

PREPRINT

Author-formatted, not peer-reviewed document posted on 02/12/2025

DOI: <https://doi.org/10.3897/arphapreprints.e181100>

Forest ecosystem restoration potential unravelled through machine learning algorithms

Jenny Vivian, Robin L. Chazdon,  Alison Shapcott,  David J Lee

Forest ecosystem restoration potential unravelled through machine learning algorithms

Jenny Vivian¹, Robin L. Chazdon¹, Alison Shapcott^{1,2,3}, David J. Lee¹

¹Forest Research Institute, University of Sunshine Coast, Sippy Downs, QLD 4556, Australia

²Centre for Bioinnovation, University of the Sunshine Coast, Sippy Downs, QLD 4556, Australia

³School of Science, Technology and Engineering, University of the Sunshine Coast, Sippy Downs, QLD 4556, Australia

Correspondence:

Jenny Vivian, Tropical Forests and People Research Centre, University of Sunshine Coast, Sippy Downs, QLD, Australia. Email:

Jenny.Vivian@research.usc.edu.au

Abstract

Decades of scientific development have led to a specialisation of environmental analyses, hampering the recognition of forest ecosystem restoration. However, an integrated approach is required to better understand the trajectory of forest recovery. In this study, we aimed to evaluate potential comprehensive approaches for forest restoration assessment by considering 184 environmental features, including geographical information, tree dimensions and biomass, soil abiotic parameters, microclimate, and soil microbial taxa and functions. All these characteristics constituted the dimensions of the ecosystem hypervolume and its emerging properties, related to naturally regenerating *Acacia mangium* plantations. Through unsupervised and supervised machine learning algorithms, we recognised the known time lag between the modification of ecosystem properties and the macroscopic aspects, such as the aboveground biomass. While the sectorial analysis highlighted the recovery of numerous characteristics by the 10th year after plantation establishment (e.g., tree biomass), the analysis integrating numerous ecosystem parameters highlighted the sequential similarity along the chronosequence, with only the 24-year-old plantation approaching recovery as measured in an adjacent remnant forest. The presence of keystone parameters (environmental variables with disproportionate effect on the ecosystem restoration relative to their value or variance) and the

lagged response of ecosystem properties to drivers of environmental changes call for a more comprehensive approach to assess the achievement of restoration goals. Together with recent efforts to include machine learning analyses to answer ecological questions, our study brings unprecedented evidence on how leveraging advances in this analytical field can prompt a better understanding and management of recovering forests, considering multiple intertwined environmental factors.

Keywords: tropical forest recovery, ecosystem properties, machine learning classification, forest ecology, environmental recovery, Machine Learning ecology, ecological hypervolume

Introduction

Forest restoration derives from the variation of multiple environmental parameters, with their monitoring being fundamental to assessing the recovery of the native characteristics or historical functions of the system (Chazdon et al. 2016). While the modification of interacting components is entangled within the forest recovery (de Jong et al. 2021; Toro et al. 2025) and fundamental to evaluating the success of restoration projects over time, long-term monitoring is often neglected (Le et al. 2012, 2014; Vivian et al. 2026). When present, the ecosystem recovery assessment often focuses on a few parameters (e.g., Adamo et al. 2021; Nunez et al. 2021) depending on the project objective, the interest of stakeholders, and the available budget (Caughlan and Oakley 2001).

Ecosystem analyses considering an integrated perspective can improve the understanding of ecosystem dynamics occurring over recovery. Specialisation of the studies (Casadevall and Fang 2014) facilitated improvements in many scientific fields while causing a gap between them. This fracture is recognised, and scientists increasingly advocate for the adoption of more comprehensive approaches in their research (e.g., Hobbs et al. 2011; Malhi et al. 2022; Vivian et al. 2025a). For instance, the “recovery wheel” developed as a tool to

51 assess the ecosystem restoration by comprehending multiple spheres (Gannet et al. 2019),
52 considers the socio-economic parameters of the recovering system along with the ecological
53 ones. Furthermore, making use of technological innovations, Woolnough et al. (2023) highlight
54 the benefits of merging fields by leveraging advances in quantum computing to understand
55 ecological processes. Forest restoration evaluation fits within this context of comprehensive
56 analysis requiring cutting-edge analytical techniques: above- and below-ground components
57 and emerging properties of the synergies within the ecosystem (e.g., functions revealed by the
58 variation of soil microbial communities) should be considered altogether to disentangle their
59 intertwined recovery (Gann et al. 2019).

60 A comprehensive evaluation of forest recovery may unravel the occurring interrelated
61 processes, allowing better management of the restoration efforts. Interests in forest services
62 (e.g., carbon sequestration), the increasing awareness of the importance of soil health (Wall et
63 al. 2015), and of microbial processes occurring in it (Bardgett and van der Putten 2014; Fierer
64 2017) promote their analysis, providing unprecedented insights into forest recovery dynamics
65 (Vivian et al. 2025b, under review). Traditional above-ground biomass assessments (e.g.,
66 Nowak and Crane 2002) may be integrated with the study of soil microbial community
67 evaluations through eDNA (genetic fragments collected from environmental substrates;
68 Thomsen and Willerslev 2015), offering early evidence for ecosystem modification (Bardgett
69 and Van Der Putten 2014; Li et al. 2020). For instance, eDNA allows for the identification of
70 numerous taxa (e.g., Allen et al. 2023; Goralczyk et al. 2025; Ke et al. 2025; López-González
71 et al. 2025; Bessey et al. 2025), such as soil microbes that respond to short- and long-term
72 environmental variations (Graham and Knelman 2023; Lemke and DeSalle 2023). This
73 characteristic makes them valuable indicators for forest ecosystem recovery (Vivian et al.
74 2025a). The consideration of analyses from complementary scientific spheres (e.g., microbial

ecology, plant ecology, botany) is crucial to reestablish the useful integrated approach, and advanced analytical techniques can be leveraged for this scope.

Machine Learning algorithms can improve the understanding of forest restoration by analysing combined ecosystem changes if a sufficient number of high-quality data are available. Multidimensional kernel density estimation has been used to explore the ecosystem modifications and ecological questions (Blonder et al. 2014; Barros et al. 2016), and different machine learning (ML) techniques were used to recognise areas with high natural regeneration potential (Crouizelles et al. 2020; Williams et al. 2024). These recent ML applications lay the ground for their application to study numerous ecological parameters (also referred to as “features”) and large datasets (Jordan et al. 2015; LeCun et al. 2015). Given these characteristics, ML techniques are valuable means to evaluate the complex environmental modifications (Tahmasebi et al. 2020), as those occurring during forest ecosystem restoration. Indeed, forest recovery involves the interplay of several abiotic and biotic factors (Chazdon 2014), composing the dimensions characterising the ecosystem “hypervolume” (Barros et al. 2016) and from which distinct environmental properties emerge (e.g., nutrient cycling).

Advanced computations have the potential to provide objective insights into forest restoration and the variation of ecosystem traits. Unlike the ecological analyses that are data-informed and usually affected by strong assumptions (e.g., Kimmel et al. 2021), ML algorithms are data-driven approaches (Jordan et al. 2015). Therefore, the interpretations of the results can generally directly emerge from patterns and characteristics of the records. Belonging to the ML domain and being assumption-free, the clustering and classification algorithms are used to recognise patterns in the data (Sarker 2021) and divide samples into a pre-defined number of categories (Hastie et al. 2009), respectively, given the traits of the samples. Reflecting the high dimensionality of the hypervolume, ML approaches fit well in a context where complex dynamics interplay, being capable of handling large datasets and numerous variables. Thus, we

used ML algorithms to identify different phases of forest restoration with real-world measurements, recognising the variation of the measurable and emerging properties over time, represented by the integrated consideration of the measured variables.

Clustering techniques are widely implemented in ecology to explore the data, making these algorithms a useful bridge from traditional to innovative analyses with ML methods. For instance, used to recognise the similarity between samples, Principal Component Analysis (PCA) is one of the most common analyses undertaken in ecology and belongs to the realm of ML algorithms (Sarker 2021). While in environmental sciences this technique is used to recognise patterns, in other scientific spheres it is often implemented to reduce the dimensions of the datasets, creating new variables (Sarker et al. 2020a), on which the analysis is applied.

Machine learning classification algorithms can provide insights into the forest restoration trajectory, enhancing the understanding of the modification of environmental traits over time. The classification of samples (also referred to as “instances”) is based on their disposition in a multidimensional space (which makes them easily applicable to the hypervolume concept), according to learned decision boundaries as limits between categories (Hastie et al. 2009). While ML classification does not substitute specialised analyses, it enables the recognition of recovery phases through an integrated perspective, bringing benefits to the management of forest restoration projects.

In our study, we aimed to recognise the efficacy of an integrated methodology to unravel patterns of forest restoration trajectories through different ML clustering and classification algorithms, by considering above-ground and below-ground parameters (e.g. biomass, species, carbon, soil attributes and microbes). Thus, we focused on assessing the benefits of ML approaches to help with the evaluation of forest restoration. The setting of this study was a degraded grassland (representing the initial state of the system), *Acacia mangium* plantations of 2-, 10-, and 24-year-old, and a reference state of remnant forest in the

Philippines. In this context, the modification of the plantations has been scarcely assessed, especially considering multiple above- and below-ground parameters, such as tree community composition, biomass stock, soil nutrients, and microbial community (but see Wei et al. 2024). This case study was adopted to fill the knowledge gap on the benefits of a more integrated approach in forest restoration evaluation, with the additional benefit of providing evidence of ecosystem changes in an understudied context. To unravel the similarity between the oldest naturally regenerating plantings and the remnant forest, we assumed that, if a trajectory of forest restoration is present, and thus samples of increasingly old plantations hold similar values to those of the aimed state, combining the instances from these classes into one category would result in a better classification. Given field observations and previous analyses on our context (see Vivian et al. 2025b, under review; Vivian et al. 2026), we expected an ongoing restoration through time, and we hypothesised a higher classification accuracy when measurements of the late stages of natural regeneration and remnant forest are considered as belonging to the same class compared to that achieved when the two classes are retained. Therefore, to test our expectations and unravel the potentialities of ML classification to understand forest restoration, we addressed the following questions: i) What is the similarity between *A. mangium* plantings and remnant forests over time based on an integrated approach? ii) What can be inferred about forest ecosystem recovery and management, with the simultaneous consideration of multiple parameters, with these reflecting the emerging ecosystem properties? While using data on tropical forest restoration, this unprecedented analysis and results create the background for the implementation of ML in monitoring outcomes of forest restoration, with important consequences for the management of such recovering ecosystems worldwide.

Methods

In our analysis, we considered: geographical information; tree dimensions (and derived biomass, referred to as Above Ground Biomass, “AGB”), species, and abundance; soil nutrients, temperature, humidity, and pH; air temperature and humidity (hereafter referred to as “microclimate”); canopy coverage; soil fungi and bacteria taxonomic and functional information as relative abundances (Table S1). We expected the analysis of 184 environmental parameters through ML algorithms to provide deep insights into the restoration trajectory, by considering the subtle changes usually overlooked in traditional approaches (evaluated in Vivian et al. 2025, under review, and Vivian et al. 2026), and the related modification of the ecosystem hypervolume.

Study site

Soil samples, and data on microclimate, tree dimensions and community composition were gathered in September 2024 on Biliran island, in the central part of the Philippines archipelago. The analysed sites comprise minimally managed *A. mangium* plantations of different ages (2, 10, and 24 years) (Fig. 1A). Due to socio-economic reasons, these *A. mangium* plantations established by local communities were usually left minimally managed after the first three years, with a lack of clearing and occasional uncontrolled harvesting (Gregorio et al. 2015). The lack of intensive management, the common placement of such plantations in proximity to forest remnants (Evans 1992), and the versatile traits of *A. mangium* (e.g., nitrogen-fixation; Krisnawati et al. 2011) led to natural regeneration of the forest (Vivian et al. 2026). As the starting state, we sampled nearby degraded grasslands of *Imperata cylindrica*, and as the reference state, we gathered data in secondary forests, hereafter named “remnant forest”, not directly affected by intensive anthropogenic use, and located in the proximity of the plantations. Altogether, the sampling sites will be referred to as “landcover types”. More details on the

location characteristics are extensively reported in Vivian et al. (2026), and the list of the variables included in this study is reported in the Supplementary Material (Table S1).

