Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2014, **5,** 1214–1224 doi: 10.1111/2041-210X.12295

A method to detect subcommunities from multivariate spatial associations

Anton J. Flügge[1,2,3], Sofia C. Olhede[2,4] and David J. Murrell[1,2,3*]

[1]*Department of Genetics, Evolution and Environment, University College London, Gower Street, London WC1E 6BT, UK;*
[2]*Centre for Mathematics and Physics in the Life Sciences and Experimental Biology, University College London, London, UK;*
[3]*Centre for Biodiversity and Environment Research, University College London, Gower Street, London WC1E 6BT, UK; and*
[4]*Department of Statistical Science, University College London, Gower Street, London WC1E 6BT, UK*

Summary

1. Species are seldom distributed at random across a community, but instead show spatial structure that is determined by environmental gradients and/or biotic interactions. Analysis of the spatial co-associations of species may therefore reveal information on the processes that helped to shape those patterns.

2. We propose a multivariate approach that uses the spatial co-associations between all pairs of species to find subcommunities of species whose distribution in the study area is positively correlated. Our method, which begins with the patterns of individuals, is particularly well-suited for communities with large numbers of species and gives rare species an equal weight. We propose a method to quantify a maximum number of subcommunities that are significantly more correlated than expected under a null model of species independence.

3. Using data on the distribution of tree and shrub species from a 50 ha forest plot on Barro Colorado Island (BCI), Panama, we show that our method can be used to construct biologically meaningful subcommunities that are linked to the spatial structure of the plant community. As an example, we construct spatial maps from the subcommunities that closely follow habitats based on environmental gradients (such as slope) as well as different biotic conditions (such as canopy gaps).

4. We discuss extensions and adaptations to our method that might be appropriate for other types of spatially referenced data and for other ecological communities. We make suggestions for other ways to interpret the subcommunities using phylogenetic relationships, biological traits and environmental variables as covariates and note that subcommunities that are hard to interpret may suggest groups of species and/or regions of the landscape that merit further attention.

Key-words: Barro Colorado Island, community ecology, spatial pattern, niche theory, neutral theory, clustering, point process

Introduction

Understanding the processes that underpin observed patterns of biodiversity and how functionally similar species coexist in close spatial proximity are among the primary challenges in ecology (Hardin 1960; Wright 2002). Biodiversity may bolster ecosystem stability and productivity (Isbell, Polley & Wilsey 2009; Cardinale *et al.* 2012) and is therefore an important aspect in the environmental services that ecosystems provide to society. However, land-use changes (Brooks *et al.* 2002), growing populations (Williams 2013) and climate change (Bellard *et al.* 2012) may all threaten biodiversity and ecosystems in general. Understanding the ecological processes that both create and maintain high biodiversity is important for protecting diverse ecosystems.

Through the Center for Tropical Forest Science and the ForestGEO initiative (2013), there are now data available on multiple large-scale forest plots for which all trees and shrubs are individually mapped and identified to species level. Spatial analyses of these individual-based tree data sets have often focussed on univariate approaches to investigate the spatial distributions and attempt to quantify the variation in species distributions that can be explained by other processes/factors such as abundance (e.g. Condit *et al.* 2000); recent changes in local abundance (Flügge, Olhede & Murrell 2012); dispersal mechanism (e.g. Muller-Landau & Hardesty 2005); conspecific density dependence (e.g. Bagchi *et al.* 2011) and habitat association (e.g. Harms *et al.* 2001; Itoh *et al.* 2010; Ledo *et al.* 2013). However, other studies also consider pairs of species and multivariate patterns to look for assemblages of species that might be related to different habitat types (e.g. Martínez *et al.* 2010; Wang *et al.* 2010; Lan *et al.* 2012; Luo *et al.* 2012; Wiegand *et al.* 2012; Baldeck *et al.* 2013; Punchi-Manage *et al.* 2013) as well as quantify the roles of habitat association and dispersal limitation in determining species area relationships, individual species area relationships (ISARS) and the spatial variation of beta diversity measures (Wiegand *et al.* 2007; Wang *et al.* 2011; Cheng *et al.* 2012; Rajala & Illian 2012; Wang *et al.* 2013).

*Correspondence author. E-mail: d.murrell@ucl.ac.uk

A multivariate spatial approach is useful because it can be used to highlight groups of species that are found together more often than expected by chance. Once these groups of species have been identified, it is possible to investigate the processes that are driving their spatial association. Recent studies have begun to explore the potential of multivariate spatial methods to uncover new biological insight by starting with locations in the landscape and using methods to group similar areas (and assemblages of species) together (e.g. Baldeck *et al*. 2013; Punchi-Manage *et al*. 2013). For example, Punchi-Manage *et al*. (2013) use the Bray–Curtis dissimilarity measure combined with a multivariate regression tree (MRT) analysis to show five distinct habitat types emerge across all life-history stages of a mixed dipterocarp forest in Sri Lanka. The added value of this approach is the ability to quantify the contribution of environmental covariates to the variation in local species composition, and in this investigation, it was estimated that approximately 25% of the variation could be attributed to topographic variables. In a separate study, Baldeck *et al*. (2013) also used the Bray–Curtis dissimilarity analysis of species composition for quadrats at the 20 m scale, but instead used principal coordinates of neighbour matrices (PCNM) to model spatial structure in the variation of community composition among quadrats (see also Borcard & Legendre 2002; Legendre *et al*. 2009). This variation was partitioned into portions explained by soil, topographic and spatial variables. Similar to Punchi-Manage *et al*. (2013), the results for eight separate mixed forests showed the soil and topographic covariates could explain 19–39% of the variation, but that spatial processes such as dispersal limitation and other unmeasured environmental variables could explain a further 19–37% of the variation. Both studies highlight the importance of small-scale environmental variation in structuring species-rich plant communities, but also that biological processes are likely to play an important role.

Although the location- or quadrat-based approach has clear benefits, one could alternatively start by focusing on individuals and consider the average biotic neighbourhood of an individual of a particular species, the so-called plant's-eye-view (Turkington & Harper 1979). This neighbourhood approach summarizes the spatial co-association of pairs of species by comparing the observed mean density of neighbours of one species around individuals of another species with that expected if the two species were arranged across the landscape independently of each other (e.g. Wiegand *et al*. 2012). These summaries of spatial co-association could be used to detect subcommunities of species that have similar spatial associations to one another. Any emergent subcommunities would naturally show spatial correlation and interpretation of their make-up, as in other multivariate pattern analyses, could be based upon both biological and abiotic variables.

The neighbourhood-based approach is strongly linked to spatial ecological theory for population and community dynamics where the spatial co-association measures can be state variables (e.g. Murrell 2010) and this may be used to better understand the interspecific patterns under investigation. For example, theory has shown that strong interspecific competition should lead to negative spatial co-associations as heterospecific individuals are removed from neighbourhoods (Murrell, Purves & Law 2001). On the other hand, positive spatial associations can occur if species interactions are positive (Callaway 1995), or if species have shared preferences in habitat. Limited dispersal, a strongly stochastic process that generates spatial structure, may also lead to some strong positive or negative co-associations, but overall, one would expect it to create spatial independence between species, and it may help to obscure the signal of spatial associations between pairs of species. Indeed, Wiegand *et al*. (2012) found that once the effects of habitat association are removed, the proportion of species-pair spatial co-associations that can be distinguished from independent may be quite low for very diverse communities (including the BCI plot). The authors attributed this to the influence of dispersal limitation combined with often low local abundances leading to a high level of statistical noise that they referred to as a dilution effect.

