# Title: Quantifying spatial niche differentiation in tropical trees

# Principal objective

Quantify the extent to which tropical trees niche-differentiate by specializing on different local environments associated with soil conditions, and connect this niche structure to species traits

# Subgoals

1. Describe and enumerate the “soil niches” in a forest.
2. Assess whether our inferred soil niches correlate with measured soil indicators (nutrient concentrations). If so, this lends credence to our inference.
3. Assess whether our inferred soil niches correlate with species traits. If so, this provides a link between species traits and the types of soil resources to which they specialize.
4. Obtain an affinity matrix to be compared with D’Andrea, Gibbs & O’Dwyer 2020 to make inferences about whether non-neutral behavior should be observed at the community level.

# Methods

Data sets

Words about the FDP on BCI. Words about nutrient dataset. Words about the traits

Data filters

We focus our analysis on adult trees by implementing a 10 cm diameter at breast height cutoff, reasoning that selection by the local environment largely operates between germination and recruitment into adult stage. This leaves circa 21,000 trees on the FDP spanning about 150 species. We also remove species that may not be found near each other due to rarity rather than disparate environmental niches, by applying the following abundance cutoff. Let p(n, d\*) be the expected number of close neighbors between two randomly distributed species of abundance n. We then restrict our analysis to species with n >= n\_min, where p(n\_min, d\*) = 1. Specifically, n\_min = 1 / sqrt(F(d\*, Lx, Ly)), where F(d\*, Lx, Ly) is the cumulative probability that a random pair of points in a rectangle of sides Lx and Ly are within distance up to d\* (CITATION). For our cutoff distance of 20 m and given our 1000 m x 500 m plot, this corresponds to a minimum abundance of 20 trees (see Supplementary Information).

Finding species groups

Our analysis is based on finding species that tend to occur near each other. First, we determined the number of close neighbors to each tree within the 50ha forest dynamics plot (FDP) on BCI. A close neighbor was defined as a tree within a cutoff distance (d\_c) of the focal tree. We used a cutoff distance of 20 m as this is the scale normally associated with plant-plant interactions (CITATIONS), but results were not sensitive to this choice (Supplementary Information). From this we obtained an adjacency matrix, whose elements specify the number of close neighbors between each species pair. Treating species as nodes in a graph with edges weighted by the corresponding elements of the adjacency matrix, we then use R package *igraph* to find modules, i.e. subsets of densely connected nodes (species) that are relatively isolated from all other nodes. The algorithm works by hierarchically grouping nodes into modules until no further grouping can be made without loss of modularity (mathematically defined as the fraction of edges found within groups minus the expected fraction if edges were randomly assigned). The outcome is therefore a maximally modular set of N groups of species found in spatial proximity.

Soil types

We hypothesize that the spatial segregation among the species groups reflects their affinity for different local soil conditions. Therefore, to each species group, we impute a matching set of local soil conditions (“soil types”, treated as a categorical variable) where group members preferentially recruit. This one-to-one mapping between species groups and soil types is then interpreted as the “soil niches” on BCI.

We infer the spatial distribution of soil types on the FDP via kernel density estimation using R package *sparr* (CITATION): assuming soil types are distributed continuously on the FDP, we use the empirical local density of trees of each group to find the smoothed probability density of each corresponding soil type across the landscape. Each 20 m x 20 m quadrat in the FDP is then assigned a probability of containing each of the N soil types.

Nutrient analysis

If our species groups reflect preferential recruitment in soil with different local conditions, there must be a spatial association between local nutrient levels and the probability density of our putative soil types. We test for this association in two ways. First, we calculate the Pearson correlation coefficient between the local levels of each nutrient and the inferred density of each soil type. If niche differentiation by specializing to different soil conditions is a dominant process on BCI, then the nutrient correlation profiles of each species group should be quite distinctive.

Second, we use supervised machine learning to measure the extent to which local nutrients predict the soil type. High predictive power would signify a tight link between nutrients and our inferred soil types. First, we label each 20m x 20 m quadrat with the soil type with highest local probability as per our kernel density estimator. Next, we train a decision tree / rule-finding algorithm to predict the soil type based on local nutrient levels. We used the C5.0 classifier in R package *caret*. This algorithm balances interpretability of results with accuracy and predictive power (PRAISE C5.0). The classifier iteratively splits the data (quadrats) based on numerical thresholds in the covariates (nutrient levels). Each split maximizes reduction of entropy in the labels (soil types) of the resulting data subsets. The C5.0 classifier converges on a set of rules/splitting decisions that allows it to guess the soil type of each quadrat based on measured local nutrients. We avoid fitting the noise by repeatedly training the learner on a subset of the quadrats and then test it on the remainder of the data (repeated cross-validation, CITATION). We asses quality of predictions using Cohen’s Kappa (CITATIONS), an index that balances sensitivity (true positives) with specificity (true negatives). This index can range from negative values (worse-than-chance predictions) to 1 (perfect predictions).

