ECOGRAPHY

Software notes

forestexplorR: an R package for the exploration and analysis of stem-mapped forest stand data

Stuart I. Graham, Ariel Rokem and Janneke Hille Ris Lambers

S. I. Graham (https://orcid.org/0000-0002-6974-9963) ✉ (sg9319@my.bristol.ac.uk) and J. Hille Ris Lambers, Dept of Biology, Univ. of Washington, Seattle, WA, USA and Inst. of Integrative Biology, ETH Zurich, Zürich, Switzerland. – A. Rokem, Dept of Psychology and eScience Inst., Univ. of Washington, Seattle, WA, USA.

Ecography
2022: e06223
doi: 10.1111/ecog.06223
Subject Editor:
Michael Krabbe Borregaard
Editor-in-Chief: Miguel Araújo
Accepted 7 June 2022

Stem-mapped forest stands offer important opportunities for investigating the fine-scale spatial processes occurring in forest ecosystems. These stands are areas of the forest where the precise locations and repeated size measurements of each tree are recorded, thereby enabling the calculation of spatially-explicit metrics of individual growth rates and of the entire tree community. The most common use of these datasets is to investigate the drivers of variation in forest processes by modeling tree growth rate or mortality as a function of these neighborhood metrics. However, neighborhood metrics could also serve as important covariates of many other spatially variable forest processes, including seedling recruitment, herbivory and soil microbial community composition. Widespread use of stem-mapped forest stand datasets is currently hampered by the lack of standardized, efficient and easy-to-use tools to calculate tree dynamics (e.g. growth, mortality) and the neighborhood metrics that impact them. We present the forestexplorR package that facilitates the munging, exploration, visualization and analysis of stem-mapped forest stands. By providing flexible, user-friendly functions that calculate neighborhood metrics and implement a recently-developed rapid-fitting tree growth and mortality model, forestexplorR broadens the accessibility of stem-mapped forest stand data. We demonstrate the functionality of forestexplorR by using it to investigate how the species identity of neighboring trees influences the growth rates of three common tree species in Mt Rainier National Park, WA, USA. forestexplorR is designed to facilitate researchers to incorporate spatially-explicit descriptions of tree communities in their studies and we expect this increased diversity of contributors to develop exciting new ways of using stem-mapped forest stand data.

Keywords: forest stand, mortality, neighborhood model, tree density, tree growth, visualization

NORDIC SOCIETY OIKOS

www.ecography.org

Introduction

Forest ecosystems represent a critical buffer against anthropogenic climate change and we therefore must develop a deep understanding of their dynamics. An important tool in this endeavor is the stem-mapped forest stand; an area of forest where the size, species identity and precise location of each tree meeting a minimum size threshold are recorded (Condit 1995). Once established, these stands are revisited periodically to document mortality, re-measure trees and record any additional trees that newly meet the minimum size threshold. In particular, stem-mapped forest stands ≥ 1 ha in area (e.g. those of the Forest Global Earth Observatory (ForestGEO; Davies et al. 2021) and Pacific Northwest Permanent Sample Plot networks (Franklin et al. 2021)) offer important opportunities for investigating the fine-scale spatial processes that drive forest dynamics (Lutz 2015). This is because they allow us to quantify the local tree community (hereafter 'neighborhood' for consistency with published literature) surrounding any particular tree in a spatially-explicit way. Neighborhood metrics have been shown to be informative covariates of spatial variation in tree growth and survival (Uriarte et al. 2004, Wiegand et al. 2017). Neighborhood metrics have also facilitated research on climate change (Buechling et al. 2017), competitive interactions (Fortunel et al. 2016) and pest-induced tree mortality (Buonanduci et al. 2020).

To extract these insights from stem-mapped forest stand datasets, a considerable amount of programming skill and computational power is required. The first task is to clean and format the data for analysis. This is a non-trivial data management problem because errors are common in field-collected data and each group of forest stands tends to have its own data structure. Next, the data must be explored to identify the important neighborhood metrics; an iterative process given that the appropriate neighborhood size is generally not known a priori, as it depends on local environment, the spatial process being explored (e.g. competition versus soil microbiomes), species identity and many other factors. In addition, calculating neighborhood metrics requires identifying all trees within the neighborhood of each tree, typically generating datasets with hundreds of thousands of rows; a daunting workflow for an inexperienced programmer. The final step is to build models for hypothesis testing that incorporate neighborhood metrics, and many of these models require high-performance computing resources for maximum likelihood estimation (Uriarte et al. 2004, Kunstler et al. 2016, Fortunel et al. 2018), which not all researchers have access to. Software tools that can alleviate these challenges would provide opportunities to increase the abundance and diversity of projects using these rich datasets.

Software tools for interacting with stem-mapped forest stand data are available, but their design restricts their use to specific dataset structures and analytical approaches. For example, the fgeo suite of R packages (Lepore et al. 2019) facilitates analyses of data from the ForestGEO network of stem-mapped forest stands (Davies et al. 2021). However, fgeo does not contain any functions for calculating neighborhood metrics, and its powerful habitat-species association analyses require individual tree-level elevation data, which are currently not available for many stands outside the ForestGEO network. The scope of the fgeo package is understandably limited in order to encourage researchers to explore all aspects of the ForestGEO network in particular, but we believe there is a great deal of insight to be gained from tools that can be applied to the data types common to all stem-mapped forest stands: the size, species identity and precise locations of trees. For example, the forestecology R package (Kim et al. 2021) implements a highly efficient neighborhood growth model that allows users to estimate species interaction coefficients, but it can only handle data from a single large stem-mapped forest stand (e.g. a ForestGEO stand) and does not allow the user to include additional covariates such as climatic data. Also of note is the rFIA package (Stanke et al. 2020), which supports the retrieval and analysis of data from the USFS Forest Inventory and Analysis (FIA) stands. However, the FIA stands, and all other circular stands we are aware of, are too small to encapsulate entire neighborhoods in most forest ecosystems (Lutz 2015), but designed instead to cover a broad distribution and thereby permit analysis of tree performance patterns across large environmental gradients.