In what follows we outline a method for grouping together species according to their interspecific (bivariate) spatial co-associations with an example where the interpretation of the groups of species is based largely upon environmental niches, although the reader should note that other covariates such as species traits could also be used to understand the membership of the subcommunities. Our method has three steps. First, the interspecific spatial co-associations are quantified, taking into account differences in abundance and within-species spatial distribution. The second step involves using a well-established clustering algorithm to group species together that have similar spatial co-associations. The final step is to then create a map denoting locations in the landscape where each subcommunity dominates.

To illustrate our approach, we use the Barro Colorado Island (BCI) forest dynamics plot (Hubbell, Condit & Foster 2005), which allows us to contrast our results with those of previous studies. In particular, we compare our results and methods to previous work on habitats at BCI (Harms *et al*. 2001; Kanagaraj *et al*. 2011) and other ForestGEO-CTFS sites (Baldeck *et al*. 2013; Punchi-Manage *et al*. 2013). While Harms *et al*. (2001) analyse the spatial pattern of individual species and test for correlations with habitats defined on the basis of environmental variables, the other three studies take the joint distribution of all species into account and are based on methods that compute the dissimilarity of the species composition at different spatial locations. Our method is complementary to those latter methods, as it does not focus on the species composition at certain locations, but on the co-associations between species across space – or in other words, where alternative methods average across species to find spatial locations with similar species assemblages, we average across space to find groups of species that co-associate with one another.

Materials and methods

DATA

We use the data from the Barro Colorado Island (BCI) 50 ha long-term forest dynamics plot in Panama (see Condit 1998; Hubbell *et al.* 1999; Hubbell, Condit & Foster 2005). The forest plot at BCI was established in 1980, and from 1985 onward; complete censuses of all trees and shrubs above 1 cm diameter at breast height (DBH) have been repeated every five years. All individuals are identified to species level, and their position and size are recorded in every census. Each individual is classified as adult or juvenile by comparing it to a species-specific DBH criteria based on estimates by Robin Foster on the typical sizes when species become reproductive (R. Foster, unpublished data). Our analysis includes 141 shrub and tree species (out of 301 species), namely those with at least 10 adults and 10 juveniles in the most recent 2010 census. Species with fewer individuals are excluded because in this instance, it is not possible to estimate reliable co-association values for both the juvenile and the adult populations. This criterion excludes very rare species, species that do not reproduce in the plot itself, and small shrub species for which all individuals included in the census are classified as adults. Considering only adults would lead to the inclusion of another 34 species (26 of which are shrub species where all individuals in the census are classified as adults). In our interpretation of the results, we also use the shade-tolerance indices (available for 124 of the 141 species) from Comita *et al.* (2010) to compare the species in different groups. In total, the analyses that follow include 153 634 trees and shrubs (35 156 adults and 118 478 juveniles) out of 207 259 individuals above 1 cm DBH in the 2010 census (see Appendix S1 for list of species, abundances, shade-tolerance indices and Robin Foster estimates).

ANALYSIS

Bivariate co-association measure

As a measure of the spatial co-association of two species a and b, we use the bivariate version of the Ripley's $K_{a,b}(r)$ standardized by the neighbourhood area (Lotwick & Silverman 1982; Wiegand & Moloney 2004). This means the expected value of the standardized Ripley's K for a random superposition is equal to one, independent of scale. We use the standardized Ripley's K with a radius of 10 m that is defined as:

$$\sigma_{a,b}(10) \equiv \frac{K_{a,b}(10)}{\pi r^2} \equiv \frac{A \sum_{i=1}^{N_a} N_{a_i,b,10}}{N_b N_a A_{10}}, \qquad \text{eqn 1}$$

where $N_{a_i,b,10}$ is the number of neighbours of species b within the interval 0–10 m from a focal individual i of species a; N_a and N_b are the total number of individuals of the respective species in the sample; A is the size of the study area and A_{10} is the size of a circle with 10 m radius r and $K_{a,b}(10)$ the bivariate Ripley's K for a radius of 10 m. Edge effects are avoided by using a buffer zone which was created by excluding individuals of species a from the sample that was closer than the neighbourhood radius to the edge of the study area (Haase 1995). We chose a buffer zone for edge correction as it gives unbiased results, is very efficient to compute and disregards few data for a radius of 10 m. Depending on the size and shape of the study area, the radius of the neighbourhood and the available computational resources, other choices of edge correction may be preferable (for a detailed discussion, see Illian *et al.* (2008)).

Co-association matrix and normalization

It is necessary to normalize $\sigma_{a,b}(10)$ because the tree and shrub species vary, both in their abundance and in their within-species spatial association (Condit *et al.* 2000; Flügge, Olhede & Murrell 2012). Consequently, it is difficult to compare the co-association measures $\sigma_{a,b}(10)$ for different pairs of species, in a meaningful way, because the marginal properties of both species vary. We therefore normalize the co-association values $\bar{\sigma}_{a,b}(10)$, accounting for marginal within-species aggregation and abundance, where we define:

$$\bar{\sigma}_{a,b}(10) \equiv \begin{cases} 0 & \text{for } a = b \\ \frac{\sigma_{a,b}(10)-1}{\mathrm{std}(\boldsymbol{\sigma}_{\mathbf{R(a,b)}}(10))} & \text{for } a \neq b \end{cases}. \qquad \text{eqn 2}$$

We produced 1000 randomized replicates of the spatial co-associations between every pair of species (a,b) by random torus translations (Lotwick & Silverman 1982; Harms *et al.* 2001) of the spatial locations of species b in relation to the spatial locations of species a. From these shifted patterns $R(a,b)$, we computed the co-associations $\sigma_{R(a,b)}(10)$ as described before and then computed the standard deviation of all 1000 values $\mathrm{std}(\boldsymbol{\sigma}_{\mathbf{R(a,b)}}(10))$. We remove unity from Equation 2 because under the assumption of random superposition (spatial independence) of two species, the expectation of $\sigma_{a,b}(10)$ is unity. By doing so, we shift the co-association values such that, compared to a null model of random superposition, negative values indicate co-segregation and positive values indicate co-aggregation.

As $\sigma_{a,b}(10)$ is identical to $\sigma_{b,a}(10)$, except for an asymmetry in the estimation introduced by the edge correction which does not affect the expected value, it is sufficient to compute the upper or lower triangle of the matrix to obtain the symmetric matrix of all pairwise co-association values. The diagonal entries of the matrix are set to zero, because we are not interested in the within-species spatial associations.

Clustering of species into subcommunities

We use the popular non-hierarchical k-means clustering algorithm (Gan, Ma & Wu 2007) to group the species into k disjunct sets of species with the most similar co-association values. Species are represented by the rows of the normalized co-association matrix. We use 100 replications of the k-means algorithm with random initialization to find the clustering that minimizes the sum of the difference $D(k)$ between the vectors of co-association values of species and the centroids of their clusters. We define $\bar{\boldsymbol{\sigma}}_{\mathbf{i}*}(10)$ as the vector of normalized co-association values between all species and species i, and $c_i{=}n$ denotes species i is assigned to the n-th cluster. The sum of the difference $D(k)$ is then computed as the sum of the Euclidean distances of each species co-association vector $\bar{\boldsymbol{\sigma}}_{\mathbf{i}*}(10)$ to the cluster centre of its cluster $c_i\left(\frac{\sum_{j\in\text{Species}} \delta_{c_i,c_j} \bar{\sigma}_{\mathbf{j}*}(10)}{\sum_{j\in\text{Species}} \delta_{c_i,c_j}}\right)$:

$$D(k) = \sum_{i\in\text{Species}} \left\| \bar{\sigma}_{\mathbf{i}*}(10) - \frac{\sum_{j\in\text{Species}} \delta_{c_i,c_j} \bar{\sigma}_{\mathbf{j}*}(10)}{\sum_{j\in\text{Species}} \delta_{c_i,c_j}} \right\|_2 \qquad \text{eqn 3}$$

with the Kronecker delta, δ_{c_i,c_j}, defined to be zero if $c_i{\neq}c_j$ and one if $c_i{=}c_j$. The result is that each group is a collection of species that show similar co-association patterns both within the group as well as with species in other groups.