Trait analysis

We examined the link between our inferred groups and plant function using three key trait syndromes: vital rates, wood density, and leaf structure. Vital rates included annual DBH growth averaged across the fastest growing trees in each species and inter-census mortality averaged across the slowest growing individuals. High vital rate scores indicate fast growth in favorable environments and high mortality in unfavorable environments. Wood density was measured as average wood specific gravity after drying at 60C or 100C. High wood density scores indicate long-living trees with high investment in structural support. Foliar structure traits included leaf area, dry matter content, mass per area, thickness, and toughness. High leaf trait scores indicate a “conservative” leaf strategy with long leaf lifespan (Diaz et al 2015). The full list of traits within each syndrome is described in the Supplementary Information. Details of how traits were measured can be found in (CITATION). As traits within each syndrome were highly correlated, we used PCA to reduce dimensionality (Supplementary Information). In all cases, the first principal component explained more than XX% of the variance across the individually measured traits. For each trait category, we performed pairwise Mann-Whitney tests of the distribution of the first principal component across species groups.

Validation process

To verify that our method successfully captures the niche structure of species segregating by preferences for local environments, we validated our method on simulation outcomes of a spatially explicit birth-death process on a heterogeneous landscape where recruitment of plants within a certain group is more likely to occur in sites with matching soil types than in mismatching environments. The model is parametrized to replicate average community size and richness observed on the FDP at BCI. Temporal iterations correspond to death and replacement of about 10% of the community, as empirically observed between censuses on BCI. The environmental landscape is generated as a binned Gaussian random field with autocorrelation scales matching the autocorrelation scale of our inferred landscape. The number of bins n is a free parameter of the model, corresponding to the number of different soil types and corresponding plant groups (niches). The degree of niche differentiation between plant groups is represented by the odds of plant recruitment in matching soil type relative to recruitment in other soil types. We set these odds as the second free parameter in our model, theta. Further details of the model are provided in the Supplementary Information.

The R code for our validation model and our clustering method and data analysis is available on GitHub (LINK) upon publication.

# Results

Validation process

We applied our method for grouping species based on spatial proximity to simulation outcomes of our model spanning different numbers of groups, n = 3, 4, 5, … 15, and levels of niche differentiation within the groups, theta = 1, 5, 10. Results indicate that our method successfully captures the group structure of the plants (Figure ?? in the Supplementary Information). Accuracy is higher when the number of groups is small and niche differentiation is high. This was expected, as all methods for grouping datapoints will run into limited statistical power as the number of groups increases and as the groups become less distinctive. Indeed, at low niche differentiation or when the number of plant groups is large, our method tends to underestimate the true number of groups. However, the groups that it identifies tend to recruit in soil types that occur near each other in our simulated soil landscape. Since spatial associations between soil types in nature is likely to reflect soils with similar nutrient composition, we conclude that our method will either correctly identify the groups, or find group “families”, i.e. groups with similar strategies whose internal substructure the method cannot resolve as sample sizes are finite and the data is noisy.

Next, we compare our species niches with the spatial distribution of soil nutrients in the FDP (data from CITATION). If there is an association between our species groups

1. We can perform network-based analyses to determine how tight the classification is, and whether it is statistically significant. For example, I found that the *modularity* of this three-niches graph is significant by randomizing the species labels of the FDP trees, finding communities among the resulting graphs, and calculating the modularity of the results.
2. John checked that this result is also robust to using different cutoffs for the interaction distance and minimum abundance. Using distance cutoff from 2 to 40 meters, abundance threshold from 0 to 400, we consistently find three clusters, with some exceptions of 2 and 4 clusters. Upon visual inspection, the predicted soil landscape looks similar over a wide range of distance (10-30 m) and minimum abundance. We ultimately use the parameters: distance = 10m, abundance cutoff = 50 for the following analysis.

# Figures

Qr code

Description automatically generated

Figure 1: A. Spatial distribution of the trees in the Forest Dynamic Plot on Barro Colorado Island, Panama. Trees are colored by their respective group. B. Adjacency matrix showing species pairs found to occur in close proximity more often than expected from a Poisson process (dark pixels). Arranging species by group membership reveals a block-diagonal strcture to the matrix.

A picture containing graphical user interface

Description automatically generated

Figure 2: A. Density field for each of the four species groups (color names), based on kernel density estimation. B. Corresponding inferred soil types on a grid of 20m x 20m cells across the FDP.

Chart

Description automatically generated

Figure 3: A. Cells of Fig 2B shown in the plane formed by the first two principal components of nutrient concentration, colored by their inferred soil type. A clear separation between the inferred soil types is visible, indicating distinct associations with nutrient levels. B. Correlation between the spatial density of each species group and local nutrient levels.

Chart, scatter chart

Description automatically generated

Figure 4: Distribution of BCI species on the plane formed the first principal component of vital rates (x axis) and wood density (y axis), colored by their respective group. Species with higher vital rates have lower wood density. In addition, there is a visible separation between species associated with high nutrient levels (red, green) and those associated with low nutrient levels (blue, yellow).

# Tables

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **vital rates** | | | **wood density** | | | **leaf traits** | | |
|  | red | green | blue | red | green | blue | red | green | blue |
| green | 0.164 |  |  | 0.182 |  |  | 1 |  |  |
| blue | 0.164 | 0.00622 |  | 0.381 | 0.00736 |  | 0.223 | 1 |  |
| yellow | 0.0000396 | 0.00000603 | 0.088 | 0.0121 | 0.0000042 | 0.381 | 0.368 | 1 | 1 |

# Supplementary Information

Table

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart, bar chart, histogram

Description automatically generated

Background pattern

Description automatically generated

Calendar

Description automatically generated

Chart, bar chart, histogram

Description automatically generated

Chart, box and whisker chart

Description automatically generated

Timeline

Description automatically generated

Diagram

Description automatically generated