To address this gap in the current technical tool landscape, we present the forestexplorR package, which facilitates the exploration, visualization and analysis of rectangular stem-mapped forest stand data. forestexplorR is specifically designed to identify and calculate important neighborhood metrics for use in further analyses and therefore requires data from forest stands large enough to capture complete neighborhoods (typically, neighborhood radius ≥ 10 m; Uriarte et al. 2004, Fortunel et al. 2016). forestexplorR is compatible with all rectangular stem-mapped forest stand datasets because it requires only the common data types of the size, species identity and position of trees, yet its functions can also combine data from many forest stands. forestexplorR provides functions for calculating common neighborhood metrics, and creates clean visualizations for data exploration. In addition, it allows the user to implement a new, efficient neighborhood model of tree growth or mortality (Graham et al. 2021) that can be used to select an appropriate neighborhood size, guide the development of a more mechanistic model (Uriarte et al. 2004), and test ecological hypotheses. forestexplorR will most significantly increase the utility of distributed networks of stem-mapped forest stands because it provides user-friendly tools for combining data from multiple stands into a single neighborhood analysis, thereby enabling users to test hypotheses on how forest processes vary between biomes and across environmental gradients. Some examples of such distributed networks are: Pacific Northwest Permanent Sample Plot network (Franklin et al. 2021), NEON Distributed and Tower Plots (NEON 2012) and both the RAINFOR and AfriTRON plots whose data are curated by <www.forestplot.net> (Lopez-Gonzalez et al. 2009, 2011).

Package structure and functionality

The functions included in the forestexplorR package are divided into five categories based on the analytical task they are designed to assist with (Fig. 1). All users should begin with the functions in the 'data formatting' category, which verify the user-provided data are in a format that other forestexplorR functions can handle. Functions of the describing neighborhoods category create neighborhoods of a user-defined size and calculate corresponding neighborhood metrics. There are also sets of functions for calculating growth rates and visualizing mapped stands. The 'modeling' category contains two functions; one that fits a neighborhood model of tree growth or mortality, and another that helps the user determine a suitable neighborhood size to use in their analyses.

In this section, we describe functions included in the package under each of these categories, and list some of their potential uses. Figure 1 provides guidance on how the function categories can be combined. forestexplorR also includes a series of built-in stem-mapped forest stand datasets from Mt Rainier National Park (Franklin et al. 2021) that can be used to test its functions (see 'examples' in each of the function help files). In addition, the forestexplorR website has a series of detailed vignettes describing how to use each of these functions (<https://sgraham9319.github.io/forestexplorR/index.html>).

Data formatting

forestexplorR requires stem-mapped forest stand data to be provided in two separate data frames; one containing tree mapping data (i.e. spatial coordinates), and the other containing tree measurement data. The decision to require separation of these two data types was made to avoid redundancy, as most trees will have multiple records of measurement data (one for each stand census) but only one record of mapping data. It follows that each tree should populate only a single

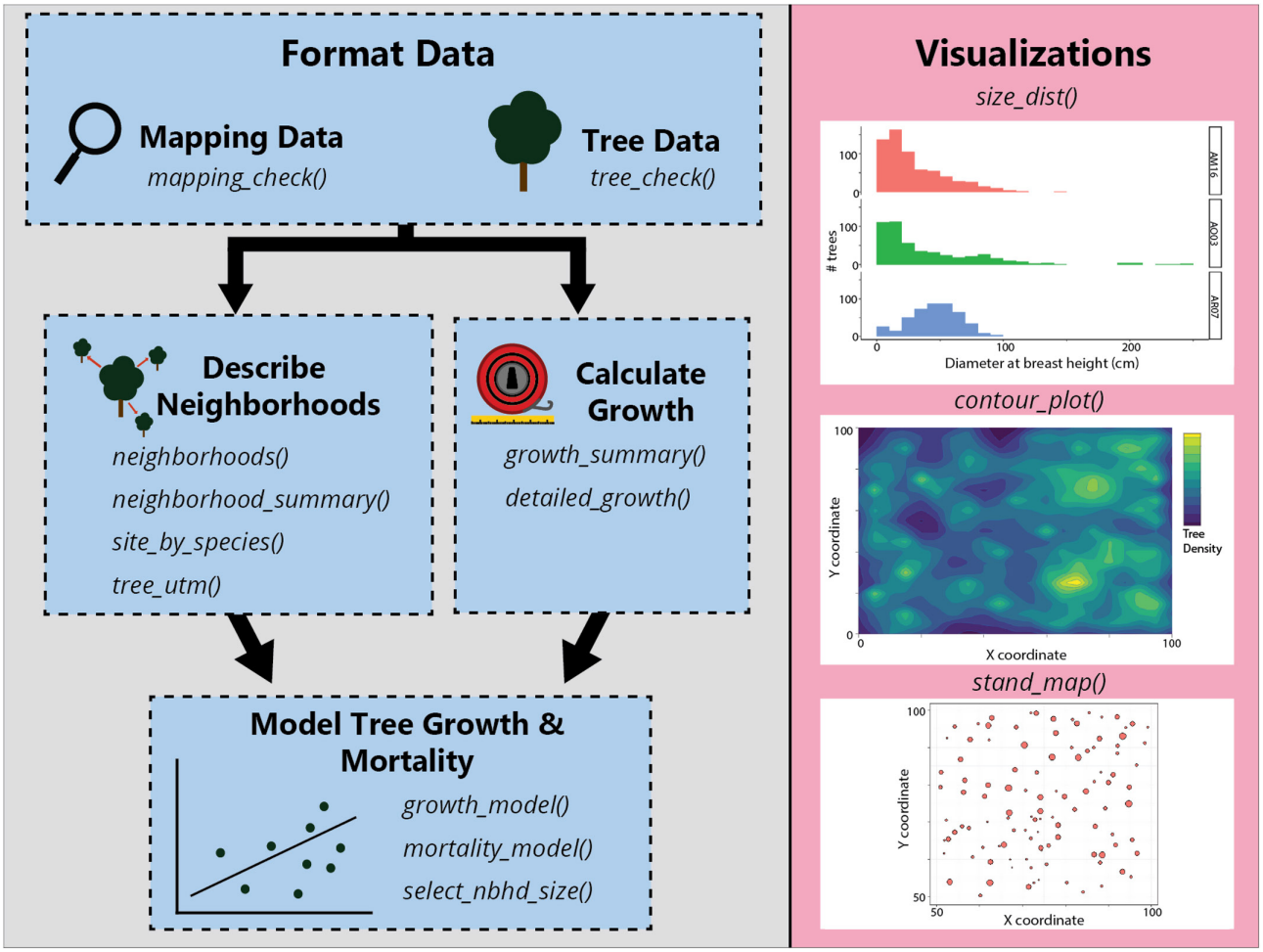

Figure 1. Schematic of forestexplorR functions. Black arrows represent a potential analysis plan that combines multiple function categories; formatted mapping and tree measurement datasets can be used to obtain quantitative neighborhood descriptions and calculate tree growth rates, which can be combined to develop tree growth and mortality models. Functions of the 'visualizations' category can be used for exploratory analysis: size_dist() creates size class distributions of one or more stands; contour_plot() uses point measurements of any variable to create an interactive contour map of the stand; stand_map() creates a map of all trees in the stand, indicating their size.