Assessment of the above and belowground ecosystem conditions

Data on tree community, soil abiotic parameters, microclimate, and aboveground biomass of standing trees were collected and calculated as thoroughly described in Vivian et al. (2026). Briefly, we sampled four transects of four plots for each of the five landcover types. These transects were ordered along the altitude axis and paired to have a comparable elevation but different light exposure (i.e., different side of the ridge). (Fig. 1A-B). Overall, we sampled a total of 80 plots across the study area, covering a maximum distance of 1800 m and a total elevation difference of 400 m. Soil, air, and microbial data were gathered in each plot (“Sample plot”; measuring 5 x 5 m; Fig. 1B), while the tree height, diameter and community composition were assessed for one larger plot in every transect (the second plot from the lowest elevation, with dimensions of 20 x 20 m; known as the “Measure plot”; Fig. 1B). Values of the vegetational data were copied for the other three plots of the same transect. For this study, we estimated aboveground biomass of trees with diameter at breast height (DBH) ≥ 5 cm using the equation of Chave et al. (2014):

$$AGB = 0.0673 \times (\rho \times D^2 \times H)^{0.976}$$

D=DBH (cm)

H=Height (m)

ρ =Wood density ($\frac{g}{cm^3}$)

The carbon concentration and stock in the 0-10 cm layer of bulk soil were assessed through the protocol of Heanes (1984), considering the soil density of the sites. Available phosphorus was calculated through the Olsen procedure (Olsen 1954), while total nitrogen was defined through the steam distillation method (Bremner et al. 1965).

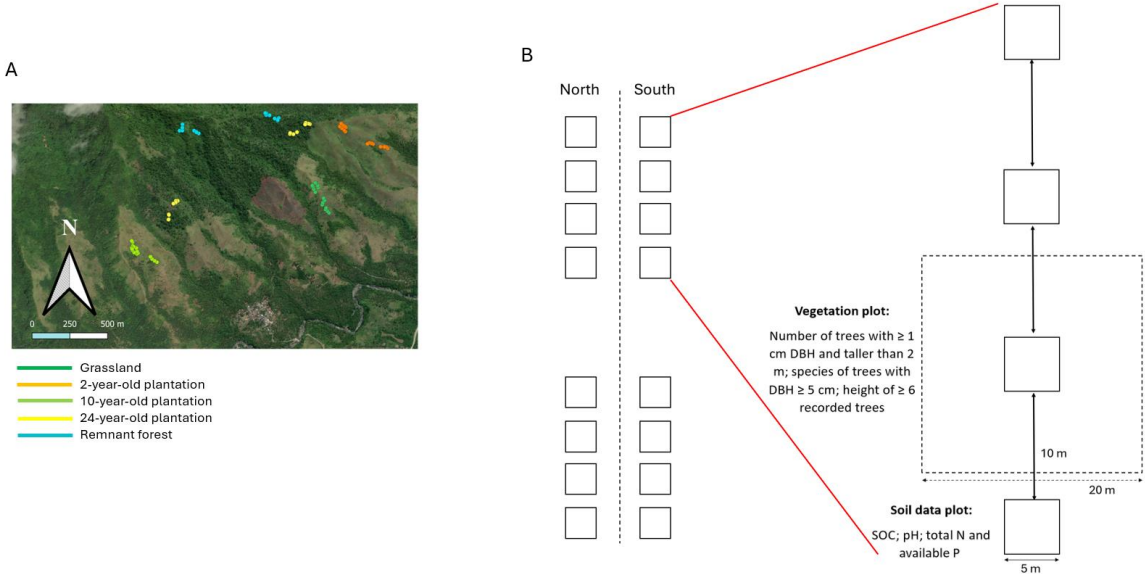


Fig. 1 (A) Study site; sampled plots (coloured points) composing the transects for each landcover type; (B) Set of transects for one landcover type, and details on the disposition of the plots and parameters collected.

Microbial communities

Fungal and bacterial communities of the topsoil (0-10 cm) were studied in all the selected landcovers. In each “soil data” plot, a soil sample was gathered, representing a replicate, totalling 4 per transect and 16 per landcover type. A detailed explanation of the sampling, sequencing, and bioinformatic pipeline is reported in Vivian et al. (2025, under review). The ITS2 rRNA and 16S rRNA gene amplicons were considered for fungi and bacteria DNA sequencing, respectively, through the Illumina NextSeq2000 platform. The sequences obtained were processed as Amplicon Sequence Variants (ASVs) (representing true and unique biological units; Callahan et al. 2017) using the “DADA2” pipeline. The taxonomic profile of the communities was obtained by comparing them with genomes present in UNITE (version 10.0) and SILVA (version 138.2) (Quast et al. 2013) databases for fungi and bacteria, respectively. To assign the functions possibly expressed by the identified microbes, we used the FUNGuild (Nguyen et al. 2016) database for fungi and that of FAPROTAX v1.2.10 (Louca et al. 2016) for bacteria. In the present study, we used the information on the observed richness

of ASVs, and on the relative abundance of each phylum and functional groups for both fungi and bacteria.

Machine Learning Analysis

Seven algorithms of supervised and unsupervised ML were implemented to unravel the temporal stages of forest recovery. These algorithms were selected based on how they handled the data and their complexity (as suggested by Alnuaimi and Albaldawi 2024), allowing a more objective interpretation of the results. Four supervised and three unsupervised ML techniques were employed to facilitate a robust exploration of the data and related ecosystem modifications, capturing general patterns among the clustering and classification methods (Tables S2-S3). By considering diverse learning models, their flaws, and advantages (e.g., Caruana and Niculescu-Mizil 2006; Sarker 2021; Alnuaimi and Albaldawi 2024, and extensively covered in the Supplementary Material section 2.2-2.3, along with details on their functioning), we hoped to gain a more comprehensive understanding of clustering and classification, and the related forest recovery. Data preprocessing and dimensionality reduction methods are reported briefly here, with the detailed explanation of the latter in the Supplementary Material section 2.1. All analyses were conducted in Python 3.12.4 (Python Software Foundation 2024). The protocol related to the adopted workflow can be found at [dx.doi.org/10.17504/protocols.io.6qpvrwj2zlmk/v1](https://doi.org/10.17504/protocols.io.6qpvrwj2zlmk/v1) (*Private link for reviewers: <https://www.protocols.io/private/E914510FCB6411F0B39B0A58A9FEAC02> to be removed before publication.*) Code and datasets are available at (https://github.com/JennyVivian99/ML_Classification_for_forest_restoration).

Data preprocessing and dimensionality reduction

Data preprocessing was conducted before running each of the selected ML algorithms. To consider the above and belowground traits of the ecosystem, we combined information on tree communities, aboveground biomass of standing trees, soil nutrients, temperature, humidity,

and pH, microclimate, and microbial (fungi and bacteria) communities characteristics (Table S1), for a total of 184 variables. Altogether, these data characterised what will be called hereon the “sample” or “instance”, one for each of the 80 plots sampled. For the microbial community characteristics, we retained the ASVs richness and the relative abundance of phyla and functional traits. According to the bioinformatic pipeline (see Vivian et al. 2025b, under review), and due to potentially low-quality results (from their low sequencing depth, affecting the reliability of the taxa identification), two samples gathered from the remnant forest were discarded for fungi communities. For bacteria, one sample from the 10-year-old plantation, two from the 24-year-old plantation, and one from the remnant forest (the same was excluded for the fungi community) were excluded from further analysis for the same reason. Thus, to avoid the inclusion of biased microbial information, we excluded the same samples across all the variables (i.e., above- and below-ground ecosystem traits), resulting in some loss of replication (i.e., from a total of 80 replicates to 75). To allow statistical computations, we numerically mapped the “character” variable of the transect disposition into numerical variables, with 0 representing the South side and 1 the North side of the hills that were sampled. Finally, as reported above, we associated the following vegetation parameters, evaluated for the second and bigger plot in each transect (“Measure plot” in Fig. 1B), to the other three plots (“Sample plot” in Fig. 1B) within the same transect: number of tree species, number of tree individuals, stems per hectare (DBH < 5cm, height > 2m), cumulative basal area (m²/ha), average DBH (cm), cumulative AGB (t/ha), “average height (m), number of native tree species, and number of exotic tree species. While for these variables the replication was reduced to four per landcover type, the integrated analysis was expected not to be affected by such a reduction in measurements, for the persistence of minimum variance and at the landscape level.

After data preprocessing, we standardized the dataset to minimize biases deriving from different feature scales of values. This step is necessary since ML algorithms are mathematical

constructs lacking the understanding of the physical unities of measure of the data (Singh and Singh 2020), and because we employed distance-based (i.e., scale-sensitive) ML techniques as the K-nearest neighbours (KNN) (Fix and Hodges 1951; Cover and Hart 1967). Before the application of unsupervised algorithms (i.e., run with unlabelled data; Hastie et al. 2009), we performed data standardisation simultaneously on the entire dataset. Only for the application of the variance threshold feature selection (explained in Supplementary Material section 2.1), the standardisation was substituted by normalization, to allow variance assessment. For supervised classification (i.e., computations occurring with labelled data; Hastie et al. 2009), data standardization (and normalization when the variance threshold was applied) was performed after splitting the dataset into training and test sets to avoid data leakage.

All the ML techniques were applied with and without dimensionality reduction through variable (also called “feature”) selection and feature extraction (Fig. 2). Dimensionality reduction was conducted as a preprocessing step to lower the dimensions of the dataset while maintaining all the relevant information (Liu and Motoda 1998; Sarker et al. 2021). Decreasing the dimensions is crucial to achieve the best functioning of distance-based ML techniques, as it mitigates the problem of the curse of dimensionality (Bellman 1966), a well-known ML issue. This limitation imposes a higher risk of overfitting, an inability of patterns and clusters distinction, and a subsequent lower efficiency of the ML algorithm as the number of features increases. In our study, the variance threshold (Pedregosa et al. 2011), was implemented to select the features before running unsupervised ML algorithms (i.e., retention of a subset of variables; Sarker 2021). Additionally, feature extraction using principal component analysis (PCA; Hotelling 1933) was employed to generate a new and smaller set of variables from the existing ones, while retaining part of the original information (Sarker 2021). Such a technique was included by taking advantage of the familiarity of the researchers in environmental sciences with PCA. Specifically, we used PCA in this study as the foundational step toward

the usage of classification techniques, which, being all part of the data-driven approaches (Hastie et al. 2009), can improve the understanding of forest restoration. These methods for dimensionality reduction were implemented before running each ML unsupervised algorithm on the entire dataset, since these methods do not require the dataset division for clustering purposes.

Supervised ML algorithms use a random subset called “training set” extracted from the complete dataset. Following the standard approach, dimensionality reduction was first conducted on the training set and not on the whole dataset to avoid data leakage. Then, the selected or extracted features of the training set were also considered for the test set, on which the dimensionality reduction is ultimately run. While the results of the dimensionality reduction are equal among different seeds in unsupervised ML (since the entire dataset is used), in supervised ML, the features selected vary according to the seed set, given its influence on the random selection of the samples for the training set. For supervised ML techniques, we also considered the Pearson correlation (Pearson 1896) threshold, selecting the variables to retain in the analysis (Sarker et al. 2020b). In this case, the correlation was computed by representing each landcover type with a numeric value: 0 for grassland, 1, 2, and 3 for the 2, 10, and 24-year-old *A. mangium* plantations, respectively, and 4 for the remnant forest. Therefore, a negative correlation coefficient indicates a diminution of the value of the feature with the recovery time since plantation establishment, with the opposite applying when a positive correlation is present. Considering the high number of features we aimed to retain, and the composition of our data, we used a $r \geq |0.20|$ value as a weak threshold to retain the variables mostly correlated with the landcover type (Evans 1996). Subsequently, a threshold of $r < |0.70|$ was used to filter out the highly autocorrelated features. Results of the feature extraction and selections are by-products usually implemented as core analysis in traditional studies to unravel patterns in the data. Indeed, the outcomes of the feature extraction and selection (see Tables

S4-S15) highlight the most important parameters for the definition of the ecosystem, making them important to discuss.

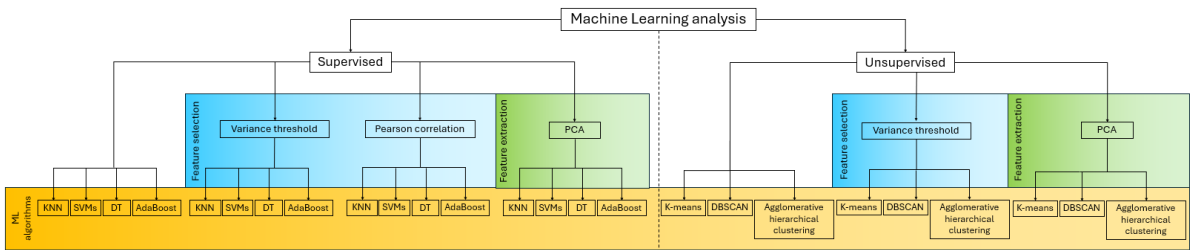


Fig. 2 Machine Learning analysis conducted in our study.