To determine the upper limit of k for which the individual clusters contain meaningful information on the spatial patterns of the species, we use the normalized co-association matrix for 1000 forests (hereafter referred to as random forests) in which the within-species pattern is held constant, but where all species are shifted relative to each other via random torus translations. For both the BCI data and each random

forest, we then compute the sum of the within-cluster species to centroid distances $D(k)$ for all k between 1 and the number of species. The sum of the difference $D(k)$ is a measure of how well the clustering fits the data (i.e. how homogeneous the species are within a cluster). With increasing k, $D(k)$ trivially gets smaller, because more clusters can always partition a set such that the sum of the within-cluster distances is smaller than with fewer clusters. However, the amount by which $D(k)$ decreases from $D(k_1)$ to $D(k_2)$ (with $k_2 > k_1$) holds information on the inherent number of clusters in the data. We therefore compare $D(k)-D(k+1)$ for all k between 1 and 140 (number of species minus 1) for the BCI data with the 1000 random forests. If the species at BCI are more likely to be found in the same or different spatial regions, we would expect $D(k)-D(k+1)$ to be larger than in a random forest for at least the first few clusters k. This would show that the structure of the forest is not random, but that there are indeed subcommunities that reduce the sum of within-cluster distances more than what would be expected in a random null model. To avoid interpreting potentially spurious effects, at most those number of clusters k that exhibit statistical significance are investigated. In the analyses below, we use a 1% significance level, that is k is significant if $D(k)-D(k+1)$ for the species at BCI is larger than for 99% of the random forests.

Finally, we do not adopt hierarchical (either aglomorative or divisive) clustering methods, but recluster all species for each possible choice of number of clusters k. This is because we want to achieve the best clustering for any choice of cluster numbers, without constraint, as this will allow us to choose an appropriate value of k. If we wished to achieve a hierarchical understanding of groupings, alternative methods could be applied (Hastie, Tibshirani & Friedman 2009). Also, there are scenarios, where spatial contiguity is enforced by the choice of clustering procedure (Gordon 1996). We chose to not apply such methods, as we wish the data to naturally reproduce spatial contiguity from unconstrained algorithms.

Density maps

Once all species are grouped, the next step is to explore the spatial distribution of the subcommunities in the landscape. For that purpose, we first use the variable bandwidth kernel density estimator by Botev, Grotowski & Kroese (2010) to estimate the relative density of each species across the 50 ha plot. We then compute the mean relative density across the 50 ha for each subcommunity. In line with our clustering method, this method of computing the relative density for subcommunities weighs each species identically, independent of its abundance, as we are less concerned with the absolute density of individuals in a certain region (in which case we should weigh the species density maps by abundance or basal area), but instead, we want to find regions at which most species in the subcommunity co-occur (but see Appendix S5 for abundance-weighted results). However, it should be noted that the variable bandwidth kernel density estimator adapts the bandwidth to the detail of the available data and therefore smooths the point pattern with a larger bandwidth for rarer species. Thus, in practice, rare species generally have less influence on the mean kernel density of a subcommunity than common species.

Subcommunity maps

The information in the subcommunity density plots can be condensed into a single panel showing the dominant cluster, that is the subcommunity with the highest mean relative density, for each 20-by-20 m quadrat in the forest plot. Below, we draw such a figure by representing each subcommunity with a different colour, and drawing a map of the 50 ha forest plot where each 20-by-20 m quadrat is coloured according to the subcommunity that has the highest mean relative density (Figs 3 and 4).

Software

All analyses are conducted using Mathwork's MATLAB R2012b, and the source code is available as supplementary material. However, the reader should note that the method could be easily implemented in R (http://www.r-project.org/) since k-means clustering functions (e.g. in the *stats* library) and kernel density estimators (e.g. in the *kernsmooth* library) are standard tools.

Results

CO-ASSOCIATION MATRIX AND NORMALIZATION

Figure 1 shows the normalized co-association matrix of the adult population in the 2010 census. Each row (and column) represents the co-association values of one species with all others. The colour-coded bars along the side of the matrix show the clustering of the species for k=5 clusters (the colours are the same as those in Fig. 3).

CLUSTERING OF SPECIES INTO SUBCOMMUNITIES

Comparing $D(k)-D(k+1)$ between the data of the adult individuals of the 141 study species at the 2010 census, and 1000

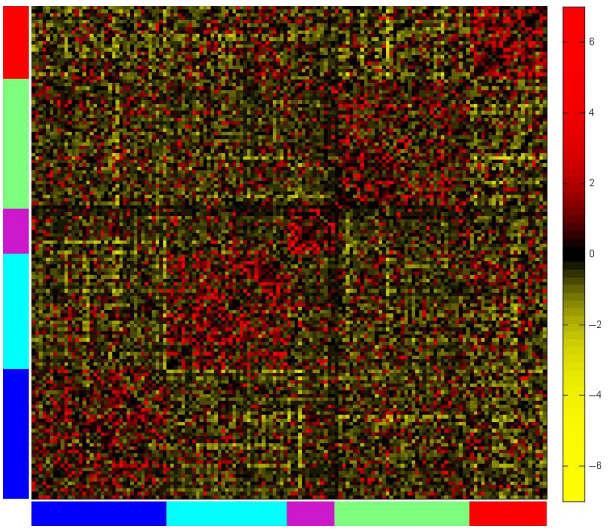

Fig. 1. The normalized co-association matrix between the adults of the 141 tree and shrub species from the BCI plot investigated here. Each row and column represents the co-associations of one species with all the others. Red indicates that two species are aggregated, and yellow that they are segregated, in comparison to a random null model. The matrix is symmetric about the diagonal, and the colours on the side show which species are grouped together in one subcommunity by the k-means algorithm (with k = 5) using the same colour coding as in Fig. 3d. Species are sorted by first sorting them according to their groups with k = 10, and then stepwise reducing the number of clusters and minimally resorting the species to group them according to the lower number of groups – the stepwise resorting is repeated until k = 5 is reached.

random forests based on the same individuals, shows that $D(k)-D(k+1)$ is larger for the true data than for 99% of the random forests up to values of $k = 10$ (see Appendix S2). This indicates that at least ten disjoint sets of species can be defined on the basis of their spatial distribution within the BCI forest plot, and within these sets, species are more correlated than expected by chance, that is assuming spatial independence between pairs of species. For $k > 10$, however, the additional fine scale structure in the data can no longer be distinguished from random effects. In order to enable better comparison of our results to those obtained by Harms *et al.* (2001), we choose to use $k = 5$ for most of our analyses (see Tables 1 and 2 for summary information on the clustering with $k = 5$), but further information on $k = 6$ to $k = 10$ clusters is given in Appendix S3.

DENSITY MAPS

The panels of Fig. 2 show the spatial variation mean density of the adult individuals of each of the five subcommunities as estimated by Botev's kernel density estimator (Botev, Grotowski & Kroese 2010).