row in the mapping data frame, but will populate n rows in the tree measurement data frame, where n is the number of censuses in which the tree was measured. If working with stem-mapped forest stands where mapping and tree measurement data are kept in a single file, the user will need to separate the two data types prior to using forestexplorR, which can be achieved using functions of the dplyr package. For example, tree x and y coordinates could be extracted from a combined dataset with: combined_dataset_name % > % group_by(tree_id) % > % summarize (x_coord = x_coord[1], y_coord = y_coord[1]). For other stem-mapped forest stands, there may be a separate file of tree measurement data for each census, but these can easily be combined into a single tree measurement file using dplyr::bind_rows.

The functions *mapping_check* and *tree_check* are provided to verify that mapping and tree measurement data frames are in a format compatible with forestexplorR functions. Both functions return a list of two elements: the first element is a data frame of the ID codes of all the trees that have a potential data issue (e.g. trees with no mapping data, a single ID code referring to multiple trees, etc.) and a short description of the issue (see function documentation for list of issues); the second element is a data frame summarizing the number and percentage of tree ID codes that have at least one issue, and that have each of the specific issues. The purpose of the second element is to help the user evaluate the amount of data that would be lost if all trees with certain issues were simply excluded from analysis. These functions do not remove any trees with data issues because it is likely that many of the data issues can be resolved through reference to field notes, etc. It is the user's responsibility to remove any trees with issues that cannot be resolved before continuing analysis.

In the interest of flexibility, the mapping and tree measurement data frames can each contain as many additional columns as desired (and these columns are retained in the output where possible; see function documentation for details), but the *mapping_check* and *tree_check* functions will stop and return warning messages if any required columns are missing. The mapping data frame must contain the columns: tree_id (code uniquely identifying each tree), stand_id (name of stand in which the tree is located), species (species identity of the tree as a string of any length containing no spaces), x_coord and y_coord (x and y coordinates of the tree, in meters from stand origin). We note here that because forestexplorR is specifically designed for rectangular stem-mapped forest stand data, the x and y coordinates represent distances from the stand corner used as the origin (as opposed to the stand center used as the origin in circular stem-mapped stands). The tree measurement data frame must contain the columns: tree_id, stand_id, species, year (year in which measurement was taken) and dbh (diameter at breast height in cm; hereafter DBH). In order to use the *mortality_model* function, the tree measurement data frame will also need to include a mort column containing mortality status of each tree at each census; *tree_check* will check this column for missing data but will not return an error if the column is missing. All forestexplorR functions, except those that create plots of a single stand (i.e. *contour_plot* and *stand_map*), accept datasets containing information from multiple stands as input.

Describing neighborhoods

The first step in calculating neighborhood metrics is to identify the set of trees in each neighborhood. The *neighborhoods* function achieves this using a mapping and a tree measurement data frame and a specified neighborhood radius (for assistance in selecting an appropriate neighborhood radius, see the *select_nbhd_size* function described in 'modeling tree growth' section). It returns a new data frame where each neighborhood ID appears on multiple rows, with each row containing information on one of the trees in that neighborhood (i.e. growing within 'radius' of the neighborhood center). By default, a neighborhood is constructed for each individual tree in the mapping data frame (and neighborhood IDs are equal to focal tree IDs), but the function can instead construct neighborhoods centered on specific locations within the stand if a data frame of these locations is provided under the argument 'coords'. This is useful if some other information has been sampled at a specific location in the stand (e.g. seed trap, soil microbial community), and the goal is to model it as a function of the local tree community. In addition, if the provided mapping dataset contains data from multiple forest stands, the user can specify for which stands the neighborhoods should be calculated. The data frame output by *neighborhoods* serves as the input for a number of other functions in forestexplorR, including *neighborhood_summary* and *site_by_species*.

The *neighborhood_summary* function is applied to the data frame output by *neighborhoods*, and calculates neighborhood metrics for each tree or user-provided set of coordinates. Specifically, it calculates species richness, Shannon–Wiener diversity, Simpson's diversity, Pielou's J as a measure of evenness, overall tree density and the density of each tree species in each neighborhood. The purpose of the density metrics is to quantify the amount of competition in the neighborhood, considering both the sizes and distances of neighboring trees. Many different competition metrics have been developed in the literature but most of these require the optimization of several parameters such as exponents that determine the relative importance of neighbor size versus neighbor distance (Canham et al. 2004, Uriarte et al. 2004). However, the goal of forestexplorR is to allow rapid analysis of stem-mapped forest stand data by calculating metrics that can act as covariates in a highly efficient linear-modeling framework (see *growth_model* and *mortality_model*), and therefore we do not explore any neighborhood metrics that require parameter optimization. This means that our neighborhood metrics cannot represent the empirically-supported non-linear relationships underlying tree performance (e.g. the relationship between neighbor distance and the competitive effect on focal growth), but despite this simplification, these linear models have been shown to predict tree growth rates with accuracy (Graham et al. 2021). This simplification does not prevent all exploration of density metrics and

neighborhood_summary provides several options for how densities can be computed. The default option is to measure density in units of m^2 basal area per hectare, but the resulting species-specific densities can also be optionally converted to the proportions of total basal area (summed across trees of all species) that they constitute. Alternatively, densities can be calculated with the angular method (Rouvinen and Kuuluvainen 1997)(Eq. 1):

$$\text{species A density} = \sum_{i=1}^{N} \arctan\left(\frac{\text{dbh}_i}{\text{distance}_i}\right) \quad (1)$$

where N is the number of neighbors of species A in the neighborhood, dbh_i is the diameter at breast height of neighbor i and distance_i is the distance of neighbor i from the neighborhood center. The angular method may be most appropriate when the differential effects of neighbor tree species depend on how close those neighbor trees are to the focal (e.g. if differences are driven by competition for soil moisture or nutrients; Contreras et al. 2011). The user also has the option to exclude neighbor trees smaller than the focal tree from neighborhood density calculations, which may be appropriate if competition from smaller neighbors is expected to be negligible (Canham et al. 2004).