Unsupervised machine learning algorithms

To answer our questions, we first explored the disposition of data into clusters through unsupervised ML algorithms. These unsupervised algorithms are partitioning, density-based and hierarchical-based methods, respectively: K-means (MacQueen 1967; Lloyd 1982; Jain and Dubes 1988), agglomerative hierarchical clustering (AHC; Ward 1963), and density-based spatial clustering of applications with noise (DBSCAN; Ester et al. 1996). K-means analysis was run considering the basic number of clusters (i.e., 2), the real number of categories (5 landcover types), and the optimal number of clusters retrieved through the elbow method (Thorndike 1953; Zhang et al. 2017). The AHC allowed for the construction of a dendrogram through which the clusters can be identified according to different levels of similarity between instances (Ward 1963; Murtagh and Legendre 2014; Nielsen 2016; Sarker 2021), and the DBSCAN analysis is flexible to the number of clusters created, without requiring their number to be set *a priori* (Deng 2020). For the AHC, the average and Ward methods for linking instances were studied (see Supplementary Material section 2.2 for details).

The silhouette score was analysed for each clustering technique. This metric is a coherence value of the clusters, spanning from -1 to 1, with higher values indicating a marked separation between groups, and therefore their better definition (Belyadi and Haghighat 2021). We also created contingency tables and plots, displaying the disposition of the samples within the identified clusters compared to the actual category (i.e., “label”) of the data.

Supervised machine learning algorithms

Through supervised ML algorithms, we aimed to clarify the recovery dynamics of *A. mangium* plantation. We chose four supervised ML algorithms (Fig. 2), based on their assumptions and data-handling characteristics (see Supplementary Materials section 2.3 for details). A multiclass classification analysis was performed using algorithms for managing high-dimensional datasets with a moderate number of samples, as techniques that perform well with a higher number of samples, such as Random Forest or deep neural networks (Krizhevsky et al. 2012; He et al. 2015; Novello and Tonello 2024), would lead to overfitting in our scenario.

Considering the true origin of the samples (known *a priori*), we envisage two possibilities from the classification results. In the first case, ML classification algorithms accurately separate the observations as belonging to different phases of forest regeneration. In the second case, the wrong allocation of observations (low algorithm accuracy) can serve as preliminary evidence of a high degree of similarity between the misclassified instances and the class they were incorrectly assigned to. Further indication of the similarity between confounded samples is the higher accuracy achieved by merging their respective categories (also referred to as “classes”), compared to that obtained when the data are labelled according to their true origin. For instance, the incorrect allocation of samples belonging to late stages of forest recovery in the class of the aimed (reference) state may prove the similarity of the former to the latter, which may be confirmed by the higher accuracy achieved by the algorithm when the two classes are considered as one.

We compared the following supervised ML classification techniques: KNN (Fix and Hodges 1951; Cover and Hart 1967), support vector machines (SVMs; Cortes and Vapnik 1995), decision tree (DT; Quinlan 1993), and adaptive boosting (AdaBoost; Freund et al. 1996) (Fig. 2). All algorithms were run with four different seed values (0, 1, 2, 3), with optimal hyperparameters set under seed 0 and retained equal for the other seeds. The analysis with

multiple seeds allowed enough replicates to calculate the mean and standard deviation for the classification metrics.

The performance of supervised ML algorithms was assessed in terms of accuracy and precision (the additional recall metric is reported in the Supplementary Material section 3). Accuracy can be defined as a measure of the correctly predicted new instances (Teng 2011; Baldi et al. 2020), while model precision averages the number of correctly assigned samples of a specific class on the total samples assigned to the same category (Hastie et al. 2009; Gangadhar and Rangaswamy 2018). For the precision, we calculated its “macro-averaging” value, which assigns an equal weight to each class (Sokolova and Lapalme 2009).

To unravel the similarity between landcover types, we observed the variation of the accuracy and confusion matrices after classification with merged categories, thus by merging the samples belonging to different categories (i.e., landcover types) (Fig. 3).

As previously reported, the idea behind merging landcover types by creating coarser classes is to further validate the difference and similarity between the phases of natural regeneration studied. A higher or lower accuracy of the classification with merged classes compared to the performance considering the original dataset was retained as an indication of their similarity or dissimilarity, respectively. To explore the results by using the original set of features, we implemented such an analysis on the dataset without reduced dimensions (hereafter referred to as “original dataset”). For this purpose, we used the best-performing classification algorithms (based on the accuracy achieved on the classification with unmerged categories), running each of them with 6 additional different random seeds (4, 5, 6, 7, 8, and 9), for a total of 10 runs. Thus, we retrieved the accuracy, precision (and recall, reported in the Supplementary Material section 3) from all the runs with differently merged datasets. Then, we calculated the accuracy difference (and of precision for the results on the late phases of recovery) between the datasets having different merged classes through the Mann-Whitney test (McKnight and Najab 2010).

The difference in the achieved accuracy was also calculated by comparing the results of the best-performing algorithms on the original and merged datasets. To make the comparison independent and considering the different number of runs done for the original and merged datasets, we calculated the statistical difference between accuracy values with the Mann-Whitney test from runs with the 0, 1, 2, 3 and 4, 5, 6, 7 seed values for the original and merged datasets, respectively. Finally, to better understand the possible forest restoration trajectory, we analysed the accuracy and precision obtained for each class considering the best-performing algorithms, adding information on the allocation of the samples into the categories by reporting part of the confusion matrices (i.e., summary table of the correctly and wrongly classified samples).

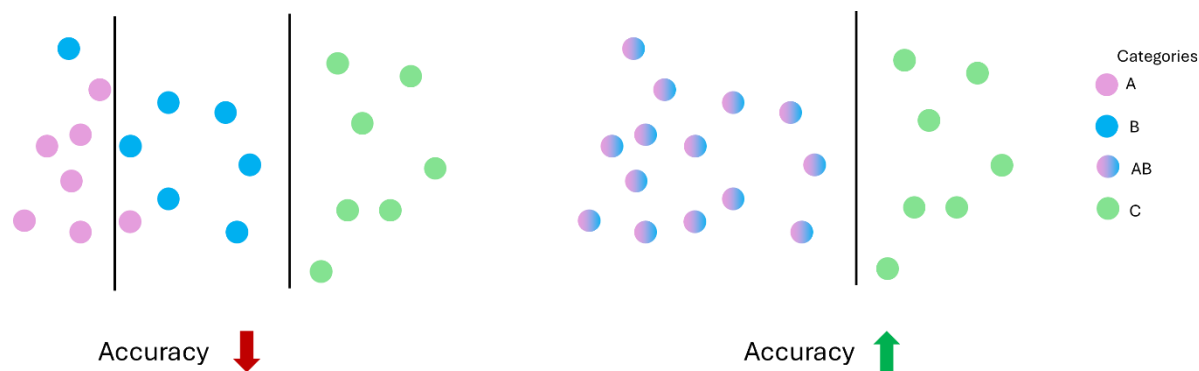


Fig. 3 Illustration of the expected increase in accuracy when merging two similar categories.

Results

Dimensionality reduction

The feature extraction and selection techniques resulted in distinct sets of variables for the unsupervised and supervised ML techniques, as well as for different seeds within the implementation of these latter algorithms. Dimensionality reduction through PCA, Pearson correlation, and variance threshold highlighted the most influential variables for the clustering of the samples. These variables were largely correlated with the landcover type, and with the highest scaled variance, respectively (Table 1).

Considering the entire dataset using unsupervised techniques, the set of ten variables that contributed most (i.e., had the highest “loading”) to the first and second principal components (PCs) was consistent across the seeds used (Table 1). The first PC (PC1) had an average total variance of 13%, and PC2 had 12.4% (the low values derived from the positive and negative influence of all the 184 features). Conversely, for the supervised methodologies, the total variance explained by the first two components was held by different variables, given the differential split of the dataset according to the seed used (Tables S4-S7). Here, PC1 only had two variables (the elevation of the sites and the number of trees present) shared among three of the four seeds (0, 1, and 2) as part of the ten most contributing. For PC2, none of the ten most contributing variables was shared between three or more sets.

Setting the variance threshold to 0.10 as an arbitrary value after scaling the data, distinct features were selected for unsupervised and supervised ML algorithms. Seven variables were retained after the dimensionality reduction of the entire dataset (i.e., for unsupervised ML techniques) (Table 1). The same threshold was applied to supervised ML algorithms to select a set of variables from the training set (then applied to the test set). Among the results for each seed (holding a different number and type of retained variables), we observed six common features: side of the hill in which the plot was assessed, average value of canopy cover, average tree height, average tree DBH, number of tree species, and longitude (Tables 2, S8, S10, S12, S14).

Dimensionality reduction through the Pearson correlation threshold showed only one feature repeatedly selected for different seeds (Tables S9, S11, S13, S15). For this technique, only used in supervised classification, the relative abundance of symbiotic fungi was the unique variable shared by three out of four seeds used, with an average correlation of $r = -0.13 \pm 0.23$ (considering the values of $r \geq |0.20|$ for this category from the analyses with the different seeds). Overall, soil nutrients, fungal and bacterial taxa were the features mostly correlated with the

variation of the landcover type, even considering the filtering of autocorrelated variables ($r < |0.70|$) (Tables S8-S15). Soil nutrients generally increased with the age of the plantation (with older plantations identified as successive phases of natural regeneration), while microbial taxa and functions displayed heterogeneous correlations (e.g., an increase in litter saprotrophic fungi and symbiotrophic fungi, and a decrease in wood saprotrophic fungi and FCPU426 bacteria phylum over the natural regeneration).

Table 1. Variables extracted from different dimensionality reduction techniques.

Unsupervised ML						Supervised ML	
PCA				Variance threshold		Variance threshold	
Variable	PC1 (13.0%) loading	Variable	PC2 (12.4%) loading	Variable	Variance	Variable	Variance
Average densiometer (%)	0.19	Bacteroidota bacteria phylum	0.18	Hill side	0.25	Hill side	0.25±0
Cumulative Basal Area (m ² /ha)	0.17	Candidatus Kryptonia bacteria phylum	0.17	Average densiometer (%)	0.22	Average densiometer (%)	0.23±0.01
Average DBH (cm)	0.17	Fibrobacterota bacteria phylum	0.16	Average tree height (m)	0.12	Average tree height (m)	0.11±0.01
Cumulative AGB (t/ha)	0.17	Anaerobic chemoheterotrophic bacteria	0.15	Average tree DBH (cm)	0.11	Average tree DBH (cm)	0.10±0.01
Average tree height (m)	0.16	Nitrogen-respiring bacteria	0.15	Number of tree species	0.11	Number of tree species	0.12±0.01
Number of trees	0.16	SAR324 bacteria phylum	0.15	Longitude	0.11	Longitude	0.11±0
Pseudomonadota bacteria phylum	0.16	Nitrate-respiring bacteria	0.15	Latitude	0.11	Latitude	0.11±0
Air T (°C)	-0.16	Bacteria ASV number	0.15				
Soil T (°C)	-0.17	NB1-j bacteria phylum	0.15				
Longitude	-0.17	MBNT15 bacteria phylum	0.14				

Variables extracted from PCA dimensionality reduction, with their relative contribution (loading) to PC1 and PC2 before implementation of unsupervised ML techniques, and variables selected through a variance threshold of 0.10 for unsupervised and supervised ML algorithms. For this latest case, the average and standard deviation of the variance of the common features selected (through four different seeds: 0, 1, 2, 3) were retained. DBH: diameter at breast height; AGB: above-ground biomass (t/ha); T: temperature (°C); Longitude and Latitude were in decimal degrees.

Clustering

The most coherent grouping among the clustering techniques was represented by two and three clusters, displaying the separation between early, late phases of natural regeneration, and the reference state of the remnant forest (Table 2 and Fig. 4). K-means, AHC, and DBSCAN

techniques without dimensionality reduction of the dataset generally resulted in poorly coherent groups. In this case, the silhouette value for the DBSCAN algorithm was higher than that of K-Means and AHC (0.35 ± 0 , 0.14 ± 0.09 , 0.30 ± 0.19 , respectively). Conversely, K-Means outcompeted the other techniques in terms of average silhouette value, when the clustering was computed on the dataset with reduced dimensions through PCA (0.62 ± 0.05 for K-means, 0.12 ± 0 for DBSCAN, and 0.61 ± 0.05 for AHC). K-Means held a higher silhouette value compared to the other techniques, also after feature selection with variance threshold analysis (0.53 ± 0.06 for K-means, -0.06 ± 0 for DBSCAN, and 0.48 ± 0 for AHC). The highest coherence of the samples considering feature extraction through PCA (Table 2, Fig. 4) derived from the K-Means distinguishing 3 clusters (obtained through the elbow method) representing early, late phases of natural regeneration and remnant forest, and from AHC, with 2 clusters (identified from dendrogram analysis) separating the plantations from the remnant forest. All algorithms highlighted the persistent diversity between plantations and remnant forests, with some techniques showing also the differences between younger and older plantations (Table 2 and Fig. 4).