SUBCOMMUNITY MAPS

Figure 3 shows the dominant subcommunities in the 50 ha forest plot for k between 2 and 5 clusters for the adult population of the 2010 census (see Appendix S3 for subcommunity maps for number of clusters up to $k = 10$). Our results largely concur with those of Harms *et al.* (2001). The first partitioning at $k = 2$ (Fig. 3a) seems to distinguish between the more wet habitat at the slopes and the drier plateau habitats. However, at $k = 2$, part of the north-western low plateau is grouped together with the slope (coloured red), rather than with the remainder of the plateau habitat (coloured green). For $k = 3$ (Fig. 3b), we do not find a distinction between the high plateau and the low plateau. Instead, the partitioning follows similar borders as the first, except that the north-western low plateau (cyan) stands out as a separate subcommunity. Thus, the remaining parts of the low plateau are still grouped together with the high plateau and the young forest. Only when increasing the number of clusters to $k = 4$ (Fig. 3c), do we find a separate high plateau/young forest subcommunity while still finding the split between the north-western and the south-eastern part of the low plateau. For $k = 5$ clusters (Fig. 3d, based on the subcommunity densities shown in Fig. 2), we find a subcommunity dominated by swamp species, together with some more widely spread shade-intolerant pioneer species. The divide between the north-western low plateau and the south-eastern low plateau seems to be mainly driven by life-history strategy, since the species of the south-eastern subcommunity have the highest mean shade-tolerance index of all clusters (Table 1), and the species of the north-western low plateau subcommunity are the second most shade intolerant on average (only the swamp/pioneer subcommunity has a lower mean shade-tolerance index).

The results for the juveniles (Fig. 4) are slightly less clear, although life-history strategy seems to be an important factor differentiating the subcommunities, suggesting light gaps drive some of the spatial structure evident in the plot. Most notably, the first grouping for k=2 (Fig. 4a) seems to be made along the line of shade tolerance (mean shade-tolerance index for the 'purple' subcommunity is $-1{\cdot}01 \pm 1{\cdot}55$; for the 'blue' subcommunity, it is $0{\cdot}50 \pm 0{\cdot}79$). For $k > 2$ (Fig. 4b–d), there always seems to be a subcommunity of highly shade-intolerant species beside those subcommunities that are more influenced by habitat and more similar to the subcommunities found for the adults (see Table 2 for summary information on the juveniles with $k = 5$). The results from the juveniles support the result from the adults that slope is the most important environmental variable to distinguish habitats with different species compositions at BCI.

Table 1. Summary information on the number of species, number of adult individuals and mean shade-tolerance index (* and ** mark mean shade-tolerance values that significantly depart from expected shade tolerance under the null hypothesis that shade tolerance is independent of grouping, on a 5% and 1% significance level, respectively; we used bootstrapping to compute the expected mean shade tolerance under the null hypothesis by randomly drawing shade-tolerance values from the set of all species) for the clustering with $k = 5$ subcommunities on the basis of the adult individuals in the 2010 census at Barro Colorado Island (BCI)

Subcommunity name	Number of species	Number of adults	Mean shade tolerance (±std)
South-eastern low plateau	37	7821	0·55** (±0·93)
North-western low plateau	33	2488	−0·63* (±1·48)
Swamp/shade-intolerant pioneers	13	1997	−1·57** (±1·96)
High plateau/young forest	37	17 164	0·17 (±0·95)
Slope	21	5686	0·45 (±0·70)

Table 2. Summary information on the number of species, number of juvenile individuals and mean shade-tolerance index (* and ** mark mean shade-tolerance values that significantly depart from expected shade tolerance under the null hypothesis that shade tolerance is independent of grouping, on a 5% and 1% significance level, respectively; we used bootstrapping to compute the expected mean shade tolerance under the null hypothesis by randomly drawing shade-tolerance values from the set of all species) for the clustering with $k = 5$ subcommunities on the basis of the juvenile individuals in the 2010 census at Barro Colorado Island (BCI)

Subcommunity name	Number of species	Number of juveniles	Mean shade tolerance (±std)
Blue/low plateau	32	41 573	0·95** (±0·51)
Light blue/mixed	37	10 498	−0·01 (±0·98)
Purple/shade-intolerant pioneers	13	3720	−2·51** (±1·34)
Green/mixed-swamp	26	17 448	−0·56 (±1·33)
Red/slope	33	41 573	0·49* (±0·50)

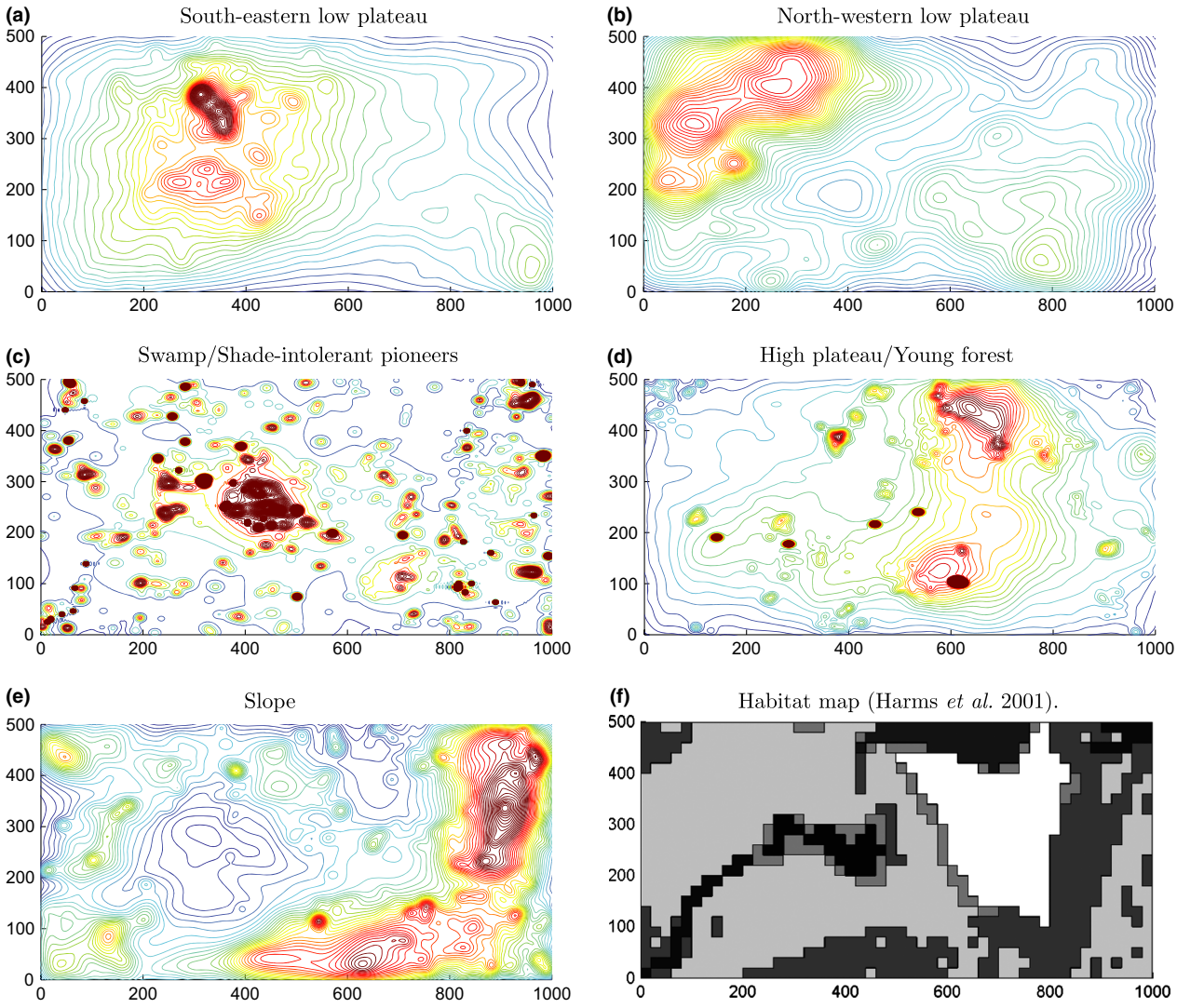

Fig. 2. Each panel shows the mean relative density of adults in the 2010 census obtained for one of five subcommunities. Red indicates that there are comparatively many individuals from that set of species while blue indicates lower relative densities. Densities were computed using Botev, Grotowski & Kroese (2010) kernel density estimator for each individual species and then averaged over all species in each subcommunity. The grey-scale map in subpanel (f) shows the different habitats at Barro Colorado Island (BCI) as defined by Harms *et al.* (2001).

