2207.14529.

Nassif, A. B., M. A. Talib, Q. Nasir, and F. M. Dakalbab. 2021. Machine learning for anomaly detection: A systematic review. *IEEE Access* 9:78658–78700. https://doi.org/10.1109/ACCESS.2021.3083060.

Rocha, S. J. S. S. da, C. M. M. E. Torres, P. H. Villanova, I. da S. Tavares Júnior, M. P. M. X. Rufino, F. M. B. Romero, L. A. G. Jacovine, et al. 2024. "Machine Learning Methods: Modeling Net Growth in the Atlantic Forest of Brazil." *Ecological Informatics* 81 (July): 102564. https://doi.org/10.1016/j.ecoinf.2024.102564.

Roesch, F. A., and P. C. Van Deusen. 2010. Anomaly detection for analysis of annual inventory data: A quality control approach. *Southern Journal of Applied Forestry* 34 (3): 131–137. https://doi.org/10.1093/sjaf/34.3.131.

Rota, F., D. Scherrer, A. Bergamini, B. Price, L. Walthert, and A. Baltensweiler. 2024. Unravelling the impact of soil data quality on species distribution models of temperate forest woody plants. *Science of the Total Environment* 944 (September): 173719. https://doi.org/10.1016/j.scitotenv.2024.173719.

Schölkopf, B., J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson. 2001. Estimating the support of a high-dimensional distribution. *Neural Computation* 13 (7): 1443–1471. https://doi.org/10.1162/089976601750264965.

Silva, G. C. C., J. C. L. Neves, G. E. Marcatti, C. P. B. Soares, N. Calegario, C. A. Araújo Júnior, D. G. E. Gonzáles, et al. 2024. Use of artificial neural networks with the physiological principles to predict

growth model. *Engineering Applications of Artificial Intelligence* 136 (October): 108914. https://doi.org/10.1016/j.engappai.2024.108914.

Taylor, P., A. C. Almeida, E. Kemmerer, and R. O. de S. Abreu. 2023. Improving spatial predictions of eucalypt plantation growth by combining interpretable machine learning with the 3-PG model. *Frontiers in Forests and Global Change* 6 (July). https://doi.org/10.3389/ffgc.2023.1181049.

Vajda, D.L., T. Van Do, T. Bérczes, and K. Farkas. 2024. Machine learning-based real-time anomaly detection using data pre-processing in the telemetry of server farms. *Scientific Reports* 14 (1): 23288. https://doi.org/10.1038/s41598-024-72982-z.

Weiskittel, A.R., D.W. Hann, J.A. Kershaw Jr., and J.K. Vanclay. 2011. *Forest growth and yield modeling*. John Wiley & Sons.

Zhang, B., G. Liu, Z. Feng, M. Zhang, T. Ma, X. Zhao, Z. Su, and X. Zhang. 2023. Constructing a model of *Poplus* spp. growth rate based on the model fusion and analysis of its growth rate differences and distribution characteristics under different classes of environmental indicators. *Forests* 14 (10): 2073. https://doi.org/10.3390/f14102073.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Gianmarco Goycochea Casas[1] · Zool Hilmi Ismail[2] · Pedro Henrique Fontes dos Santos[1] · Leonardo Ippolito Rodrigues[3] · Tassius Menezes Araújo[1] · Helio Garcia Leite[1]

✉ Gianmarco Goycochea Casas
gianmarco.casas@ufv.br

Zool Hilmi Ismail
zool@utm.my

Pedro Henrique Fontes dos Santos
pedro.fontes@ufv.br

Leonardo Ippolito Rodrigues
leoippef@gmail.com

Tassius Menezes Araújo
tassius.araujo@ufv.br

Helio Garcia Leite
hgleite@ufv.br

[1] Department of Forest Engineering, Federal University of Viçosa, Viçosa, MG 36570-900, Brazil

[2] Center for Artificial Intelligence and Robotics, Universiti Teknologi Malaysia, Jalan Sultan Yahya Petra, 54100 Kuala Lumpur, Malaysia

[3] Forestry Consultant, Bairro Imbaubas, Rua Xenônio, N 49, Ipatinga, MG 35160-284, Brazil