Table 2. Average silhouette coefficient and related standard deviation of the different clustering techniques considered.

Original	K-Means		Agglomerative Hierarchical Clustering			DBSCAN	
	2 clusters	0.20±0.11	Average linkage	2 clusters	0.48±0	1 cluster	0.35±0
PCA	5 clusters	0.08±0.04	Ward linkage	2 clusters	0.12±0	2 clusters	0.12±0
	3 clusters	0.13±0.07					
	5 clusters	0.58±0.01	Average linkage	2 clusters	0.65±0	8 clusters	-0.06±0
VT	3 clusters	0.66±0	Ward linkage	2 clusters	0.57±0		
	5 clusters	0.59±0.01	Average linkage	2 clusters	0.48±0	8 clusters	-0.06±0
	2 clusters	0.48±0	Ward linkage	2 clusters	0.48±0		

Average silhouette coefficient and related standard deviation of the different clustering techniques considered, for each of the datasets analysed, and derived from the usage of four seeds to include variability and statistical significance, while controlling for randomness. Original: dataset without reduced dimensions; PCA: dataset with dimensions reduced through Principal Component Analysis; VT: dataset with dimensions reduced through Variance Threshold. Highlighted in italics are the highest silhouette values and the relative number of clusters.

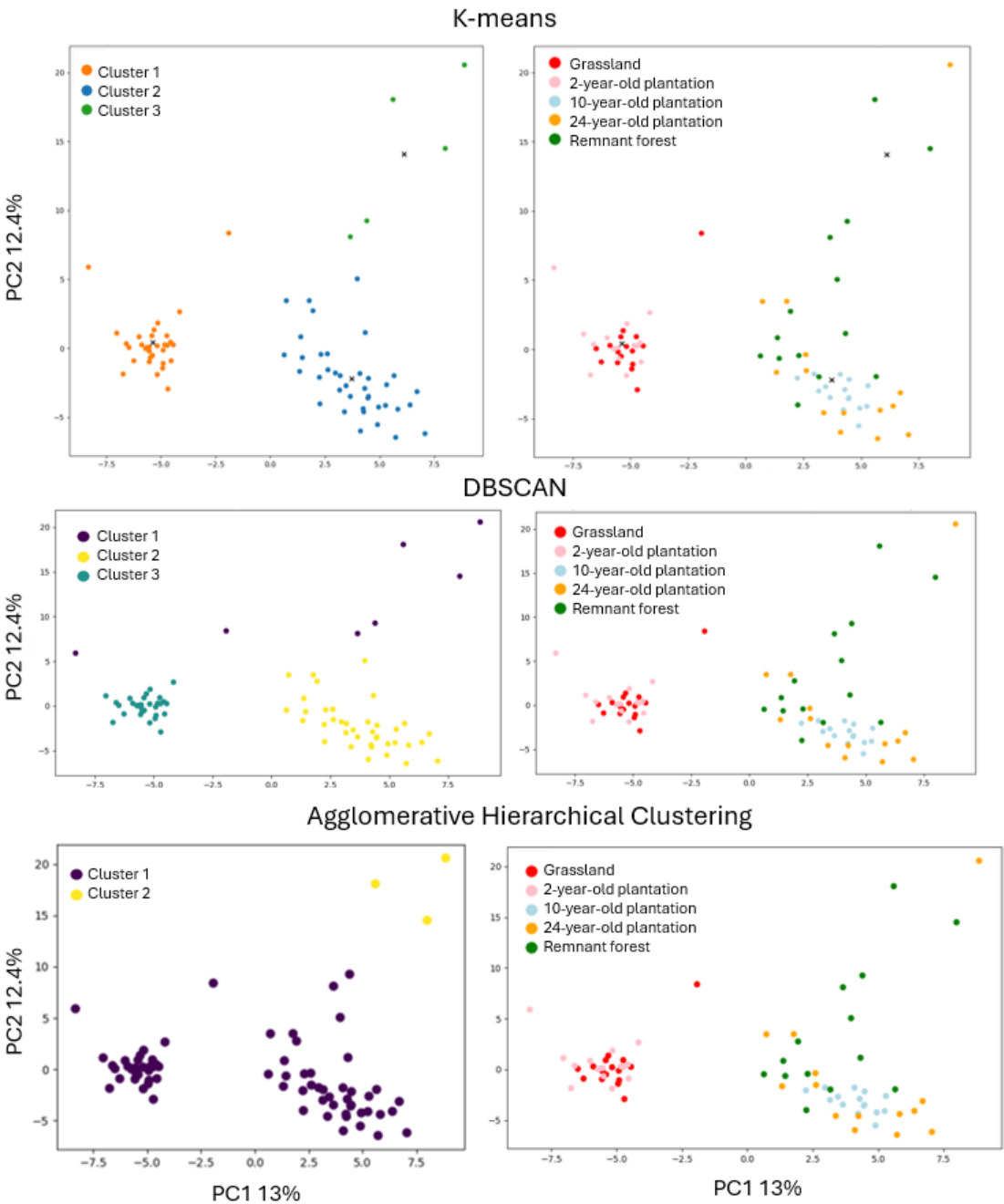


Fig. 4 Unsupervised classification with three different machine learning algorithms, considering the dataset with reduced dimensions through PCA, and seed 0.

Classification

Different techniques achieved perfect classification of the samples, with the original set of variables, or only the features holding the highest variance (Table 3). While KNN never achieved the test maximum accuracy and precision, SVMs correctly classified all the instances when only features with the highest variance (> 0.10) were used. This result displays the

possible confounding effect of considering numerous variables altogether to recognise the true origin of the samples. Still, the DT and AdaBoost algorithms achieved the maximum test accuracy with the original dataset, or, as SVMs, by computing the classification using the most variable features, making these techniques not affected by the high dimensionality for the allocation of the samples to their corresponding landcover type. The perfect classification indicated the evident distinction between landcover types, and therefore phases of natural regeneration, while still leaving the possibility for a trajectory of forest recovery to be observed with the merged categories classification. The additional recall metric values are reported in the Supplementary Material (Table S16) for all the supervised ML techniques implemented.

Table 3. Average accuracy and precision, and related standard deviation of the different classification algorithms considered (from seeds 0, 1, 2, and 3), for each of the datasets analysed.

		KNN	SVMs	DT	AdaBoost	Average
Original	Accuracy	0.65±0.14	0.78±0.04	1±0	1±0	0.86±0.17 (A)
	Precision	0.60±0.16	0.78±0.05	1±0	1±0	
PCA	Accuracy	0.53±0.12	0.52±0.03	0.52±0.06	0.52±0.06	0.52±0.01 (C)
	Precision	0.57±0.22	0.32±0.04	0.47±0.12	0.47±0.12	
VT	Accuracy	0.90±0.07	1±0	1±0	1±0	0.98±0.05 (A)
	Precision	0.90±0.13	1±0	1±0	1±0	
P	Accuracy	0.60±0.05	0.60±0.12	0.71±0.29	0.75±0.11	0.67±0.08 (B)
	Precision	0.49±0.12	0.61±0.14	0.66±0.29	0.75±0.12	

SVMs: Support Vector Machines; DT: Decision tree; AdaBoost: Adaptive boosting; Original: dataset without reduced dimensions; PCA: dataset with dimensions reduced through Principal Component Analysis; VT: dataset with dimensions reduced through Variance Threshold; P: dataset with dimensions reduced through Pearson correlation threshold. Capital letters indicate the significant difference between accuracy values using different dataset.

Classification with merged landcover types

Classification with merged categories allowed for the recognition of the similarity between successive phases of natural regeneration and their relative reference states. To classify samples considering merged classes, we used the best-performing algorithms according to the accuracy achieved with unmerged classes: DT and AdaBoost (Table 4). A perfect classification of the samples was achieved by both algorithms when merging the samples belonging to the grassland

and 2-year-old plantation (G-2), then to the 2-year-old plantation with those of the 10-year-old plantation (2-10), and by collapsing into one category the instances of this latest landcover type with that of the 24-year-old plantation (10-24). With the AdaBoost algorithm, a higher accuracy was achieved by grouping the observations belonging to the remnant forest with those of the “24-year-old plantation” class (24-R), compared to merging the instances of the goal reference state with the 10-year-old plantation (10-R) ones (Table 4), exposing their similar values.

Table 4. Average values and standard deviation of the accuracy and precision values achieved by decision tree (DT) and Adaptive Boosting (AdaBoost) algorithms, considering seven datasets having differently merged categories.

	Decision Tree						
	G-R	2-R	G-2	2-10	10-24	10-R	24-R
Accuracy	0.95±0.07 (A)	0.87±0.09 (B)	1±0 (A)	1±0	1±0	0.91±0.09 (a)	0.93±0.16 (a)
Precision	0.93±0.12	0.83±0.16	1±0	1±0	1±0	0.89±0.14	0.92±0.19
	AdaBoost						
	G-R	2-R	G-2	2-10	10-24	10-R	24-R
Accuracy	1±0 (AA)	0.94±0.06 (BB)	1±0 (AA)	1±0	1±0	0.91±0.11 (aa)	0.99±0.03 (bb)
Precision	1±0	0.93±0.11	1±0	1±0	1±0	0.93±0.10	0.99±0.02

G-R: dataset with samples of the grassland and the remnant forest merged; 2-R: dataset with instances of the 2-year-old plantation and remnant forest considered as part pf the same category; G-2: dataset with grassland and 2-year-old samples as part of the same class; 2-10: dataset with 2- and 10-year-old plantations instances collapsed into a unique category; 10-24: dataset in which the samples belonging to the 10- and 24-year-old plantations were labelled similarly; 10-R: dataset with samples of the remnant forest and 10-year-old plantations labelled equally; 24-R: dataset with samples of the 24-year-old plantation and remnant forest labelled as belonging to the same class. Statistically significant differences (Mann-Whitney test; $p < 0.05$) in accuracy are displayed, considering, for DT results, single and double capital letters for distinct pair comparisons. The same was applied to highlight accuracy differences between AdaBoost results, adopting single and double lowercase letters for distinct pairwise comparisons.

Merging the classes of “Grassland” with “Remnant forest” (G-R), and “2-year-old plantation” with “Remnant forest” (2-R) resulted in a lower accuracy compared to the original dataset using DT, but not AdaBoost, highlighting the classification power of this latest

technique, through the distinction of common patterns among the high-dimensional data. For both DT and AdaBoost, merging the samples of the “10-year-old plantation” class, or those of the “24-year-old plantation” with the instances belonging to the remnant forest, did not bring different values of accuracy compared to the original dataset, where the samples of these classes are separated in each respective category (Mann-Whitney test; $p > 0.05$).

To better disentangle the similarity between the 10-, 24-year-old plantations and remnant forest (through classes merging), we analysed the accuracy and precision obtained by DT and AdaBoost for each class (Table 5) and observed the samples allocation (i.e., confusion matrix; Table S18). Merging the 24-year-old plantation with remnant forest resulted in almost a perfect allocation of the grouped observations, while collapsing samples of the 10-year-old plantation with those of the reference state induced misclassifications. For the DT algorithm, significantly lower accuracy was achieved with the 10-R dataset compared to that calculated using the 24-R dataset. The lower score was caused by the misclassification of the samples belonging to the merged classes (10-year-old plantation and remnant forest) and to the 24-year-old plantation. In the case of the 24-R dataset, a low average accuracy was triggered by the misclassification of samples belonging to the 2-year-old plantation, allocated into the 10-year-old class, highlighting their similarity. The AdaBoost algorithm wrongly classified samples of the 10-R merged class in the 24-year-old plantation category and vice-versa, and of the 2-year-old plantation in the 10-year-old plantings when the 24-R classes were united. Recall metric scores are reported in the Supplementary Material (Table S19) for all the explored scenarios.

Table 5. Accuracy and precision values for each class, achieved by decision tree (DT) and Adaptive Boosting (AdaBoost) algorithms, considering the two different datasets.