When quantifying neighborhoods we must also decide how to deal with trees whose neighborhoods overlap the stand boundary. As we do not have data on the entire neighborhoods of these trees, we must either exclude them from our analyses as focal trees, or estimate the characteristics of the missing portions of their neighborhoods. The *neighborhood_summary* function supports both of these approaches with the optional argument 'edge_correction'. When this option is set to true, neighborhood metrics are estimated for these partial neighborhoods by calculating the proportion of the neighborhood contained in the plot, and then multiplying the density values by the inverse of this proportion.

forestexplorR also contains some additional functions that can help the user acquire further neighborhood metrics. The *site_by_species* function takes the output of *neighborhoods* as input and creates a site-by-species matrix (presence/absence or abundance-weighted) where each site is a neighborhood. The resulting matrix could be used in conjunction with other datasets to calculate either phylogenetic or functional diversity (see R packages picante (Kembel et al. 2010) and FD (Laliberté and Legendre 2010, Laliberté et al. 2014)). The *tree_utm* function takes a mapping data frame and information on the orientation and location of stands (latitude and longitude of stand origins; see function documentation for details) and outputs universal transverse mercator (UTM) coordinates for each individual tree in the mapping data frame. The user could then connect their stem-mapped forest stand dataset with remote-sensing data and extract, for each individual tree, detailed topographical, crown-structure or normalized difference vegetation index (NDVI) data.

Calculating growth rates

The repeated censuses of stem-mapped forest stands enable the calculation of growth rate for each individual tree. forestexplorR provides two functions that implement these calculations using a tree measurement data frame. The *growth_summary* function extracts the earliest and most recent measurements of each tree and calculates an average annual growth rate across this period (both DBH change and basal area increment). It is not uncommon to obtain illogical negative annual growth rates due to measurement error of tree DBH in stem-mapped forest stands, and so the function returns a warning stating the number of trees that exhibited a negative growth rate. The output also contains, for each tree that had a non-negative annual growth rate, size-corrected average annual diameter and basal area increment growth rates. The formula for the size-corrected annual diameter growth rate is (Eq. 2):

$$\text{size corrected annual diameter growth rate} = \sqrt{\frac{\text{average annual DBH change}}{\text{first DBH measurement}}} \quad (2)$$

These transformed growth rate variables partially account for the non-linear relationship between tree size and growth rate and tend to more closely follow a normal distribution than raw growth rates, thereby enabling linear modeling of tree growth (Graham et al. 2021). The *detailed_growth* function works similarly to *growth_summary* but instead returns separate growth rates for each tree between each pair of consecutive stand censuses. The greater temporal resolution in the growth rates calculated by the *detailed_growth* function enables exploration of how tree growth rates have changed over time, perhaps in relation to climatic events.

Visualizing mapped stands

forestexplorR contains several functions for visualizing stem-mapped forest stands, which facilitate exploratory data analysis. The *size_dist* function uses a tree measurement data frame to graph the size-class distribution of one or more mapped stands. These distributions are frequently used in forest ecology and management to understand forest stand structure dynamics (i.e. stand succession). The ability to construct these size-class distributions for multiple stands simultaneously is valuable because *size_dist* outputs a multi-panel plot that facilitates comparison.

The *contour_plot* function creates an interactive contour map of the stand based on the value of any quantitative variable measured at multiple locations in the stand. For example, the *neighborhoods* and *neighborhood_summary* functions can be combined to calculate tree density at many locations within a given mapped stand, and then *contour_plot* can be used to generate a map of spatial variation in tree density across the stand. The resulting map could be used to identify suitable sampling locations for a study investigating effects of canopy density. If

the plotting variable for *contour_plot* is a neighborhood metric (e.g. tree density), it is important to recognize that trees whose neighborhoods overlap the stand boundary have not had their entire neighborhood sampled. In these cases, the user should implement the edge correction feature of *neighborhood_summary* in order to more accurately estimate neighborhood metrics for these edge trees. The *contour_plot* function could also be used to generate contour plots of edaphic variables measured on a grid in the stand, which could be used to estimate edaphic conditions for each tree (or any sampling location in the stand) at high spatial resolution.

To facilitate repeat censuses of stem-mapped forest stands or the identification of suitable sample sites, forestexplorR can also produce stem maps of stem-mapped forest stands. The *stand_map* function takes as input a mapping and a tree census data frame, each representing a single mapped stand, and returns a stem map of either the entire stand or a specified subsection of it, with individual trees represented by points with size proportional to their most recent DBH measurement.

Modeling tree growth and mortality

A major research opportunity afforded by stem-mapped forest stands is the ability to model tree growth or mortality as a function of the spatial structure of the surrounding neighborhood of trees. Tree growth and mortality models have provided insights on the processes underlying variation in tree dynamics, but common modeling techniques rely on the computationally-intensive process of maximum likelihood estimation to estimate their coefficients. This means that such models take a long time to fit and their use is therefore restricted to research teams with access to high-performance computing resources. To encourage broader use of neighborhood models, the forestexplorR functions *growth_model* and *mortality_model* implement a new regularized regression model of tree growth and mortality, respectively, that can be run in minutes on a personal laptop (Graham et al. 2021); less than 1% of the time required to fit other neighborhood models (Uriarte et al. 2004).

In these models, the independent variable of either tree growth rate (calculated using *growth_summary*) or tree mortality status (the mort column in the tree measurement data frame) is modeled as a weighted sum of the following variables: the size, distance and species identities of neighboring trees; species diversity, overall tree density and species-specific tree densities in the neighborhood (i.e. neighborhood metrics computed by *neighborhood_summary*); and any additional variables the user wishes to provide (e.g. abiotic data collected at the stand level). To correctly interpret the estimated coefficients of these models, it is necessary to understand the structure of the underlying design matrix, which differs from those used in most regression analyses. We are trying to model the growth or mortality of individual trees as a function of their neighborhood, and therefore it seems logical that our design matrix would contain a single row for each focal tree and columns that describe various aspects of its neighborhood. However, this structure poses a problem if we want to estimate coefficients for the size, distance and species identity of neighboring trees; most focal trees have multiple neighbors with differing sizes, distances and species identities and we therefore cannot describe neighbor size (or distance or species identity) using a single column. Our models resolve this problem by splitting the data for each focal tree across n rows in the design matrix, where each row represents an interaction between this focal tree and one of its n neighbors. This allows us to have a single column describing neighbor size and distance. For neighbor species identity, our design matrix contains a separate column for each potential neighbor species, and the species identity of the neighbor for a given row is indicated by a 1 in the relevant neighbor species column and 0 in every other neighbor species column. This means that a separate coefficient is estimated for each neighbor species and those coefficients represent the average difference in the dependent variable (growth or mortality) between when the neighbor is of that species versus when it is not of that species.