Discussion

There are an increasing number of data sets available that provide rich spatial data for communities with large numbers of species and individuals (ForestGEO-CTFS 2013). While early studies have mainly focused on the spatial distributions of individual species, only aggregating the results to summarize the number of species that show certain spatial associations (e.g. Harms *et al.* 2001) or reporting a median value (e.g. Condit *et al.* 2000), an increasing number of studies use methods to draw information from the joint spatial distribution of species (e.g. Martínez *et al.* 2010; Wang *et al.* 2010; Lan *et al.* 2012; Luo *et al.* 2012; Wiegand *et al.* 2012; Baldeck *et al.* 2013; Punchi-Manage *et al.* 2013). The method we introduce makes use of an individual-based spatial co-association measure to group species together based upon their co-occurrence in the landscape, and we believe it provides a valuable addition to the growing toolkit of multivariate methods in spatial pattern analysis. In what follows we will discuss the strengths and limitations of our method and explore some of the possible extensions and adaptations that may be required for different data sets and biological questions.

The first step of our method is to calculate a normalized co-association matrix. The matrix of co-associations (Fig. 1) is itself an interesting object that could be used for other analyses such as the comparison of the degree of segregation and aggregation between different groups of species or communities. The normalization procedure we outline in Equation 2 is necessary to make the individual entries of the matrix comparable, but the precise measure that is used to compute co-associations will differ depending on the scale of the processes of interest, in the particular community and the available data. So, for example, while the BCI data set holds information on individual tree locations using precise x- and y-coordinates, our method could

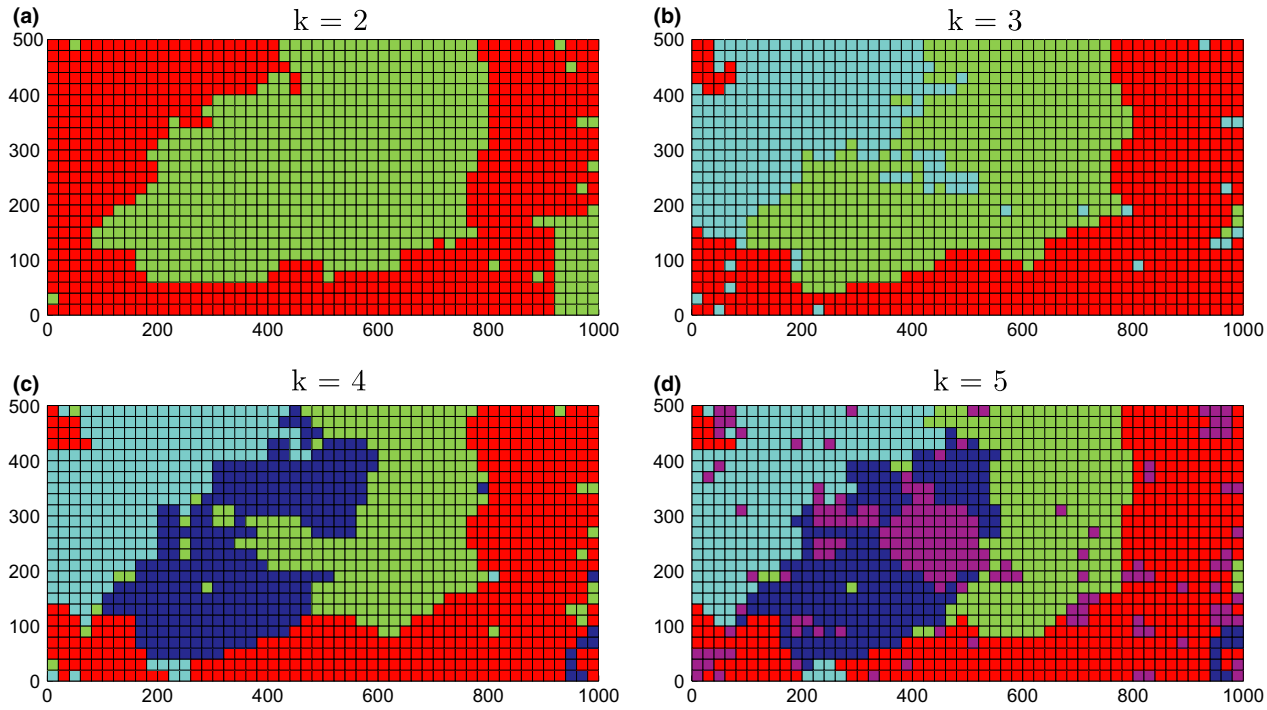

Fig. 3. Depicted are the subcommunities with the highest mean relative density for each 20-by-20 m quadrat for the number of cluster k between $k = 2$ and $k = 5$ (top-left to bottom-right) for the adult plants in the 2010 census.

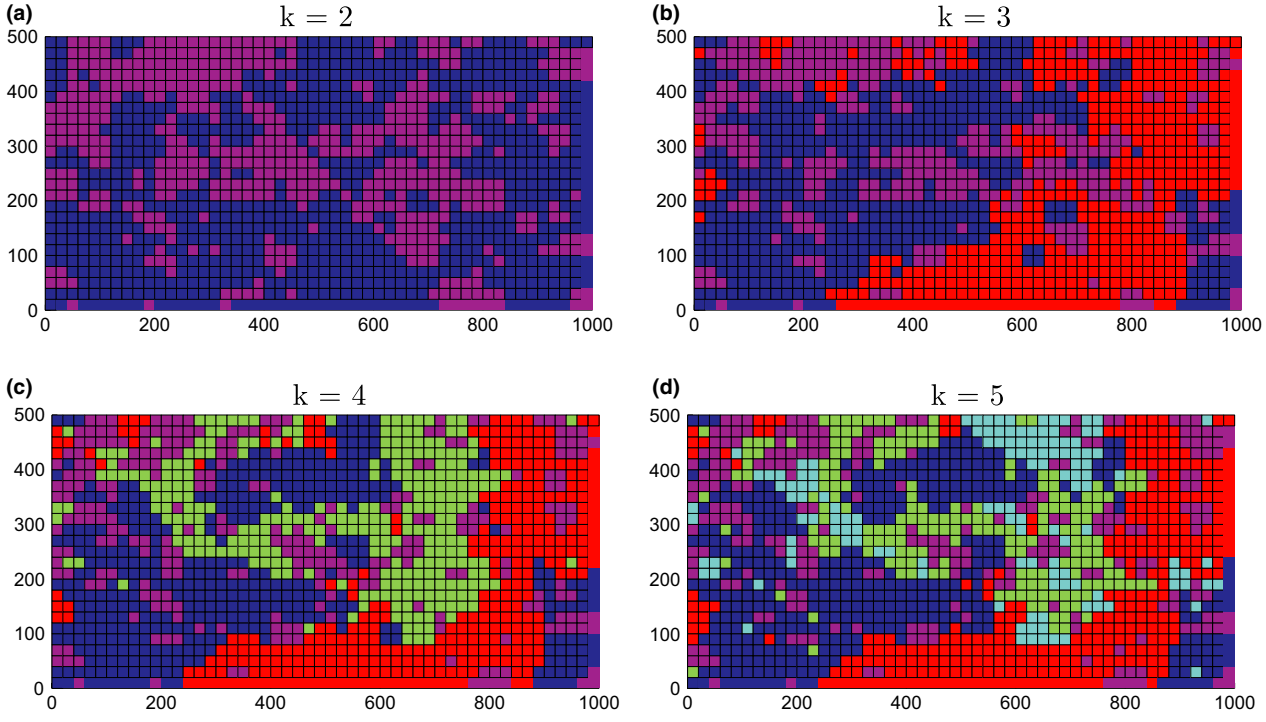

Fig. 4. Depicted are the subcommunities with the highest mean relative density for each 20-by-20 m quadrat for the number of clusters k between $k = 2$ and $k = 5$ (top-left to bottom-right) for the juvenile plants in the 2010 census.