	Decision Tree			
	Grassland	2-year-old	10-year-old	24-R
Accuracy	1±0	0.90±0.32	1±0	1±0 (A)
Precision	0.88±0.26	0.80±0.42	1±0	1±0 (a)
	Grassland	2-year-old	10-R	24-year-old

Accuracy	1±0	1±0	0.91±0.14 (B)	0.68±0.43
Precision	1±0	1±0	0.85±0.20 (b)	0.72±0.42
	AdaBoost			
	Grassland	2-year-old	10-year-old	24-R
Accuracy	1±0	0.98±0.08	0.97±0.11	1±0 (A)
Precision	1±0	0.98±0.08	0.98±0.06	1±0 (a)
	Grassland	2-year-old	10-R	24-year-old
Accuracy	1±0	1±0	0.93±0.12 (B)	0.79±0.27
Precision	0.95±0.16	1±0	0.87±0.17 (b)	0.91±0.15

Datasets considered: the “10-R”, in which the samples of the remnant forest and 10-year-old plantations are merged in the same class; the “24-R”, where samples of the 24-year-old plantation and remnant forest are combined in the same category. Capital letters highlight the statistical difference (Mann-Whitney test; $p < 0.05$) between precision values of the merged classes of 10, 24-year-old plantation and remnant forest, while lowercase letters highlight the statistical difference (Mann-Whitney test; $p < 0.05$) of precision for the same combination of comparisons between classes.

Discussion

Discrepancy between ecosystem properties and macro-aspects restoration is unravelled through ML clustering and classification

Naturally regenerating *A. mangium* plantings increasingly resembled the remnant forest, without fully recovering. Unsupervised ML techniques are important to recognise patterns in the data without biases related to *a priori* knowledge of their true origin (Hastie et al. 2009). In our study, these algorithms highlighted how the ecosystem characteristics of the 10 and 24-year-old plantations can appear similar to those of the remnant forest while remaining differentiated from it (i.e., with instances belonging to different clusters), emphasising the transitional nature of these intermediate phases of natural regeneration. Through AHC, the highest distinction among landcover types occurred between the earliest years of *A. mangium* growth (0-2 years) and the peak of its dominant abundance in the ecosystem (around the 10th year; Krisnawati et al. 2011). This latter phase corresponds to a closer canopy compared to the early plantation development, which persists after *A. mangium* decay (30th-50th year, Guzman

et al. 1997; Krisnawati et al. 2011), for the presence of other tree species entering the ecosystem (Vivian et al. 2026). Therefore, this algorithm seems to have captured the macroscopic aspect of the system variation as an indication of their similarity. Conversely, K-means and DBSCAN techniques distinguished the plantings from the remnant forest, highlighting the unique composition of characteristics of the reference state. In this context, the combination of multiple biotic and abiotic environmental parameters (e.g., microclimate, tree community diversity) may have imposed a higher influence on the results compared to the similar macroscopic aspect of the canopy coverage. Overall, we argue that being sensitive to outliers (Balcan et al. 2014), the AHC algorithm may have failed to unravel the discrepancies between similar yet distinct landcover types compared to K-means and DBSCAN.

The results from unsupervised algorithms and supervised classification emphasise the known lag between the evident change of ecosystem properties and that of macroscopic parameters (Fig. 5A). The unsupervised technique of AHC, being sensitive to outliers and thus being affected by their value for the disposition of the samples in the multidimensional space, arranged the instances into 2 clusters. In this case, samples of the old plantations were clustered together with those of the remnant forest, suggesting their similarity. However, the less outlier-biased K-Means and DBSCAN algorithms showed the best performance with 3 groups, highlighting the persisting differences between the monocultures and the remnant forest, even for the 24-year-old planting, which is consistent with the restoration trajectory expectations. The supervised techniques further stressed the outcomes of the unsupervised algorithms. Precisely, the high accuracy achieved by merging subsequent categories (e.g., grassland and 2-year-old plantation, 2- and 10-year-old plantations) unravelled the similarity of the samples of the collapsed categories, which also showed the delay in the recovery of the ecosystem properties (represented by the consideration of the integrated variables analysis) (Odum and Barret 1971; Mitchell et al. 2023). Thus, the ecosystem state at different phases of modification

(in a trajectory of forest regeneration) is similar to the state it follows and precedes, despite how the environmental components vary at a different pace (Fig. 5B). Underscoring the need for a comprehensive method, the reliance on a sectorial analysis of only a few ecosystem characteristics can hinder the recognition of broader patterns and dynamics (Levins 1992; Loehle 2004).

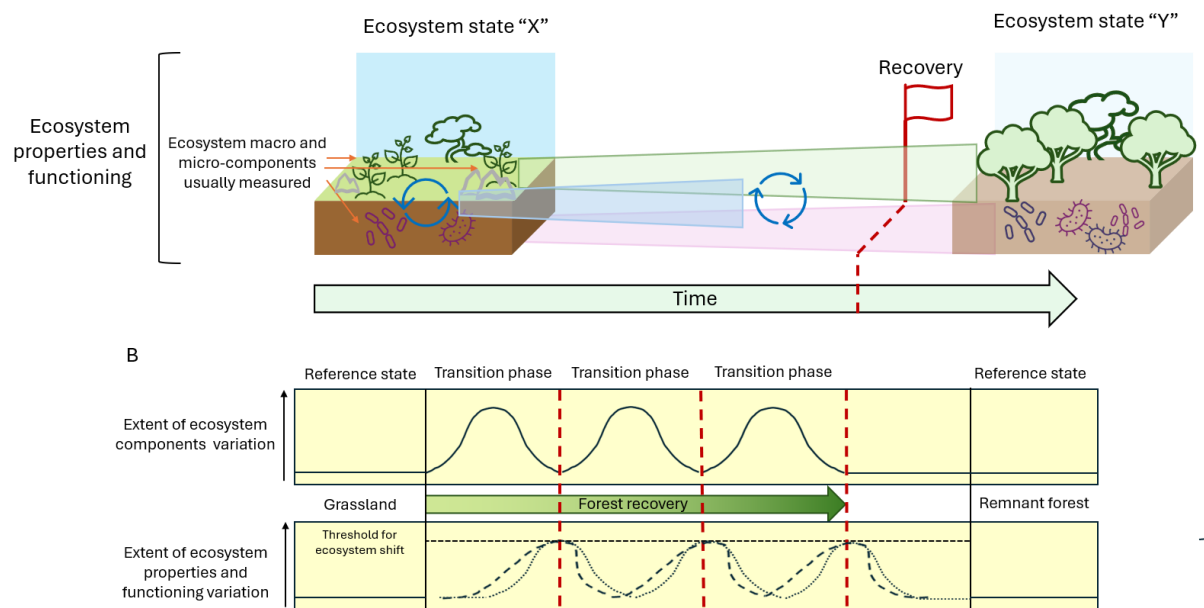


Fig. 5 (A) Representation of the ecosystem components and properties (circular arrows) variation over forest recovery. Diagram in (B) explains the differential time required for the variation of the components and properties.

The integrated analysis seems to confirm the different time scales of recovery for ecosystem properties and components (e.g., Odum and Barret 1971; Mitchell et al. 2023), improving the understanding of forest restoration. For instance, in our context, AGB certainly recovers from the 10th year of the plantation establishment (Vivian et al. 2026). According to such a result, the accuracy achieved with the classification of the samples of the 2- and 10-year-old plantation merged should have demonstrated a low value, especially compared to the performance of the algorithms with these originally separated classes, but this was not the case. The accuracy achieved was the highest possible, showing the similarity of the two categories when numerous ecosystem characteristics, even at the “fine scale”, are considered (e.g., the relative abundance of bacteria and fungi taxa). Examples like this underscore the need for a

comprehensive method to better assess the restoration trajectory and show that relying on detailed studies of only a few ecosystem characteristics (e.g. AGB, tree community) can hinder the recognition of broader patterns and dynamics (Levins 1992; Loehle 2004).

The outcomes of our study are consistent with previous research on ecosystem shifts along with their resistance to change, stability and resilience, which are related to their abiotic and biotic characteristics (Holling 1973; Tilman and Downing 1994; Scheffer et al. 2001). Scheffer et al. (2001) explain the possibility of different ecosystem responses to multiple disturbances and parameter changes, with a shift toward an alternative state occurring at variable times according to the resilience of the system, type, and extent of modifications. Such results are reflected in our study, considering the modification over time of plantings of *A. mangium*, a fast-growing tree (Krisnawati et al. 2011). We speculate that this biological characteristic of the planted species, together with the climatic and anthropogenic disturbances (e.g. typhoons, fires, and occasional harvesting), can prompt steady changes to the macroscopic aspects of the environment, with the ecosystem properties varying at a different pace. This dynamic of ecosystem components and properties change is extensively studied (e.g., Attiwill and Adams 1993; Pastor and Walker 2006; Watts et al. 2020; Pigani et al. 2022) and is thought to lead to delayed ecosystem shifts (e.g., Scheffer et al. 2001; Adrian et al. 2012; Hughes et al. 2013). For example, the development of a dense and closed canopy (i.e., steady variation of a macroscopic aspect) induces a lower mean air temperature and reduced light on the undergrowth, affecting the composition of the bacterial community. Consequently, nutrient cycling and decomposition rates shift, but these alterations become apparent only later, as it takes time for significant changes to accumulate and prompt the modification of these ecosystem properties. Because of dynamics, we emphasise the need to consider the different reaction times between ecosystem components and properties to environmental variations (Moreno-Mateos et al. 2020; Watts et al. 2020) in forest ecosystem monitoring.

When numerous variables are considered, which can be viewed as the multiple dimensions of a hypervolume, the study computes changes for the ecosystem components and linked properties rather than observing the modification of each feature independently.

Our results stress the discrepancy between the response time of the ecosystem hypervolume, represented by the integrated analysis of numerous variables, and the related properties compared to macro-scale environmental variables. Observing the values of environmental characteristics, it is evident that many of them recover by 10 or 24 years (e.g., soil nutrients, canopy coverage, biomass of trees and tree community composition (Fig. 5A; see Vivian et al. 2025b, under review; Vivian et al. 2026). Still, the total accuracy obtained for differently merged samples unravelled the comparability between the grassland and 2-year-old plantation, the 2- and 10-year-old plantations, highlighting how the ecosystem is still recovering. Furthermore, the wrong class attribution occurring when samples of the 24-year-old plantation and remnant forest are labelled equally was related to the 2-, 10-year-old plantation categories (i.e., not the united class), emphasising the common characteristics among these two landcovers.

Integrated approach in tropical forest restoration

Our results suggest that a general approach is beneficial to recognise distinct phases of restoration when the considered parameters are highly variable. Differently from KNN, which is instance-based (Fix and Hodges 1951; Cover and Hart 1967), and thus rely on similarities to the training data to classify the unknown cases (Mitchell 1997 Section 8.6), the DT, AdaBoost, and SVMs algorithms are model-based (Quinlan 1993; Cortes and Vapnik 1995; Freund et al. 1996). During the training phase, these later algorithms construct a general model through which new instances are classified (Mitchell 1997, Section 8.6). Our results suggest that allocating new samples based on their similarity to training data (i.e., through KNN) is less effective than classifying them using general patterns (i.e., with DT, AdaBoost, and SVMs).

We argue that the high variability of each feature may stand at the core of such a result, emphasising the utility of an integrated, model-based approach to unravel patterns of ecosystem forest restoration, considering multiple characteristics and their trends rather than their absolute value.

Interestingly, the high accuracy achieved by AdaBoost when samples belonging to the grassland and remnant forest are allocated to the same category emphasises underlying common characteristics between the two landcover types and the classification power of this algorithm compared to DT. This result may derive from the iterations and complex structure of AdaBoost, considering multiple decision trees, which allows for a refined analysis of the subtle dissimilarities and commonalities between the samples (Gallatin and Albon 2023; Hussain and Zaidi 2024). This characteristic may have also influenced the results of AdaBoost in our study, which recognised a higher similarity for grassland than for the 2-year-old plantation when compared to the remnant forest. In this case, the underlying similarities could be linked to particular microbial taxa and functional groups (as reported in Vivian et al. 2025b, under review) rather than “macro-scale” features defining the diverse ecosystem structure (e.g., tree canopy). Nevertheless, the similarity of grassland to the 2-year-old plantation was also revealed, given the high accuracy achieved by both DT and AdaBoost when these classes are merged, sustaining the expected and previously assessed natural regeneration occurring in the studied landcover types (Vivian et al. 2025b, under review; Vivian et al. 2026).