The structure of the growth and mortality models results in two estimated coefficients that provide information on how neighbors of each species influence the dependent variable (i.e. growth or mortality). The first coefficient is that described in the previous paragraph, which represents the average impact of the neighbor being of that species versus a different species. The second informative coefficient is that of the density of the neighbor species in the neighborhood (i.e. the species-specific density output by *neighborhood_summary*), which indicates the impact of increasing local density of that neighbor species on the dependent variable. In our experience, the species identity and density coefficients referring to the same neighbor species have always had the same direction. As a result, and for the purposes of clean visualization, in the case study below we summarize the effect of each neighbor species as the average of its two coefficients. However, we urge users to carefully consider the implications of this in their own data before following this protocol.

Previous work has shown this model to be capable of capturing the inherently non-linear mechanisms underlying tree growth through a transformation of the dependent growth variable that accounts for the major non-linear relationship; that between focal tree size and growth rate (Graham et al. 2021). This is why we encourage users to use the size-corrected growth rates output by *growth_summary* as the dependent variable in *growth_model*. forestexplorR does not currently offer a transformation of mortality status data to normalize model residuals of the logistic regression mortality model that *mortality_model* implements. However, we encourage users to experiment with transformations of mortality status data if they believe that will improve their model fits.

The efficiency of the growth and mortality models leads to a number of potential uses. First, the resulting model can be used to test hypotheses regarding the processes underlying variation in tree growth (Graham et al. 2021) or mortality. Second, the model can be run multiple times using different neighborhood sizes and plotting mean square error of the

resulting models to select a suitable neighborhood size for the model that is to be interpreted; a process that takes considerable time when using traditional modeling methods. Third, the regularization component of the model means that it is robust to correlated independent variables. This is because regularization shrinks the coefficients of variables that provide only marginal explanatory power down to zero and thereby automatically removes all but the most influential of a series of correlated variables during model fitting (Tibshirani 1996). This means that these models could be used to select the most important of several correlated climate variables to include in a more complex non-linear tree growth model (Uriarte et al. 2004) and thereby streamline the model building process (see Graham et al. 2021 for details). We also note here that one of the most fruitful use cases of forestexplorR may be to combine data from multiple stem-mapped stands in a single growth or mortality model. However, it is important to note that the *growth_model* and *mortality_model* functions do not currently account for spatial autocorrelation between the combined stands. For this reason, we recommend that users attempt to control for spatial autocorrelation by including explanatory variables in their model that capture the major axes of variation among stands, such as climatic conditions, latitude and elevation. We point users interested in quantifying spatial autocorrelation to other R packages that include this functionality, such as spdep and ncf.

To use the *growth_model* or *mortality_model* functions, a single data matrix containing neighborhood data and either growth or mortality data needs to be constructed. To create this matrix, the user must independently join the outputs of the neighborhoodsand *neighborhood_summary* functions, and then add growth data (output by *growth_summary*) or mortality data (from the tree measurement data frame). There is also an option to handle rare neighbor tree species (defined as species appearing as neighbors less than x times, where x is a user-provided argument) by grouping them together under the species identity of 'RARE'. Finally, if the user splits their data into a training and test set, these functions can conduct test set validation of the final model, which is important if the model is intended for use as a predictive tool (Tredennick et al. 2021). The output contains information on model fit and estimated coefficients, and contains the model object that can be used to make predictions when passed to stats::predict. Further details on the use of *growth_model* and *mortality_model*, as well as detailed examples of how to build the required design matrix, are provided in the case study below and in the 'modeling tree growth and mortality' vignette of forestexplorR.

An important step in building a neighborhood model is the selection of an appropriate neighborhood size. The neighborhood size that corresponds to neighborhood metrics of maximal explanatory power for tree growth or mortality will vary among ecosystems, forest successional stages and study species (Lutz 2015), and it is therefore strongly advised that a sensitivity analysis be conducted for neighborhood size. The forestexplorR function *select_nbhd_size* helps the user make an informed decision on the appropriate neighborhood size by fitting either *growth_model* or *mortality_model* using different neighborhood sizes and comparing model fit. The case study presented in the next section provides more details on how to use this function and select a neighborhood size.

Case study

To demonstrate the utility of forestexplorR, we investigate how tree growth is influenced by the species identity of neighboring trees at Mt Rainier National Park, WA, USA. This analysis uses the built-in datasets of forestexplorR, which originate from 15 rectangular 1 ha stem-mapped forest stands at Mt Rainier that are managed by the Pacific Northwest Permanent Sample Plot Network (Franklin et al. 2021). A similar analysis has already been conducted on these data with forestexplorR (Graham et al. 2021), but the analysis presented here differs by: focusing on a subset of the focal species (*Abies amabilis*, *Pseudotsuga menziesii*, *Tsuga heterophylla*), using the entire dataset for fitting each model (i.e. no data put aside for validation), and allowing neighborhood size to differ between focal species.

It is generally expected that tree growth rates are influenced by the species identity of neighboring trees, but the nature of these interactions can range from facilitative to competitive. Of particular interest to ecologists studying coexistence is how tree growth is influenced by conspecific neighbors, because reduced growth in the presence of conspecifics indicates negative density-dependent growth – an important mechanism for the maintenance of biodiversity (Canham et al. 2006, Fortunel et al. 2018). The biggest challenges in such analyses are: 1) calculating neighborhood metrics; 2) determining the appropriate neighborhood size; and 3) optimizing the typically complex models of tree growth. forestexplorR provides simple functions that address each of these challenges. The following paragraphs describe how this analysis was conducted with forestexplorR (for underlying code see: <https://github.com/sgraham9319/forestexplorR_case_study>).