easily be adapted if the data are for the presence/absence in a grid since the key ingredient required to group the species is a matrix of spatial co-associations. Similarly, the matrix can be computed for a variety of spatial scales. We use circles of 10 m diameter to compute co-association values for our analyses, as this is a scale at which many important ecological processes are happening in our study system (Uriarte *et al.* 2004). This scale also provides a good balance between covering a wide enough

area to achieve stable numerical results (i.e. for most species pairs, we find individuals of the other species in at least some of the neighbourhoods around the focus species) while still capturing fine-scaled differences in species distribution (Flügge, Olhede & Murrell 2012). Different species may show spatial correlations at different spatial scales, but because we were looking at the spatial structure of the plant community with all species, we used the same scale for all species pairs to keep the results comparable. However, the co-associations between species are relatively stable for small changes to neighbourhood diameter since the neighbourhood densities at different scales are highly correlated (Condit *et al.* 2000). Our experience shows that species that are assigned to a different cluster at a different scale are also likely to be the least typical species for that subcommunity. A detailed analysis of how the species co-associations change with scale is beyond the scope of this study (see Appendix S4 for cursory results on the 5 and 20 m scale), but we note that as an alternative to choosing a particular scale of analysis, a different approach could be to use a measure of spatial association such as the cross-pair overlap that averages the pair-correlation function over a range of distances (Brown *et al.* 2011).

The second step defines subcommunities from the co-association matrix and explores the number of statistically significant clusters in the data. Giving an upper limit for the number of subcommunities that can be distinguished from random is an important aspect of our method. By torus translating the full patterns for each species, we vary only the between-species patterns, while keeping the within-species patterns fixed. The advantages of this model are that it is very quick to compute randomized patterns and that it does not rely on fitting a model to describe the within-species pattern. The disadvantage, however, is that it only breaks the observed relationships of patterns between species, but does not produce all the variability present in stochastic realizations of within-species patterns. More sophisticated methods that fit models to the within-species variability of spatial patterns are available (Illian *et al.* 2008; Wiegand, He & Hubbell 2013).

The method to cluster species into subcommunities could also be adapted where appropriate. For example, if the expectation is that clusters breakdown into subclusters (e.g. groups relating to slope habitat break down into upper and lower slope groups), then hierarchical clustering methods (Gan, Ma & Wu 2007) could be used. We also note that our measure for co-association is not symmetrical for co-aggregation and segregation and our clusters are therefore potentially more strongly driven by positive associations between species.

Our results suggest there is statistically significant structure in the co-association matrix for up to ten subcommunities for the adults at BCI (see Appendix S2), but we argue that below this cut-off, there is no *a priori* correct number of clusters to analyse. By considering different numbers of subcommunities, we can explore which spatial structures are most prevalent and are therefore the strongest. In our main analyses, we concentrated on $k = 5$ clusters because this allowed easy comparison with previous analyses using different methods. However, further in-depth analysis of the characteristics of the species in the full $k = 10$ clusters could lead to new insights into ecologically important factors structuring the plant community at BCI. More information about the species and/or the abiotic environment is likely to be required to explain large number of clusters. A possible extension of our study could be to use the data on soil chemicals available for the BCI plot (Dalling *et al.* 2009; Condit *et al.* 2013). Baldeck *et al.* (2013) and John *et al.* (2007) indicate that soil properties can explain a significant part of the spatial distribution of species at BCI, and this might explain the additional structure when k ranges between 6 and 10. We also note that instead of looking at density maps and subcommunity maps, one could also stop at the clustering stage and analyse the attributes of species in the various subcommunities to explore what they have in common and what distinguishes them. In this case, the focus would be on species traits such as wood density, seed size, maximum adult size, etc. (Wright *et al.* 2010), or investigating the within and between subcommunities pattern of phylogenetic relatedness.

Although the interpretation of our species clusters focuses mainly on the role of environmental niches, other biotic and abiotic processes may be influential and the subcommunities represent realized rather than fundamental niches. In our example, it is clear that subcommunities are influenced by both environmental gradients such as slope and elevation, but also biotic conditions such as canopy gaps caused by treefall. The biotic factors may include both positive as well as negative forces acting on species spatial pattern since species are clustered according to similar positive and negative interspecific associations. A canopy gap, for example, provides particularly advantageous conditions for shade-intolerant species (Wright *et al.* 2003). On the other hand, shared pathogens or superior competitors could conceivably restrict the range of some species to those parts of the forest where the pathogen or competitor is not present. Spatial clustering of groups of species could also arise from positive interactions between the species, and the challenge is to be able to differentiate between, or quantify the roles of these candidate processes in generating the spatial associations, especially in species-rich communities where the spatial ecological signals of biotic interactions may be quite weak (Wiegand *et al.* 2012).

The third step takes the species clusters and computes relative density maps and a number of adaptations could be required depending on the data used. Firstly, we use a density estimation kernel that smooths the individual stem map to produce a continuous density distribution over the whole area. If the data were based on presence/absence in a grid, then a different kernel density estimator would be appropriate. Secondly, because we are particularly interested in rare species, we weight each species equally in the calculation of the mean subcommunity density, but other methods of weighting the contribution of species depending on abundance, biomass or other measure of relevance or reliability of the data are possible. Weighting by stem abundance might bias the results towards species that produce lots of juveniles, whereas weighting by biomass would bias towards species that produce large individuals and both might lead to different, but biologically informative interpretations of the spatial density maps.

The choice of weighting in abundance to produce the density maps allows us to detect patterns not dictated by a single, or a few, very abundant species. Kanagaraj *et al.* (2011), for example, found the result that juveniles showed more habitat partitioning than adults at BCI, despite adults being exposed to habitat filtering processes over a longer time. This might be explained by the fact that juvenile results are dominated by shade-tolerant species, which generally have size distributions that are more juvenile dominated than in shade-intolerant species and at the same time might be more dependent on numerous soil variables compared to shade-intolerant species that mainly rely on gaps in the canopy for recruitment.

In the final step, we summarize the information from the density maps of the subcommunities into a single map for the whole community, but we could have instead analysed the correlation between the density of the various subcommunities to the value of other continuous variables, such as soil nutrients. One key advantage of the multivariate approach is increased statistical power compared to looking for such correlations on a species-by-species basis, and this also allows the inclusion of relatively rare species into the analysis. By condensing the subcommunity maps into one, we get a result that is comparable to methods by Kanagaraj *et al.* (2011), Punchi-Manage *et al.* (2013) and Baldeck *et al.* (2013) that focus on the similarity of species compositions at spatial locations within the study area. In contrast to those studies, we start with groups of species and then consider how these subcommunities are spread over the landscape, and this will lead to different, but potentially complementary, results and interpretations. For example, if there are subcommunities that partially overlap in space, for example two disjoint sets of species A and B that partially overlap, then Kanagaraj *et al.* (2011), Punchi-Manage *et al.* (2013) and Baldeck *et al.* (2013) might detect three different spatial areas, two areas where only species of group A or B occur and a third area where individuals of species from both sets A and B are present. In contrast, a neighbourhood-based approach such as that introduced here would detect the two subcommunities, but would then assign the area where they overlap to one or the other depending on which group is more dominant. The choice of approach would depend upon whether locations or groups of species are the objects of interest. This can be interpreted as a choice of trade-off between bias and variance.