Sustaining our hypothesis on the benefits of undertaking an integrated approach, the results suggest the analysis of all the available environmental parameters for a better recognition of the forest ecosystem regeneration. With supervised ML algorithms, the highest accuracy was achieved considering all the features or, in all but one case, the most variable ones. The maximum accuracy achieved by DT and AdaBoost through the usage of the parameters with high variance suggests their role as “macro” characteristics to assess in each

monitoring effort (e.g., geographic characteristics and ecosystem structure). In our case, the variables with the highest variance are ecosystem characteristics commonly assessed in forest restoration (e.g., tree richness, DBH), but other parameters may have played a fundamental role in ecosystem restoration despite their variance. In addition, such traits may not express a large variation in other contexts, making the usage of all features preferable to avoid misleading inferences. For example, when stages of forest restoration are assessed in close succession, tree species richness or canopy coverage may not vary enough to inform on the recovery phase of the system. In those cases, bacterial and fungal communities may be the most distinguishable features with their sensitivity to small ecosystem changes (Yan et al. 2018). Thus, while in our case the environmental “macro” characteristics were enough to classify the samples in three out of four algorithms, we argue that the consideration of multiple “micro-scale” parameters would improve the understanding of the restoration trajectory, especially in the case of unclear distinction between ecosystemic changes, or monitoring efforts timed close to each other. In practice, the environmental assessment can mirror this process by undertaking a sequential distinction between categories through macro-aspects, small-scale properties and, when these parameters are taken altogether, ecosystem properties.

We believe that the discrepancy between the ecosystem response and that of many parameters (e.g., tree biomass, soil nutrients) is caused by the influence of micro-scale characteristics, which may respond slowly to environmental changes and are often overlooked. These aspects, despite their abundance or variance, can be highly influential for the restoration of the ecosystem. Comparable to keystone species (species with disproportionate effect relative to their abundance; Paine 1969), we define such ecosystem characteristics as “keystone parameters”. Examples may be bacterial taxa and functional groups, which, despite their low relative abundance value (e.g., under 1% for the Candidatus Kryptonia and SAR324 bacteria

phylum), are particularly important for their contribution in PCs, thus for the definition of data structure (Jolliffe 2002; Peres-Neto et al. 2003; Abdi and Williams 2010).

The presence of keystone parameters and the related discrepancy in the time response of ecosystem hypervolume to environmental changes corroborates the expected benefits of an integrated approach. Assessing only the macro-aspects of the system, it may be speculated the realisation of the restoration, while ecosystem properties could still be adjusting to the variation. Therefore, we recommend a conservation effort involving a lack of intensive anthropogenic disturbance, or careful management to ensure enough time for the ecosystem to adapt to the new conditions. For example, when forest restoration seems to be achieved from a few macro-aspects characterisation, the conservation of the area without extensive harvesting should be promoted for a certain period, ensuring the effective transition of the ecosystem properties (even unassessed ones) to the aimed state. Such management would result in avoided delays in ecosystem recovery or unexpected variations (e.g., return to degraded state) due to the fragility of the system in this transition phase (e.g., Scheffer et al. 2001; Folke et al. 2004).

Important for the recognition of different landcover types studied were soil microbial communities, canopy coverage, and geographical information (exposure and location coordinates). Specifically, the number of bacteria ASVs and functional groups commonly composed the first two principal components. In addition, the symbiotic fungi group was repeatedly highly influential in the dimensionality reduction through Pearson correlation, with a decreasing relationship with the landcover type over the natural regeneration. Such a result is particularly important considering the statistical similarity of the relative abundances of this functional group in old plantations and remnant forests (see Vivian et al. 2025b, under review). Given these results, we ML classification allowed in our case the recognition of a distinct trait of the remnant forest compared to old plantations, expressed by a modification of the interactions between plants and symbiotic fungi, but not by the simple evaluation of the group

relative abundance between landcover types. Besides the variability of the hillside being an artifact of the value representing the North and South side of the plot disposition (1 and 0, respectively), the location coordinates, canopy coverage (represented by average densiometer percentage) and tree dimensions were usually the most variable parameters, indicating their role as “macro-scale” indicators for the distinction of phases of natural regeneration. As discussed, while these features are particularly important in the definition of the landcover type, the observation of small-scale traits may be fundamental to recognising restoration phases when more evident ecosystem structural or macroscopic characteristics are similar, or when they deviate from the true recognition of inherently similar landcover types. As reported by the example of the fungi symbiotic group, the importance of small-scale variables is unravelled by the highest Pearson correlation with landcover types displayed by numerous microbial taxa, functional groups, and soil nutrients.

Conclusions

Unsupervised and supervised ML algorithms allowed for the recognition of a forest restoration trajectory yet to be completed, starting from naturally regenerating *A. mangium* plantations. Such transitions occurred with an ecosystem (hypervolume) shift lagging compared to the changes of environmental macroscopic and usually evaluated components (e.g., AGB, tree community). The sectoral evaluation of environmental aspects may induce the premature declaration of forest restoration, with environmental properties still having to fully recover, and ultimately hampering the achievement of the aimed state. The analysis through ML unsupervised and supervised algorithms allowed for the recognition of the discrepancy between the recovery time of the ecosystem components and properties, adding a new fundamental perspective and analytic tool, the avoidance of unintended consequences from anthropogenic interventions, and promoting a better management of recovering forests.

Funding

This research was undertaken with financial support provided by the University of the Sunshine Coast.

Conflict of interest

Research results and analyses presented in this manuscript are independent from the reforestation Project funded activities. The Project funder had no influence over the design of the research and any of the research findings. The Project funder does not necessarily agree with the analysis presented nor with any of the recommendations made because of the research.

Acknowledgements

The authors are thankful to Nicola Novello for the feedback and suggestions provided on the Machine Learning analysis.

References

- Abdi H, Williams LJ (2010) Principal component analysis. Wiley Interdisciplinary Reviews: Computational Statistics 2(4): 433–459. <https://doi.org/10.1002/wics.101>
- Adamo I, Ortiz-Malavasi E, Chazdon R, Chaverri P, Ter Steege H, Geml J (2021) Soil fungal community composition correlates with site-specific abiotic factors, tree community structure, and forest age in regenerating tropical rainforests. Biology 10(11): 1120. <https://doi.org/10.3390/biology10111120>

- 769 Adrian R, Gerten D, Huber V, Wagner C, Schmidt SR (2012) Windows of change: Temporal
770 scale of analysis is decisive to detect ecosystem responses to climate change. Marine Biology
771 159: 2533–2542. <https://doi.org/10.1007/s00227-012-1938-1>
- 772 Allen MC, Kwait R, Vastano A, Kisurin A, Zoccolo I, Jaffe BD, Angle JC, Maslo B, Lockwood
773 JL (2023) Sampling environmental DNA from trees and soil to detect cryptic arboreal
774 mammals. Scientific Reports 13(1). <https://doi.org/10.1038/s41598-023-27512-8>
- 775 Alnuaimi AF, Albaldawi TH (2024) An overview of machine learning classification
776 techniques. BIO Web of Conferences 97: 00133. <https://doi.org/10.1051/bioconf/20249700133>
- 777 Attiwill PM, Adams MA (1993) Nutrient cycling in forests. New Phytologist 124(4): 561–582.
778 <https://doi.org/10.1111/j.1469-8137.1993.tb03847.x>
- 779 Balcan MF, Liang Y, Gupta P (2014) Robust hierarchical clustering. The Journal of Machine
780 Learning Research 15(1): 3831–3871.
- 781 Baldi P, Brunak S, Chauvin Y, Andersen CA, Nielsen H (2000) Assessing the accuracy of
782 prediction algorithms for classification: An overview. Bioinformatics 16(5): 412–424.
783 <https://doi.org/10.1093/bioinformatics/16.5.412>
- 784 Bardgett RD, Van Der Putten WH (2014) Belowground biodiversity and ecosystem
785 functioning. Nature 515(7528): 505–511. <https://doi.org/10.1038/nature13855>
- 786 Barros C, Thuiller W, Georges D, Boulangeat I, Münkemüller T (2016) N-dimensional
787 hypervolumes to study stability of complex ecosystems. Ecology Letters 19(7): 729–742.
788 <https://doi.org/10.1111/ele.12617>
- 789 Bellman R (1966) Dynamic programming. Science 153(3731): 34–37.
790 <https://doi.org/10.1126/science.153.3731.34>

- 791 Belyadi H, Haghighat A (2021) Chapter 4 - Unsupervised machine learning: Clustering
792 algorithms. In: Belyadi H, Haghighat A (Eds) Machine Learning Guide for Oil and Gas Using
793 Python, 125–168. Gulf Professional Publishing. [https://doi.org/10.1016/B978-0-12-821929-](https://doi.org/10.1016/B978-0-12-821929-4.00002-0)
794 [4.00002-0](https://doi.org/10.1016/B978-0-12-821929-4.00002-0)
- 795 Bessey C, Jarman SN, Berry O, Olsen YS, Bunce M, Simpson T, ... Keesing J (2020)
796 Maximizing fish detection with eDNA metabarcoding. Environmental DNA 2(4): 493–504.
797 <https://doi.org/10.1002/edn3.74>
- 798 Blonder B, Lamanna C, Violle C, Enquist BJ (2014) The n-dimensional hypervolume. Global
799 Ecology and Biogeography 23(5): 595–609. <https://doi.org/10.1111/geb.12146>
- 800 Bremer LL, Farley KA (2010) Does plantation forestry restore biodiversity or create green
801 deserts? A synthesis of the effects of land-use transitions on plant species rich-ness.
802 Biodiversity and Conservation 19: 3893–3915. <https://doi.org/10.1007/s10531-010-9936-4>
- 803 Callahan BJ, McMurdie PJ, Holmes SP (2017) Exact sequence variants should replace
804 operational taxonomic units in marker-gene data analysis. ISME J. 11: 2639–2643.
805 <https://doi.org/10.1038/ismej.2017.119>
- 806 Caruana R, Niculescu-Mizil A (2006) An empirical comparison of supervised learning
807 algorithms. In: Proceedings of the 23rd International Conference on Machine Learning, 161–
808 168. <https://doi.org/10.1145/1143844.1143865>
- 809 Casadevall A, Fang FC (2014) Specialized science. Infection and Immunity 82(4): 1355–1360.
810 <https://doi.org/10.1128/IAI.01530-13>
- 811 Caughlan L, Oakley KL (2001) Cost considerations for long-term ecological monitoring.
812 Ecological Indicators 1(2): 123–134. [https://doi.org/10.1016/S1470-160X\(01\)00015-2](https://doi.org/10.1016/S1470-160X(01)00015-2)

813 Chave J, Réjou-Méchain M, Búrquez A, Chidumayo E, Colgan MS, Delitti WB, Duque A, Eid
814 T, Fearnside PM, Goodman RC, Henry M, Martínez-Yrizar A, Mugasha WA, Muller-Landau
815 HC, Mencuccini M, Nelson BW, Ngomanda A, Nogueira EM, Ortiz-Malavassi E, Péliissier R,
816 ... Vieilledent G (2014) Improved allometric models to estimate the aboveground biomass of
817 tropical trees. *Global Change Biology* 20(10): 3177–3190. <https://doi.org/10.1111/gcb.12629>

818 Chazdon RL (2014) *Second growth: The promise of tropical forest regeneration in an age of*
819 *deforestation*. University of Chicago Press.

820 Chazdon RL, Brancalion PH, Laestadius L, Bennett-Curry A, Buckingham K, Kumar C, Moll-
821 Rocek J, Vieira IC, Wilson SJ (2016) When is a forest a forest? Forest concepts and definitions
822 in the era of forest and landscape restoration. *Ambio* 45(5): 538–550.
823 <https://doi.org/10.1007/s13280-016-0772-y>

824 Cortes C, Vapnik V (1995) Support-vector networks. *Machine Learning* 20: 273–297.
825 <https://doi.org/10.1007/BF00994018>

826 Cover T, Hart P (1967) Nearest neighbor pattern classification. *IEEE Transactions on*
827 *Information Theory* 13(1): 21–27. <https://doi.org/10.1109/TIT.1967.1053964>

828 Crouzeilles R, Beyer HL, Monteiro LM, Feltran-Barbieri R, Pessôa ACM, Barros FSM,
829 Lindenmayer DB, Lino EDM, Grelle CEV, Chazdon RL, Matsumoto M, Rosa M, Latawiec
830 AE, Strassburg BBN (2020) Achieving cost-effective landscape-scale forest restoration
831 through targeted natural regeneration. *Conservation Letters* 13(3): e12709.
832 <https://doi.org/10.1111/conl.12709>

833 de Jong W, Liu J, Long H (2021) The forest restoration frontier. *Ambio* 50(12): 2224–2237.
834 <https://doi.org/10.1007/s13280-021-01614-x>