The first step is to select the appropriate neighborhood size, which can be achieved using the *select_nbhd_size* function. This function applies *growth_model* multiple times, with each run using neighborhood metrics calculated according to a different neighborhood size, and calculates the mean square error of each model. For this case study, we tried neighborhood sizes of 2, 4, 6, 8, 10, 12, 14, 16, 18 and 20 m in radius. To ensure that models using different neighborhood sizes maintain equal sample size, *select_nbhd_size* excludes all trees whose neighborhood overlaps the stand boundary according to the largest neighborhood size to be tested (i.e. in this case, all trees within 20 m of a stand boundary were excluded). We also specified that species-specific tree densities in the neighborhoods should be calculated according to the proportional method of *neighborhood_summary* and included stand-level abiotic variables (i.e. the built-in dataset *stand_abiotic*) as covariates in the model. The stand-level abiotic data are not required for growth or mortality modeling, but they are a useful example of how easy it is to incorporate

potential explanatory variables beyond those calculated in forestexplorR.

This plots of mean square error versus neighborhood size output by *select_nbhd_size* are shown in Fig. 2A–C. It is expected that a larger neighborhood size will generally lead to a better fitting model because it contains more information on each neighborhood. There is a tradeoff, however, in that the larger the neighborhood size, the more focal trees need to be excluded from the analysis as a result of their neighborhood overlapping the stand boundary (unless their neighborhood metrics are corrected e.g. by the 'edge_handling' argument in *select_nbhd_size*). It is desirable to model growth of as many focal trees as we can without dramatically increasing mean square error because this gives us greater confidence that our findings will be applicable beyond the sampled stands. Therefore, we recommend choosing the neighborhood size that, based on visual observation, lies at the 'elbow' of these plots, which represents where further increases in neighborhood size lead to only small improvements in model fit. Following this protocol, we selected neighborhood sizes of: *A. amabilis* = 12 m, *P. menziesii* = 10 m, *T. heterophylla* = 12 m.

Next, we fit the final model for each species using the selected neighborhood size and including all trees whose neighborhood of this size does not overlap a stand boundary. Each model outputs two coefficients that are indicative of the effect of a particular neighbor species on the focal species; one for the neighbor species identity variable and another for the neighborhood metric of that neighbor species' density. To obtain a single value for the effect of each neighbor species on the growth of each focal species, we averaged these two related coefficients – these are the values shown in Fig. 2D–F. This coefficient averaging procedure is appropriate because all predictor variables in this model are rescaled in order to be comparable and, in this case, the two coefficients corresponding to the same neighbor species never differed in direction. However, users should confirm this is also the case in their data before using this coefficient averaging procedure.