Perhaps a more similar method is presented by Legendre (2005) who adapts Kendall's coefficient of concordance to look for groups of species that are positively associated with one another across a number of discrete sites in the landscape. Here, for each study species, sample sites are ranked according to abundance observed, with the first ranked site having the highest abundance and the last ranked the lowest abundance. Kendall's coefficient of concordance is then computed to test whether the rankings are all independent of one another, and if the null hypothesis is rejected, *post hoc* tests are required to see which species positively associate across the sites. This method has the advantage of being relatively simple and works well for data where discrete locations are sampled (such as soil cores) rather than the locations of all individuals such as in the BCI data set, whereas our method works well for point process data covering one sample area. Legendre (2005) also points out Kendall's coefficient of concordance ranges from 0 (all species are independent) to 1 (all species are perfectly associated according to the site ranking) and does not work so well for assemblages where there are strong negative associations, something which cannot be ruled out *a priori* for most plant communities. In contrast, our method uses both positive and negative associations to cluster species into groups, and the test for statistical significance is on the subcommunities rather than the large number of pairwise associations.

In conclusion, we have outlined a novel method to compare spatial co-associations between different pairs of species of different abundances and within-species aggregation. Using the resultant matrix of normalized co-association values, we propose a method to group species into subcommunities of spatially co-associated species and provide a measure of the number of statistically significant subcommunities. The interpretation of these subcommunities depends on the system under study and the information available for this purpose. However, even when interpretation is difficult due to a lack of relevant covariates, the methods will suggest groups of species and areas of the landscape that merit further investigation. Moreover, by defining subcommunities, we are able to incorporate relatively rare species that might not be sufficiently abundant to be included in traditional species-focussed habitat association studies. As such, we believe our method is a useful addition to existing methods for multivariate spatial pattern analysis and can increase the understanding of communities that exhibit high biodiversity and for which the processes that structure the communities are not obvious to the human observer or not as yet well understood.

Acknowledgements

We thank the anonymous reviewers for helpful suggestions and especially suggestions of additional material to reference. This work was funded by a scholarship from the Engineering and Physical Sciences Research Council (EPSRC) to AF and the grant EP/I005250/1 and EP/L001519/1 from the EPSRC to SO. Part of the work was also funded by the Natural Environment Research Council (UK) blue skies fellowship NE/D009367/1to DM. The BCI forest dynamics research project was made possible by National Science Foundation grants to Stephen P. Hubbell – DEB-0640386, DEB-0425651, DEB-0346488, DEB-0129874, DEB-00753102, DEB-9909347, DEB-9615226, DEB-9615226, DEB-9405933, DEB-9221033, DEB-9100058, DEB-8906869, DEB-8605042, DEB-8206992, DEB-7922197 – support from the Center for Tropical Forest Science, the Smithsonian Tropical Research Institute, the John D. and Catherine T. MacArthur Foundation, the Mellon Foundation, the Celera Foundation, and numerous private individuals, and through the hard work of over 100 people from 10 countries over the past three decades. The plot project is part of the Center for Tropical Forest Science, a global network of large-scale demographic tree plots.

Data accessibility

The data used in this publication are available from the Center for Tropical Forest Science upon request (2013).

References

Bagchi, R., Henrys, P.A., Brown, P.E., Burslem, D.F.P., Diggle, P.J., Gunatilleke, C.S. *et al.* (2011) Spatial patterns reveal negative density dependence and habitat associations in tropical trees. *Ecology*, **92**, 1723–1729.

Baldeck, C.A., Harms, K.E., Yavitt, J.B., John, R., Turner, B.L., Valencia, R. *et al.* (2013) Soil resources and topography shape local tree community structure in tropical forests. *Proceedings of the Royal Society B: Biological Sciences*, **280**, 2012–2532.

Bellard, C., Bertelsmeier, C., Leadley, P., Thuiller, W. & Courchamp, F. (2012) Impacts of climate change on the future of biodiversity. *Ecology Letters*, **15**, 365–377.

Borcard, D. & Legendre, P. (2002) All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices. *Ecological Modelling*, **153**, 51–68.

Botev, Z.I., Grotowski, J.F. & Kroese, D.P. (2010) Kernel density estimation via diffusion. *Annals of Statistics*, **38**, 2916–2957.

Brooks, T., Mittermeier, R., Mittermeier, C., Da Fonseca, G., Rylands, A., Konstant, W. *et al.* (2002) Habitat loss and extinction in the hotspots of biodiversity. *Conservation Biology*, **16**, 909–923.

Brown, C., Law, R., Illian, J.B. & Burslem, D.F. (2011) Linking ecological processes with spatial and non-spatial patterns in plant communities. *African Journal of Ecology*, **99**, 1402–1414.

Callaway, R.M. (1995) Positive interactions among plants. *The Botanical Review*, **61**, 306–349.

Cardinale, B., Duffy, J., Gonzalez, A., Hooper, D., Perrings, C., Venail, P. *et al.* (2012) Biodiversity loss and its impact on humanity. *Nature*, **486**, 59–67.

Cheng, J., Mi, X., Nadrowski, K., Ren, H., Zhang, J. & Ma, K. (2012) Separating the effect of mechanisms shaping species-abundance distributions at multiple scales in a subtropical forest. *Oikos*, **121**, 236–244.

Comita, L.S., Muller-Landau, H.C., Aguilar, S. & Hubbell, S.P. (2010) Asymmetric density dependence shapes species abundances in a tropical tree community. *Science*, **329**, 330–332.

Condit, R. (1998) *Tropical Forest Census Plots*. Springer-Verlag and R. G. Landes Company, Berlin, Germany, and Georgetown, Texas, USA.

Condit, R., Ashton, P.S., Baker, P., Bunyavejchewin, S., Gunatilleke, S., Gunatilleke, N. *et al.* (2000) Spatial patterns in the distribution of tropical tree species. *Science*, **288**, 1414–1418.

Condit, R., Engelbrecht, B.M., Pino, D., Pérez, R. & Turner, B.L. (2013) Species distributions in response to individual soil nutrients and seasonal drought across a community of tropical trees. *Proceedings of the National Academy of Sciences of the United States of America*, **110**, 5064–5068.

Dalling, J., John, R., Harms, K., Stallard, R. & Yavitt, J. (2009) Soil maps of Barro Colorado Island 50 ha plot. http://ctfs.arnarb.harvard.edu/webatlas/datasets/bci/soilmaps/BCIsoil.html [Online; accessed 06-February-2013].

Flügge, A.J., Olhede, S.C. & Murrell, D.J. (2012) The memory of spatial patterns: changes in local abundance and aggregation in a tropical forest. *Ecology*, **93**, 1540–1549.

ForestGEO-CTFS (2013) Center for tropical forest science. [online] http://www.ctfs.si.edu. Accessed: 2013-06-15.

Gan, G., Ma, C. & Wu, J. (2007) Data clustering: theory, algorithms, and applications, volume 20. SIAM.

Gordon, A.D. (1996) A survey of constrained classification. *Computational Statistics & Data Analysis*, **21**, 17–29.

Haase, P. (1995) Spatial pattern analysis in ecology based on Ripley's K-function: introduction and methods of edge correction. *Journal of Vegetation Science*, **6**, 575–582.

Hardin, G. (1960) The competitive exclusion principle. *Science*, **131**, 1292–1297.

Harms, K.E., Condit, R., Hubbell, S.P. & Foster, R.B. (2001) Habitat associations of trees and shrubs in a 50-ha neotropical forest plot. *African Journal of Ecology*, **89**, 947–959.