- 835 Deng D (2020) DBSCAN clustering algorithm based on density. In: 2020 7th International
836 Forum on Electrical Engineering and Automation (IFEEA), 949–953. IEEE.
837 <https://doi.org/10.1109/IFEEA51475.2020.00199>
- 838 Ester M, Kriegel H-P, Sander J, Xu X (1996) A density-based algorithm for discovering
839 clusters in large spatial databases with noise. In: Pfitzner DW, Salmon JK (Eds) Second
840 International Conference on Knowledge Discovery and Data Mining (KDD'96), 226–331.
- 841 Evans J (1992) Plantation forestry in the tropics: Tree planting for industrial, social,
842 environmental, and agroforestry purposes. Oxford University Press, USA.
- 843 Evans JD (1996) Straightforward statistics for the behavioral sciences. Thomson Brooks/Cole
844 Publishing Co.
- 845 Fernandez Nuñez N, Maggia L, Stenger P-L, Lelievre M, Letellier K, Gigante S, Manez A,
846 Mournet P, Ripoll J, Carriconde F (2021) Potential of high-throughput eDNA sequencing of
847 soil fungi and bacteria for monitoring ecological restoration in ultramafic substrates: The case
848 study of the New Caledonian biodiversity hotspot. Ecological Engineering 173: 106416.
849 <https://doi.org/10.1016/j.ecoleng.2021.106416>
- 850 Fierer N (2017) Embracing the unknown: Disentangling the complexities of the soil
851 microbiome. Nature Reviews. Microbiology 15(10): 579–590.
852 <https://doi.org/10.1038/nrmicro.2017.87>
- 853 Fix E, Hodges JL (1951) Discriminatory analysis. Nonparametric discrimination: Small sample
854 performance (Report No. A, 193008).
- 855 Folke C, Carpenter S, Walker B, Scheffer M, Elmqvist T, Gunderson L, Holling CS (2004)
856 Regime shifts, resilience, and biodiversity in ecosystem management. Annual Review of

- 857 Ecology, Evolution, and Systematics 35: 557–581.
858 <https://doi.org/10.1146/annurev.ecolsys.35.021103.105711>
- 859 Freund Y, Schapire RE (1996) Experiments with a new boosting algorithm. In: Proceedings of
860 the Thirteenth International Conference on Machine Learning (ICML), 148–156.
861 <https://doi.org/10.4236/iim.2010.26047>
- 862 Gallatin K, Albon C (2023) Machine learning with Python cookbook. O'Reilly Media, Inc.
- 863 Gangadhar S, Rangaswamy S (2018) Machine learning. In: Gudivada VN, Rao CR (Eds)
864 Handbook of Statistics 38: Data Science, 197–228. Elsevier.
865 <https://doi.org/10.1016/bs.host.2018.07.004>
- 866 Gann GD, McDonald T, Walder B, Aronson J, Nelson CR, Jonson J, Hallett JG, Eisenberg C,
867 Guariguata MR, Liu J, Hua F, Echeverría C, Gonzales E, Shaw N, Decler K, Dixon KW
868 (2019) International principles and standards for the practice of ecological restoration. Second
869 edition. Restoration Ecology 27(S1): S1–S46. <https://doi.org/10.1111/rec.13035>
- 870 Gann GD, McDonald T, Walder B, Aronson J, Nelson CR, Jonson J, ... Dixon K (2019)
871 International principles and standards for the practice of ecological restoration. Restoration
872 Ecology 27(S1): S1–S46. <https://doi.org/10.1111/rec.13035>
- 873 Goralczyk M, Lyet A, Naidoo R, Burton C, Pellissier L, Carraro L (2025) Combining eDNA
874 metabarcoding, hydrology-based modeling and camera trap datasets to assess the potential of
875 river eDNA in monitoring terrestrial mammals. Environmental DNA 7(2): e70089.
876 <https://doi.org/10.1002/edn3.70089>
- 877 Graham EB, Knelman JE (2023) Implications of soil microbial community assembly for
878 ecosystem restoration: Patterns, process, and potential. Microbial Ecology 85(3): 809–819.
879 <https://doi.org/10.1007/s00248-022-02155-w>

- 880 Gregorio NO, Herbohn JL, Harrison SR, Pasa A, Fernandez J, Tripoli R, Polinar B (2015)
881 Evidence-based best practice community-based forest restoration in Biliran: Integrating food
882 security and livelihood improvements into watershed rehabilitation in the Philippines.
883 Enhancing Food Security through Forest Landscape Restoration: Lessons from Burkina Faso,
884 Brazil, Guatemala, Viet Nam, Ghana, Ethiopia and Philippines, 177–217.
885 <https://doi.org/10.2305/IUCN.CH.2015.FR.2.en>
- 886 Guzman LC, Lawrence JH, Marler TE (1997) Growth and development responses of three
887 Acacia species to long-term light regimes. *Journal of Sustainable Forestry* 4(1–2): 119–130.
888 https://doi.org/10.1300/J091v04n01_06
- 889 Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: Data mining,
890 inference, and prediction (2nd ed.). Springer. <https://doi.org/10.1007/978-0-387-84858-7>
- 891 He K, Zhang X, Ren S, Sun J (2015) Deep residual learning for image recognition. arXiv
892 preprint arXiv:1512.03385. <https://arxiv.org/abs/1512.03385>
- 893 Heanes DL (1984) Determination of total organic-C in soils by an improved chromic acid
894 digestion and spectrophotometric procedure. *Communications in Soil Science and Plant*
895 *Analysis* 15(10): 1191–1213. <https://doi.org/10.1080/00103628409367551>
- 896 Hobbs RJ, Hallett LM, Ehrlich PR, Mooney HA (2011) Intervention ecology: Applying
897 ecological science in the twenty-first century. *BioScience* 61: 442–450.
898 <https://doi.org/10.1525/bio.2011.61.6.6>
- 899 Holling CS (1973) Resilience and stability of ecological systems.
- 900 Hotelling H (1933) Analysis of a complex of statistical variables into principal components.
901 *Journal of Educational Psychology* 24(6): 417–441. <https://doi.org/10.1037/h0071325>

- 902 Hughes TP, Linares C, Dakos V, Van De Leemput IA, Van Nes EH (2013) Living dangerously
903 on borrowed time during slow, unrecognized regime shifts. *Trends in Ecology & Evolution*
904 28(3): 149–155. <https://doi.org/10.1016/j.tree.2012.08.022>
- 905 Hussain SS, Zaidi SSH (2024) Adaboost ensemble approach with weak classifiers for gear fault
906 diagnosis and prognosis in DC motors. *Applied Sciences* 14(7): 3105.
907 <https://doi.org/10.3390/app14073105>
- 908 Jain AK, Dubes RC (1988) Algorithms for clustering data. Prentice-Hall, Inc.
- 909 Jolliffe IT (2002) Principal component analysis for special types of data. In: *Principal*
910 *Component Analysis*, 338–372. Springer, New York, NY. [https://doi.org/10.1007/0-387-](https://doi.org/10.1007/0-387-22440-8_13)
911 [22440-8_13](https://doi.org/10.1007/0-387-22440-8_13)
- 912 Jordan MI, Mitchell TM (2015) Machine learning: Trends, perspectives, and prospects. *Science*
913 349(6245): 255–260. <https://doi.org/10.1126/science.aaa8415>
- 914 Ke Y, Liu T, Han C, Yu X, Wang J, Ding L, Pan H, Mo X, Lu X (2025) A review of eDNA
915 technology in avian monitoring: Current status, challenges and future perspectives. *Avian*
916 *Research* 16(2): 100235. <https://doi.org/10.1016/j.avrs.2025.100235>
- 917 Kimmel K, Dee LE, Avolio ML, Ferraro PJ (2021) Causal assumptions and causal inference
918 in ecological experiments. *Trends in Ecology & Evolution* 36(12): 1141–1152.
919 <https://doi.org/10.1016/j.tree.2021.08.008>
- 920 Krisnawati H, Kallio M, Kanninen M (2011) *Acacia mangium* Willd.: Ecology, silviculture
921 and productivity. CIFOR.
- 922 Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet classification with deep convolutional
923 neural networks. In: Pereira F, Burges CJ, Bottou L, Weinberger KQ (Eds) *Advances in Neural*
924 *Information Processing Systems* 25: 1097–1105. Curran Associates, Inc.

- 925 https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c
926 [45b-Paper.pdf](#)
- 927 Le HD, Smith C, Herbohn J (2014) What drives the success of reforestation projects in tropical
928 developing countries? The case of the Philippines. *Global Environmental Change* 24: 334–348.
929 <https://doi.org/10.1016/j.gloenvcha.2013.09.010>
- 930 Le HD, Smith C, Herbohn J, Harrison S (2012) More than just trees: Assessing reforestation
931 success in tropical developing countries. *Journal of Rural Studies* 28(1): 5–19.
932 <https://doi.org/10.1016/j.jrurstud.2011.07.006>
- 933 LeCun Y, Bengio Y, Hinton G (2015) Deep learning. *Nature* 521(7553): 436–444.
934 <https://doi.org/10.1038/nature14539>
- 935 Lemke M, DeSalle R (2023) The next generation of microbial ecology and its importance in
936 environmental sustainability. *Microbial Ecology* 85(3): 781–795.
937 <https://doi.org/10.1007/s00248-023-02185-y>
- 938 Levin SA (1992) The problem of pattern and scale in ecology: The Robert H. MacArthur award
939 lecture. *Ecology* 73(6): 1943–1967. <https://doi.org/10.2307/1941447>
- 940 Li J, Shangguan Z, Deng L (2020) Dynamics of soil microbial metabolic activity during
941 grassland succession after farmland abandonment. *Geoderma* 363: 114167.
942 <https://doi.org/10.1016/j.geoderma.2019.114167>
- 943 Liu H, Motoda H (Eds) (1998) Feature extraction, construction and selection: A data mining
944 perspective 453. Springer Science & Business Media.
- 945 Lloyd S (1982) Least squares quantization in PCM. *IEEE Transactions on Information Theory*
946 28(2): 129–137. <https://doi.org/10.1109/TIT.1982.1056489>

- 947 Loehle C (2004) Challenges of ecological complexity. *Ecological Complexity* 1(1): 3–6.
948 <https://doi.org/10.1016/j.ecocom.2003.09.003>
- 949 López-González M, Piquet JC, Maestresalas B, López-Darias M (2025) Validation of eDNA
950 methods for managing the terrestrial invasive snake *Lampropeltis californiae* on the Canary
951 Islands. *Scientific Reports* 15(1): 14116. <https://doi.org/10.1038/s41598-025-96387-8>
- 952 Louca S, Parfrey LW, Doebeli M (2016) Decoupling function and taxonomy in the global
953 ocean microbiome. *Science* 353(6305): 1272–1277. <https://doi.org/10.1126/science.aaf4507>
- 954 MacQueen J (1967) Some methods for classification and analysis of multivariate observations.
955 In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*,
956 Volume 1: Statistics 5: 281–298. University of California Press.
- 957 Malhi Y, Riutta T, Wearn OR, Deere NJ, Mitchell SL, Bernard H, ... Struebig MJ (2022)
958 Logged tropical forests have amplified and diverse ecosystem energetics. *Nature* 612(7941):
959 707–713. <https://doi.org/10.1038/s41586-022-05523-1>
- 960 McKnight PE, Najab J (2010) Mann-Whitney U test. *The Corsini Encyclopedia of Psychology*,
961 1–1. <https://doi.org/10.1002/9780470479216.corpsy0524>
- 962 Mitchell JC, Kashian DM, Chen X, Cousins S, Flaspohler D, Gruner DS, ... Buma B (2023)
963 Forest ecosystem properties emerge from interactions of structure and disturbance. *Frontiers*
964 in *Ecology and the Environment* 21(1): 14–23. <https://doi.org/10.1002/fee.2589>
- 965 Mitchell TM (1997) *Machine learning*. McGraw-Hill.
- 966 Moreno-Mateos D, Alberdi A, Morriën E, van der Putten WH, Rodríguez-Uña A, Montoya D
967 (2020) The long-term restoration of ecosystem complexity. *Nature Ecology & Evolution* 4(5):
968 676–685. <https://doi.org/10.1038/s41559-020-1154-1>