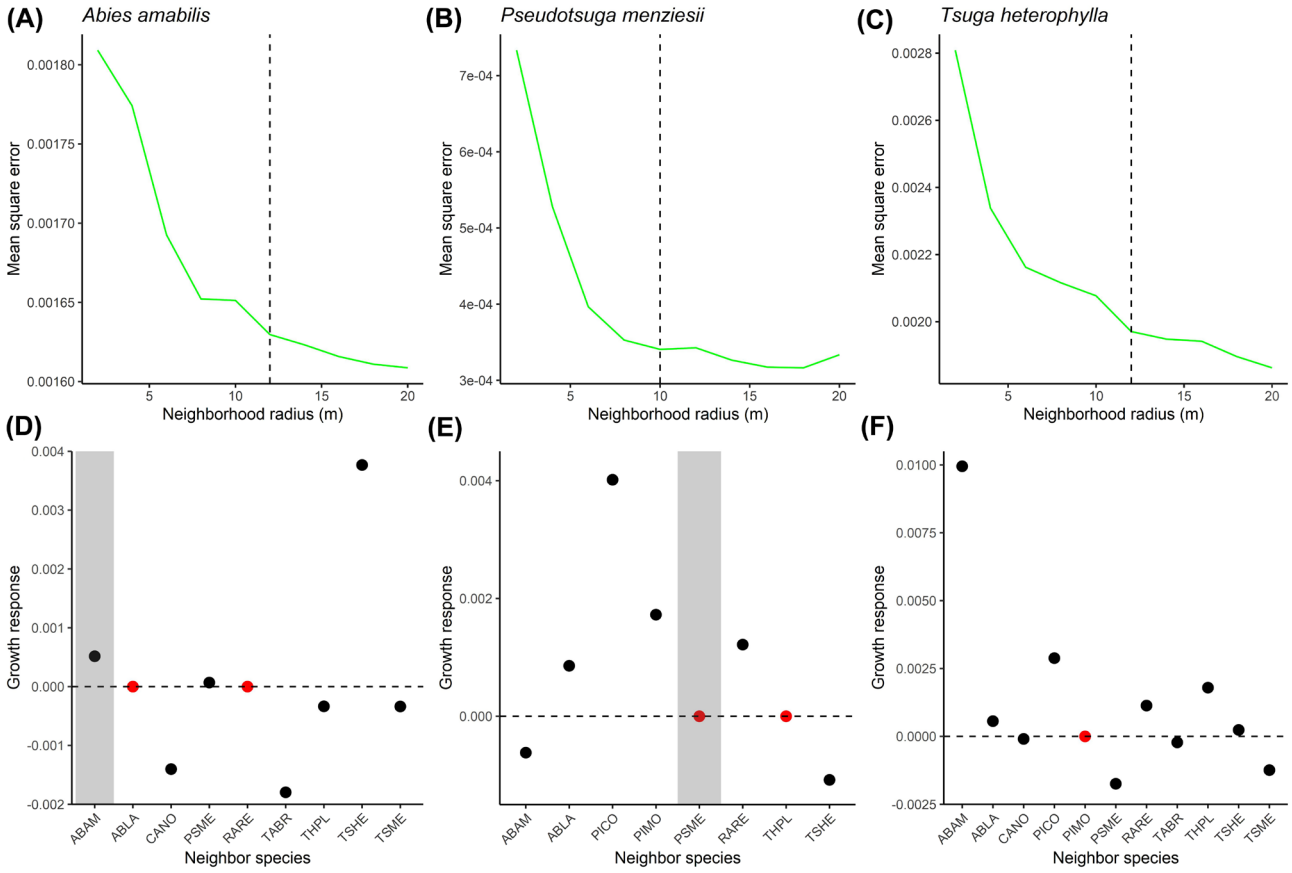

Figure 2. Using forestexplorR to investigate how the growth rates of common tree species in Mt Rainier National Park are influenced by the species identity of neighboring trees. (A–C) show the relationship between model fit (mean square error) and size of tree neighborhoods, with dashed vertical lines indicating the neighborhood size selected for further analysis. (D–F) show the growth response of each focal species to neighbors of particular species, with conspecific neighbors shaded in gray and the points for neighbor species whose effect was dropped during model fitting colored in red. Plots are aligned according to the focal species they represent: (A and D) = *A. amabilis*, (B and E) = *P. menziesii*, (C and F) = *T. heterophylla*. Codes for competitor species are defined as follows: ABAM = *A. amablis*, ABLA = *Abies lasiocarpa*, CANO = *Callitropsis nootkatensis*, PICO = *Pinus contorta*, PIMO = *Pinus monitcola*, PSME = *P. menziesii*, RARE = all trees of species represented by < 100 interactions, TABR = *Taxus brevifolia*, THPL = *Thuja plicata*, TSHE = *T. heterophylla*, TSME = *Tsuga mertensiana*.

We find that *A. amabilis* grows more quickly in the presence of conspecifics, indicating positive density-dependent growth, but that *A. amabilis* responds even more positively to neighboring *T. heterophylla*. By contrast, *P. menziesii* and *T. heterophylla* do not exhibit either positive or negative density-dependent growth, but instead grow particularly quickly in the presence of *Pinus contorta* and *A. amabilis* neighbors respectively. These findings are mostly, but not entirely consistent with those reported previously (Graham et al. 2021), but the present models are likely more accurate because they use a larger dataset and a neighborhood size tailored to the focal species.

Discussion

forestexplorR equips forest ecology researchers of all skillsets with the tools they need to explore and analyze rectangular stem-mapped forest stand datasets, and thereby has the potential to spur exciting new ways of utilizing this rich data source. Widespread use of stem-mapped forest stand datasets is currently hampered by the high programming and computational demands involved, particularly in the quantitative description of neighborhoods, selecting an appropriate neighborhood size and fitting neighborhood models of tree performance. forestexplorR removes these barriers by providing flexible and user-friendly functions for describing neighborhoods, selecting neighborhood size and implementing a rapid-fitting neighborhood model of tree growth or mortality (Graham et al. 2021) that can be used to investigate species interactions. Moreover, by requiring only the data types common to all stem-mapped forest stands (species identity, location, DBH measurements), forestexplorR is compatible with all rectangular stem-mapped forest stand datasets. It will therefore now be easier to combine data from multiple stem-mapped forest stands in a single analysis, which could greatly expand the utility of networks of plots distributed over environmental gradients such as those of the Pacific Northwest Permanent Sample Plot networks (Franklin et al. 2021), the NEON Distributed and Tower Plots (NEON 2012), the RAINFOR and AfriTRON networks curated by ForestPlots.net (Lopez-Gonzalez et al. 2009, 2011), and likely many others. This also means that stem-mapped forest stand datasets are now accessible to all researchers, regardless of programming experience and access to computing resources, which should greatly expand the diversity of projects using these datasets.

forestexplorR could be used to replicate and strengthen tests of previously posed research questions and encourage the development of new research directions relating to stem-mapped forest stand data. The package allows a variety of spatially-explicit neighborhood metrics to be calculated with ease, and their variation across stands to be explored through visualizations. Consequently, it allows any researchers collecting data in stem-mapped forest stands to include neighborhood metrics as covariates in their models. This could lead to investigations of tree growth and mortality responses to climate change (Buechling et al. 2017) and competitive interactions (Fortunel et al. 2016) being replicated in a greater diversity of systems. Further, the inclusion of spatially-explicit neighborhood metrics could permit deeper investigation of spatial variation in soil microbial communities (Otsing et al. 2021), and improve the accuracy of models predicting carbon storage dynamics (Martínez Cano et al. 2020, Ma et al. 2021). We also hope, that by encouraging a greater diversity of researchers to use stem-mapped forest stand data, forestexplorR will pave the way for many novel and creative research questions to be asked with this data type.

This initial version of forestexplorR is designed to facilitate what we expect to be the most common uses of stem-mapped forest stand data, but we plan to expand its functionality over time with input from the research community. Some potential expansions include: 1) the ability to extract average annual growth rates over a user-defined time period, and 2) functions that connect to other packages or publicly available datasets to pull in stand-level climatic data (e.g. WorldClim data obtained through raster::getData) or tree-level topographical data (e.g. lidar). However, as the code underlying forestexplorR is openly available on GitHub, we also hope that the growing community of researchers working with stem-mapped forest stand data will contribute improvements and additions that we have not considered.

The forestexplorR package can be downloaded at <https://github.com/sgraham9319/forestexplorR> and a website containing detailed vignettes for the common uses of the package is available at <https://sgraham9319.github.io/forestexplorR/index.html>. Users of the package are encouraged to report difficulties and suggest improvements through pull requests and by posting issues on GitHub.

To cite forestexplorR or acknowledge its use, cite this Software note as follows, substituting the version of the application that you used for 'ver. 1.0.0':

Graham, S. I. et al. 2022. forestexplorR: an R package for the exploration and analysis of stem-mapped forest stand data. – Ecography 2022: 1–10 (ver. 1.0.0).

Acknowledgements – We thank the handling editor for invaluable feedback that has greatly improved the functionality of the package and the clarity of this manuscript. We also thank Joe Ammirati, Michele Buonanduci, Tony Cannistra, Brian Harvey, Aji John, Rubén Delgado Manzanedo, Kavya Pradhan, Meera Lee Sethi, Hunter Stanke, Kristiina Visakorpi, the ETH Zürich Plant Ecology Group and the UW eScience Incubator program for support and feedback throughout this project. The tree growth data included in the package were provided courtesy of the Pacific Northwest Permanent Sample Plot Program, in partnership with the HJ Andrews Experimental Forest and Long Term Ecological Research (LTER) program, which are administered cooperatively by the USDA Forest Service Pacific Northwest Research Station, Oregon State Univ. and the Willamette National Forest.

Funding – During this project, SIG was funded by the Dept of Biology at the Univ. of Washington, and AR was funded through grants from the Alfred P. Sloan Foundation and the Gordon and Betty Moore Foundation to the Univ. of Washington eScience Inst. The Pacific Northwest Permanent Sample Plot Program is funded by the National Science Foundation, grant no. LTER8 DEB-2025755 (2020–2026) and LTER7 DEB-1440409 (2012–2020).