Hastie, T., Tibshirani, R. & Friedman, J. (2009) *The Elements of Statistical Learning*. Springer, New York, New York, USA.

Hubbell, S.P., Foster, R.B., O'Brien, S.T., Harms, K.E., Condit, R., Wechsler, B., Wright, S.J. & de Lao, S.L. (1999) Light-gap disturbances, recruitment limitation, and tree diversity in a neotropical forest. *Science*, **283**, 554–557.

Hubbell, S.P., Condit, R. & Foster, R.B. (2005) Barro colorado forest census plot data. [online] https://ctfs.arnarb.harvard.edu/webatlas/datasets/bci.

Illian, J., Penttinen, A., Stoyan, H. & Stoyan, D. (2008) *Statistical Analysis and Modelling of Spatial Point Patterns, Volume 70*. John Wiley & Sons, Chichester.

Isbell, F., Polley, H. & Wilsey, B. (2009) Biodiversity, productivity and the temporal stability of productivity: patterns and processes. *Ecology Letters*, **12**, 443–451.

Itoh, A., Ohkubo, T., Nanami, S., Tan, S. & Yamakura, T. (2010) Comparison of statistical tests for habitat associations in tropical forests: a case study of sympatric dipterocarp trees in a Bornean forest. *Forest Ecology and Management*, **259**, 323–332.

John, R., Dalling, J.W., Harms, K.E., Yavitt, J.B., Stallard, R.F., Mirabello, M. *et al.* (2007) Soil nutrients influence spatial distributions of tropical tree species. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 864–869.

Kanagaraj, R., Wiegand, T., Comita, L.S. & Huth, A. (2011) Tropical tree species assemblages in topographical habitats change in time and with life stage. *African Journal of Ecology*, **99**, 1441–1452.

Lan, G., Getzin, S., Wiegand, T., Hu, Y., Xie, G., Zhu, H. & Cao, M. (2012) Spatial distribution and interspecific associations of tree species in a tropical seasonal rain forest of china. *PLoS ONE*, **7**, e46074.

Ledo, A., Burslem, D.F., Condés, S. & Montes, F. (2013) Micro-scale habitat associations of woody plants in a neotropical cloud forest. *Journal of Vegetation Science*, **24**, 1086–1097.

Legendre, P. (2005) Species associations: the kendall coefficient of concordance revisited. *Journal of Agricultural, Biological, and Environmental Statistics*, **10**, 226–245.

Legendre, P., Mi, X., Ren, H., Ma, K., Yu, M., Sun, I.F. & He, F. (2009) Partitioning beta diversity in a subtropical broad-leaved forest of china. *Ecology*, **90**, 663–674.

Lotwick, H. & Silverman, B. (1982) Methods for analysing spatial processes of several types of points. *Journal of the Royal Statistical Society Series B (Methodological)*, **44**, 406–413.

Luo, Z.R., Yu, M.J., Chen, D.L., Wu, Y.G. & Ding, B.Y. (2012) Spatial associations of tree species in a subtropical evergreen broad-leaved forest. *Chinese Journal of Plant Ecology*, **5**, 346–355.

Martínez, I., Wiegand, T., Gonzalez-Taboada, F. & Obeso, J.R. (2010) Spatial associations among tree species in a temperate forest community in north-western spain. *Forest Ecology and Management*, **260**, 456–465.

Muller-Landau, H.C. & Hardesty, B.D. (2005) Seed dispersal of woody plants in tropical forests: concepts, examples and future directions. *Biotic Interactions in the Tropics* (eds D. Burslem, M. Pinard & S. Hartley), pp. 267–309. Cambridge University Press, Cambridge.

Murrell, D.J. (2010) When does local spatial structure hinder competitive coexistence and reverse competitive hierarchies? *Ecology*, **91**, 1605–1616.

Murrell, D.J., Purves, D.W. & Law, R. (2001) Uniting pattern and process in plant ecology. *Trends in Ecology & Evolution*, **16**, 529–530.

Punchi-Manage, R., Getzin, S., Wiegand, T., Kanagaraj, R., Savitri Gunatilleke, C., Nimal Gunatilleke, I., Wiegand, K. & Huth, A. (2013) Effects of topography on structuring local species assemblages in a Sri Lankan mixed dipterocarp forest. *African Journal of Ecology*, **101**, 149–160.

Rajala, T. & Illian, J. (2012) A family of spatial biodiversity measures based on graphs. *Environmental and Ecological Statistics*, **19**, 545–572.

Turkington, R. & Harper, J.L. (1979) The growth, distribution and neighbour relationships of Trifolium repens in a permanent pasture: I. ordination, pattern and contact. *The Journal of Ecology*, **67**, 201–218.

Uriarte, M., Condit, R., Canham, C.D. & Hubbell, S.P. (2004) A spatially explicit model of sapling growth in a tropical forest: does the identity of neighbours matter? *African Journal of Ecology*, **92**, 348–360.

Wang, X., Wiegand, T., Hao, Z., Li, B., Ye, J. & Lin, F. (2010) Species associations in an old-growth temperate forest in north-eastern china. *African Journal of Ecology*, **98**, 674–686.

Wang, X., Wiegand, T., Wolf, A., Howe, R., Davies, S.J. & Hao, Z. (2011) Spatial patterns of tree species richness in two temperate forests. *African Journal of Ecology*, **99**, 1382–1393.

Wang, X., Swenson, N.G., Wiegand, T., Wolf, A., Howe, R., Lin, F. *et al.* (2013) Phylogenetic and functional diversity area relationships in two temperate forests. *Ecography*, **36**, 883–893.

Wiegand, T. & Moloney, K. (2004) Rings, circles, and null models for point pattern analysis in ecology. *Oikos*, **104**, 209–229.

Wiegand, T., Gunatilleke, C.S., Gunatilleke, I.N. & Huth, A. (2007) How individual species structure diversity in tropical forests. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 19029–19033.

Wiegand, T., Huth, A., Getzin, S., Wang, X., Hao, Z., Gunatilleke, C.S. & Gunatilleke, I.N. (2012) Testing the independent species arrangement assertion made by theories of stochastic geometry of biodiversity. *Proceedings of the Royal Society B: Biological Sciences*, **279**, 3312–3320.

Wiegand, T., He, F. & Hubbell, S.P. (2013) A systematic comparison of summary characteristics for quantifying point patterns in ecology. *Ecography*, **36**, 092–103.

Williams, J.N. (2013) Humans and biodiversity: population and demographic trends in the hotspots. *Population and Environment*, **34**, 510–523.

Wright, J. (2002) Plant diversity in tropical forests: a review of mechanisms of species coexistence. *Oecologia*, **130**, 1–14.

Wright, S.J., Muller-Landau, H.C., Condit, R. & Hubbell, S.P. (2003) Gap-dependent recruitment, realized vital rates, and size distributions of tropical trees. *Ecology*, **84**, 3174–3185.

Wright, S.J., Kitajima, K., Kraft, N.J., Reich, P.B., Wright, I.J., Bunker, D.E. *et al.* (2010) Functional traits and the growth-mortality trade-off in tropical trees. *Ecology*, **91**, 3664–3674.

Received 4 April 2013; accepted 16 September 2014
Handling Editor: Helen Freckleton

Supporting Information

Additional Supporting Information may be found in the online version of this article.

Appendix S1. Species list.

Appendix S2. Structure at Barro Colorado Vs. Randomised forests.

Appendix S3. Subcommunity maps for k=6 to k=10.

Appendix S4. Subcommunity maps for 5 and 20 metres scale.

Appendix S5. Abundance-weighted cluster densities.

Data S1. Matlab source code.

Data S2. Matlab code readme file.