- 969 Murtagh F, Legendre P (2014) Ward's hierarchical agglomerative clustering method: Which
970 algorithms implement Ward's criterion?. *Journal of Classification* 31: 274–295.
971 <https://doi.org/10.1007/s00357-014-9161-z>
- 972 Nguyen NH, Song Z, Bates ST, Branco S, Tedersoo L, Menke J, Schilling JS, Kennedy PG
973 (2016) FUNGuild: An open annotation tool for parsing fungal community datasets by
974 ecological guild. *Fungal Ecology* 20: 241–248. <https://doi.org/10.1016/j.funeco.2015.06.006>
- 975 Nielsen F (2016) Hierarchical clustering. *Introduction to HPC with MPI for Data Science*, 195–
976 211. https://doi.org/10.1007/978-3-319-21903-5_8.
- 977 Novello N, Tonello AM (2024) $\$f\$$ -Divergence based classification: Beyond the use of cross-
978 entropy. In: Salakhutdinov R, Kolter Z, Heller K, Weller A, Oliver N, Scarlett J, Berkenkamp
979 F (Eds) *Proceedings of the 41st International Conference on Machine Learning* 235: 38448–
980 38473. PMLR. <https://proceedings.mlr.press/v235/novello24a.html>
- 981 Nowak DJ, Crane DE (2002) Carbon storage and sequestration by urban trees in the USA.
982 *Environmental Pollution* 116(3): 381–389. [https://doi.org/10.1016/S0269-7491\(01\)00214-7](https://doi.org/10.1016/S0269-7491(01)00214-7)
- 983 Odum EP, Barrett GW (1971) *Fundamentals of ecology*. Saunders.
- 984 Oksanen J, Simpson G, Blanchet F, Kindt R, Legendre P, Minchin P, O'Hara R, Solymos P,
985 Stevens M, Szoecs E, Wagner H, Barbour M, Bedward M, Bolker B, Borcard D, Carvalho G,
986 Chirico M, De Caceres M, Durand S, Evangelista H, FitzJohn R, Friendly M, Furneaux B,
987 Hannigan G, Hill M, Lahti L, McGlinn D, Ouellette M, Ribeiro Cunha E, Smith T, Stier A, Ter
988 Braak C, Weedon J (2024) *Vegan: Community ecology package (Version 2.6-8) [R package]*.
989 <https://CRAN.R-project.org/package=vegan>

- 990 Olsen SR, Cole CV, Watanabe FS (1954) Estimation of available phosphorus in soils by
991 extraction with sodium bicarbonate (USDA Circular No. 939). U.S. Government Printing
992 Office.
- 993 Paine RT (1969) A note on trophic complexity and community stability. The American
994 Naturalist 103(929): 91–93. <https://doi.org/10.1086/282586>
- 995 Pastor J, Durkee Walker R (2006) Delays in nutrient cycling and plant population oscillations.
996 Oikos 112(3): 698–705. <https://doi.org/10.1111/j.0030-1299.2006.14478.x>
- 997 Pearson K (1896) VII. Mathematical contributions to the theory of evolution.—III. Regression,
998 heredity, and panmixia. Philosophical Transactions of the Royal Society of London. Series A,
999 Containing Papers of a Mathematical or Physical Character 187: 253–318.
1000 <https://doi.org/10.1098/rsta.1896.0007>
- 1001 Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M,
1002 Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M,
1003 Perrot M, Duchesnay E (2011) Scikit-learn: Machine learning in Python. Journal of Machine
1004 Learning Research 12: 2825–2830.
1005 <https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf>
- 1006 Peres-Neto PR, Jackson DA, Somers KM (2003) Giving meaningful interpretation to
1007 ordination axes: Assessing loading significance in principal component analysis. Ecology
1008 84(9): 2347–2363. <https://doi.org/10.1890/00-0634>
- 1009 Pigani E, Sgarbossa D, Suweis S, Maritan A, Azalee S (2022) Delay effects on the stability of
1010 large ecosystems. Proceedings of the National Academy of Sciences of the United States of
1011 America 119(45): e2211449119. <https://doi.org/10.1073/pnas.2211449119>

- 1012 Python Software Foundation (2024) Python (Version 3.12.4) [Computer software].
1013 <https://www.python.org/>
- 1014 Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO (2013)
1015 The SILVA ribosomal RNA gene database project: Improved data processing and web-based
1016 tools. Nucleic Acids Research 41(Database issue): D590–D596.
1017 <https://doi.org/10.1093/nar/gks1219>
- 1018 Quinlan JR (1993) C4.5: Programs for machine learning. Morgan Kaufmann Publishers.
- 1019 R Core Team (2022) R: A language and environment for statistical computing (Version 4.2.2)
1020 [Computer software]. R Foundation for Statistical Computing. <https://www.R-project.org/>
- 1021 Sarker IH (2021) Machine learning: Algorithms, real-world applications and research
1022 directions. SN Computer Science 2(3): 160. <https://doi.org/10.1007/s42979-021-00592-s>
- 1023 Sarker IH, Abushark YB, Khan AI (2020a) ContextPCA: Predicting context-aware smartphone
1024 apps usage based on machine learning techniques. Symmetry 12(4): 499.
1025 <https://doi.org/10.3390/sym12040499>
- 1026 Sarker IH, Alqahtani H, Alsolami F, Khan AI, Abushark YB, Siddiqui MK (2020b) Context
1027 pre-modeling: An empirical analysis for classification based user-centric context-aware
1028 predictive modeling. Journal of Big Data 7: 1–23. [https://doi.org/10.1186/s40537-020-00328-](https://doi.org/10.1186/s40537-020-00328-3)
1029 3
- 1030 Note: The duplicate reference (Sarker et al. 2020, J Big Data) was merged into the 2020b entry.
- 1031 Scheffer M, Carpenter S, Foley JA, Folke C, Walker B (2001) Catastrophic shifts in
1032 ecosystems. Nature 413(6856): 591–596. <https://doi.org/10.1038/35098000>
- 1033 Singh D, Singh B (2020) Investigating the impact of data normalization on classification
1034 performance. Applied Soft Computing 97: 105524. <https://doi.org/10.1016/j.asoc.2019.105524>

- 1035 Sokolova M, Lapalme G (2009) A systematic analysis of performance measures for
1036 classification tasks. *Information Processing & Management* 45(4): 427–437.
1037 <https://doi.org/10.1016/j.ipm.2009.03.002>
- 1038 Tahmasebi P, Kamrava S, Bai T, Sahimi M (2020) Machine learning in geo-and environmental
1039 sciences: From small to large scale. *Advances in Water Resources* 142: 103619.
1040 <https://doi.org/10.1016/j.advwatres.2020.103619>
- 1041 Teng CM (2011) Data, data, everywhere: Statistical issues in data mining. In: *Philosophy of*
1042 *Statistics*, 1099–1117. North-Holland.
- 1043 Thomsen PF, Willerslev E (2015) Environmental DNA—an emerging tool in conservation for
1044 monitoring past and present biodiversity. *Biological Conservation* 183: 4–18.
1045 <https://doi.org/10.1016/j.biocon.2014.11.019>
- 1046 Thorndike RL (1953) Who belongs in the family? *Psychometrika* 18(4): 267–276.
1047 <https://doi.org/10.1007/BF02289263>
- 1048 Tilman D, Downing JA (1994) Biodiversity and stability in grasslands. *Nature* 367(6461): 363–
1049 365. <https://doi.org/10.1038/367363a0>
- 1050 Toro L, Werden LK, Addo-Danso SD, Andersen KM, Batterman S, Bragadini MM, Choksi P,
1051 Cole RJ, Comita LS, Cusack D, Dent DH, Dietterich LH, Fisher JB, Fleischer K, Fuchslueger
1052 L, Huanca-Nunez N, Lienau JR, McCulloch LA, Morrissey EM, Powers JS, . . . Wong MY
1053 (2025) Integrating belowground recovery into tropical forest restoration design and
1054 monitoring. *BioScience*. <https://doi.org/10.1093/biosci/biaf097>
- 1055 Vivian J, Chazdon RL, Catling AA, Lee DJ (2025a) Early evidences of links between soil
1056 microbes and forest restoration through multiple ecosystem metrics. *Frontiers in Forests and*
1057 *Global Change* 8: 1540513. <https://doi.org/10.3389/ffgc.2025.1540513>

- 1058 Vivian J, Scriver M, Zaiko A, Shapcott A, Chazdon RL, Catling AA, Lee DJ (2025b) Recovery
1059 of soil microbiota in naturally regenerating *Acacia mangium* ecosystems. PeerJ. Under review.
- 1060 Vivian J, Chazdon RL, Catling AA, Shapcott A, Herbohn J, Lee DJ (2026) *Acacia mangium*
1061 monocultures can catalyse the recovery of the tree community and aboveground carbon stock
1062 in the Philippines. *Forest Ecology and Management* 601: 123357.
1063 <https://doi.org/10.1016/j.foreco.2025.123357>
- 1064 Wall DH, Nielsen UN, Six J (2015) Soil biodiversity and human health. *Nature* 528(7580): 69–
1065 76. <https://doi.org/10.1038/nature15744>
- 1066 Ward JH (1963) Hierarchical grouping to optimize an objective function. *Journal of the*
1067 *American Statistical Association* 58(301): 236–244.
1068 <https://doi.org/10.1080/01621459.1963.10500845>
- 1069 Watts K, Whytock RC, Park KJ, Fuentes-Montemayor E, Macgregor NA, Duffield S,
1070 McGowan PJK (2020) Ecological time lags and the journey towards conservation success.
1071 *Nature Ecology & Evolution* 4(3): 304–311. <https://doi.org/10.1038/s41559-019-1087-8>
- 1072 Wei L, Bergeron Y, De Frenne P, Verheyen K, Tian L, Ren H, Jian S (2024) Above-and
1073 belowground composition and diversity of subtropical plantations and their relationships with
1074 soil nutrient stocks. *Plant and Soil* 495(1): 235–252. [https://doi.org/10.1007/s11104-023-](https://doi.org/10.1007/s11104-023-06317-8)
1075 [06317-8](https://doi.org/10.1007/s11104-023-06317-8)
- 1076 Williams BA, Beyer HL, Fagan ME, Chazdon RL, Schmoeller M, Sprenkle-Hyppolite S,
1077 Griscom BW, Watson JEM, Tedesco AM, Gonzalez-Roglich M, Daldegan GA, Bodin B,
1078 Celentano D, Wilson SJ, Rhodes JR, Alexandre NS, Kim D-H, Bastos D, Crouzeilles R (2024)
1079 Global potential for natural regeneration in deforested tropical regions. *Nature* 636(8041): 131–
1080 137. <https://doi.org/10.1038/s41586-024-08106-4>

1081 Woolnough AP, Hollenberg LCL, Cassey P, Prowse TAA (2023) Quantum computing: A new
1082 paradigm for ecology. Trends in Ecology & Evolution 38(8): 727–735.
1083 <https://doi.org/10.1016/j.tree.2023.04.001>

1084 Yan D, Mills JG, Gellie NJ, Bissett A, Lowe AJ, Breed MF (2018) High-throughput eDNA
1085 monitoring of fungi to track functional recovery in ecological restoration. Biological
1086 Conservation 217: 113–120. <https://doi.org/10.1016/j.biocon.2017.10.035>

1087 Zhang Y, Mańdziuk J, Quek CH, Goh BW (2017) Curvature-based method for determining the
1088 number of clusters. Information Sciences 415: 414–428.
1089 <https://doi.org/10.1016/j.ins.2017.05.024>

1090 Figures legends

1091 **Fig. 1** (A) Study site; sampled plots (coloured points) composing the transects for each landcover type; (B) Set
1092 of transects for one landcover type, and details on the disposition of the plots and parameters collected.

1093 DBH: diameter at breast height; SOC: soil organic carbon; N: nitrogen; P: phosphorus.

1094 **Fig. 2** Machine Learning analysis conducted in our study.

1095 KNN: K-nearest neighbours; SVM: Support vector machines; AdaBoost: Adaptive boosting; DT: Decision tree;
1096 PCA: Principal component analysis.

1097 **Fig. 3** Illustration of the expected increase in accuracy when merging two similar categories.

1098 On the left, individual data points (observations or samples composed of the value of each parameter assessed)
1099 for three distinct classes (pink, blue, and green) are displayed. Their proximity demonstrates their high similarity
1100 (blue points within the section of the pink class and vice versa), resulting in a low accuracy. On the right, the pink
1101 and blue classes are merged. In this scenario, the simplified boundary computation allows for an accuracy
1102 improvement, demonstrating the similarity between the A and B classes.

1103 **Fig. 4** Unsupervised classification with three different machine learning algorithms, considering the dataset with
1104 reduced dimensions through PCA, and seed 0.

1105 Left column: plots with points representing the samples, spatially organised and labelled according to the clusters
1106 identified by the algorithm. Right column: plots with points representing the samples (i.e., comprehending the
1107 values of the variables), spatially organised according to the clusters identified by the algorithms, labelled
1108 considering the landcover type to which they belong.

1109 **Fig. 5** (A) Representation of the ecosystem components and properties (circular arrows) variation over forest
1110 recovery. Diagram in (B) explains the differential time required for the variation of the components and properties.

1111 In (A), the ecosystem properties shift to a different state (i.e., state “Y”) only after the variation of the
1112 environmental components (i.e., tree community, soil microbial communities) from which the properties emerge.
1113 Diagram in (B) displays the variation of the ecosystem (when it approaches the threshold for the ecosystem shift)
1114 into different phases of forest recovery, only after the modification of the ecosystem parameters, occurring during
1115 transition phases. Dotted curves represent the shift of a more resilient ecosystem compared to one represented by
1116 dashed curves.