Author contributions

Stuart Graham: Conceptualization (lead); Data curation (lead); Formal analysis (lead); Funding acquisition (lead); Investigation (lead); Methodology (lead); Project administration (lead); Software (lead); Validation (equal); Visualization (lead); Writing – original draft (lead); Writing – review and editing (supporting). **Ariel Rokem**: Conceptualization (supporting); Data curation (supporting); Formal analysis (supporting); Investigation (supporting); Methodology (supporting); Software (supporting); Supervision (supporting); Writing – review and editing (supporting). **Janneke Hille Ris Lambers**: Project administration (supporting); Supervision (lead); Validation (equal); Writing – review and editing (lead).

Transparent peer review

The peer review history for this article is available at <https://publons.com/publon/10.1111/ecog.06223>.

Data availability statement

Data are available from the Zenodo Digital Repository: <https://doi.org/10.5281/zenodo.6634609> (Graham et al. 2022).

References

Buechling, A. et al. 2017. Climate and competition effects on tree growth in Rocky Mountain forests. – J. Ecol. 105: 1636–1647.

Buonanduci, M. S. et al. 2020. Neighborhood context mediates probability of host tree mortality in a severe bark beetle outbreak. – Ecosphere 11: e03236.

Canham, C. D. et al. 2004. A neighborhood analysis of canopy tree competition: effects of shading versus crowding. – Can. J. For. Res. 34: 778–787.

Canham, C. D. et al. 2006. Neighborhood analyses of canopy tree competition along environmental gradients in New England forests. – Ecol. Appl. 16: 540–554.

Condit, R. 1995. Research in large, long-term tropical forest plots. – Trends Ecol. Evol. 10: 18–22.

Contreras, M. A. et al. 2011. Evaluating tree competition indices as predictors of basal area increment in western Montana forests. – For. Ecol. Manage. 262: 1939–1949.

Davies, S. J. et al. 2021. ForestGEO: understanding forest diversity and dynamics through a global observatory network. – Biol. Conserv. 253: 108907.

Fortunel, C. et al. 2016. Functional trait differences influence neighbourhood interactions in a hyperdiverse Amazonian forest. – Ecol. Lett. 19: 1062–1070.

Fortunel, C. et al. 2018. Topography and neighborhood crowding can interact to shape species growth and distribution in a diverse Amazonian forest. – Ecology 99: 2272–2283.

Franklin, J. F. et al. 2021. Long-term growth, mortality and regeneration of trees in permanent vegetation plots in the Pacific Northwest, 1910 to present ver 18. – Environmental Data Initiative, <https://doi.org/10.6073/pasta/45a1f16b3d8ecd0585a2d2e115c07d41>, accessed 24 September 2021.

Graham, S. I. et al. 2021. Regularized regression: a new tool for investigating and predicting tree growth. – Forests 12: 1283.

Graham, S. I. et al. 2022. Data from: forestexplorR: an R package for the exploration and analysis of stem-mapped forest stand data. – Zenodo Digital Repository, <https://doi.org/10.5281/zenodo.6634609>.

Kembel, S. W. et al. 2010. Picante: R tools for integrating phylogenies and ecology. – Bioinformatics 26: 1463–1464.

Kim, A. Y. et al. 2021. The forestecology R package for fitting and assessing neighborhood models of the effect of interspecific competition on the growth of trees. – Ecol. Evol. 11: 15556–15572.

Kunstler, G. et al. 2016. Plant functional traits have globally consistent effects on competition. – Nature 529: 204–207.

Laliberté, E. and Legendre, P. 2010. A distance-based framework for measuring functional diversity from multiple traits. – Ecology 91: 299–305.

Laliberté, E. et al. 2014. FD: measuring functional diversity from multiple traits, and other tools for functional ecology. – R package ver. 1.0-12, <https://cran.r-project.org/package=FD>.

Lepore, M. et al. 2019. fgeo: analyze forest diversity and dynamics. R package ver. 1.1.4. – <https://CRAN.R-project.org/package=fgeo>.

Lopez-Gonzalez, G. et al. 2009. ForestPlots.net database. – <www.forestplots.net>, accessed 12 April 2022.

Lopez-Gonzalez, G. et al. 2011. ForestPlots.net: a web application and research tool to manage and analyse tropical forest plot data. – J. Veg. Sci. 22: 610–613.

Lutz, J. A. 2015. The evolution of long-term data for forestry: large temperate research plots in an era of global change. – Northw. Sci. 89: 255–269.

Ma, L. et al. 2021. High-resolution forest carbon modelling for climate mitigation planning over the RGGI region, USA. – Environ. Res. Lett. 16: 045014.

Martínez Cano, I. et al. 2020. Allometric constraints and competition enable the simulation of size structure and carbon fluxes in a dynamic vegetation model of tropical forests (LM3PPA-TV). – Global Change Biol. 26: 4478–4494.

NEON (National Ecological Observatory Network) 2012. Vegetation structure (DP1.10098.001), RELEASE-2022. – <https://doi.org/10.48443/re8n-tn87>, accessed 12 April 2022.

Otsing, E. et al. 2021. Tree species richness and neighborhood effects on ectomycorrhizal fungal richness and community structure in boreal forest. – Front. Microbiol. 12: 567961.

Rouvinen, S. and Kuuluvainen, T. 1997. Structure and asymmetry of tree crowns in relation to local competition in a natural mature Scots pine forest. – Can. J. For. Res. 27: 890–902.

Stanke, H. et al. 2020. rFIA: an R package for estimation of forest attributes with the US Forest Inventory and Analysis database. – Environ. Model. Softw. 127: 104664.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. – J. R. Stat. Soc. B 58: 267–288.

Tredennick, A. T. et al. 2021. A practical guide to selecting models for exploration, inference and prediction in ecology. – Ecology 102: e03336.

Uriarte, M. et al. 2004. A neighborhood analysis of tree growth and survival in a hurricane-driven tropical forest. – Ecol. Monogr. 74: 591–614.

Wiegand, T. et al. 2017. Spatially explicit metrics of species diversity, functional diversity and phylogenetic diversity: insights into plant community assembly processes. – Annu. Rev. Ecol. Evol. Syst. 48: 329–351.