Oecologia
DOI 10.1007/s00442-007-0923-8

COMMUNITY ECOLOGY - ORIGINAL PAPER

# Explaining variation in tropical plant community composition: influence of environmental and spatial data quality

**Mirkka M. Jones · Hanna Tuomisto · Daniel Borcard · Pierre Legendre · David B. Clark · Paulo C. Olivas**

Received: 9 March 2007 / Accepted: 12 November 2007


**Abstract** The degree to which variation in plant community composition (beta-diversity) is predictable from environmental variation, relative to other spatial processes, is of considerable current interest. We addressed this question in Costa Rican rain forest pteridophytes (1,045 plots, 127 species). We also tested the effect of data quality on the results, which has largely been overlooked in earlier studies. To do so, we compared two alternative spatial models [polynomial vs. principal coordinates of neighbour matrices (PCNM)] and ten alternative environmental models (all available environmental variables vs. four subsets, and including their polynomials vs. not). Of the environmental data types, soil chemistry contributed most to explaining pteridophyte community variation, followed in decreasing order of contribution by topography, soil type and forest structure. Environmentally explained variation increased moderately when polynomials of the environmental variables were included. Spatially explained variation increased substantially when the multi-scale PCNM spatial model was used instead of the traditional, broad-scale polynomial spatial model. The best model combination (PCNM spatial model and full environmental model including polynomials) explained 32% of pteridophyte community variation, after correcting for the number of sampling sites and explanatory variables. Overall evidence for environmental control of beta-diversity was strong, and the main floristic gradients detected were correlated with environmental variation at all scales encompassed by the study (c. 100–2,000 m). Depending on model choice, however, total explained variation differed more than fourfold, and the apparent relative importance of space and environment could be reversed. Therefore, we advocate a broader recognition of the impacts that data quality has on analysis results. A general understanding of the relative contributions of spatial and environmental processes to species distributions and beta-diversity requires that methodological artefacts are separated from real ecological differences.

**Keywords** Environmental control · Model specification · Spatial structure · Species composition · Variation partitioning

Communicated by Katherine Gross.

M. M. Jones (✉) · H. Tuomisto
Department of Biology, University of Turku,
20014 Turku, Finland
e-mail: mirkka.jones@utu.fi

D. Borcard · P. Legendre
Département des sciences biologiques,
Université de Montréal, C.P. 6128,
succursale Centre-ville, H3C 3J7 Montreal, QC, Canada

D. B. Clark
Department of Biological Sciences,
University of Missouri-St. Louis, St. Louis,
MO 63121, USA

P. C. Olivas
Department of Biological Sciences,
Florida International University,
Miami, FL 33199, USA

## Introduction

Several studies have documented that plant species composition and abundances within tropical forest landscapes respond to heterogeneity in soil properties, topography and forest successional stage (e.g. Denslow 1987; Dirzo et al. 1992; Tuomisto et al. 1995; Clark et al. 1999; Tuomisto and Poulsen 2000; Harms et al. 2001; Duque et al. 2002;

Potts et al. 2002; Tuomisto et al. 2003a, b; Cannon and Leighton 2004; Valencia et al. 2004; John et al. 2007). However, it is debated to what degree floristic composition depends on environmental factors relative to other processes, such as dispersal limitation and biotic interactions (Hubbell 2001; Dalling et al. 2002; Fine et al. 2004; Wyatt and Silman 2004).

Many factors that influence plant distributions will generate spatial pattern in community composition. Dispersal, biotic interactions, and gap dynamics are likely to produce spatial structure most evident at relatively fine scales, whereas edaphic or topographic variation may produce structure at different scales depending on underlying geology and geomorphology.

There has been a lot of recent interest in modelling species abundances using both environmental and spatial explanatory variables to study their relative contributions to explaining beta-diversity, i.e. variation in community composition. This can be done using canonical analysis, such as redundancy analysis (RDA) or canonical correspondence analysis (CCA) in a variation partitioning framework (Borcard et al. 1992). In theory, floristic composition can exhibit two different kinds of spatial structure: (1) autogenous structure, independent of any environmental variation; and (2) exogenous structure, which results when species respond to environmental variables that themselves are spatially structured. In practice, interpretation is complicated by the fact that spatially structured but unmeasured environmental variables may also affect floristic composition.

Variation partitioning has been used in numerous studies on plant species composition, with the total proportion of variation explained ranging from 20 to 72% in some recent temperate forest studies (Borcard et al. 1992; Gilbert and Lechowicz 2004; Cottenie 2005; Karst et al. 2005; Svenning and Skov 2005; Thomsen et al. 2005; Corney et al. 2006), and from 16 to 86% in studies in tropical forests (Duivenvoorden 1995; Balvanera et al. 2002; Dalle et al. 2002; Arbeláez and Duivenvoorden 2004; Svenning et al. 2004; Duque et al. 2005; Chust et al. 2006).

Ecologically meaningful comparison of the results of different studies is difficult because the amount of variation in community composition explained by "space" and "environment" will depend on how these are modelled. The spatial model has usually been based on either the $x$ and $y$ coordinates of the sampling sites, or on the coordinates and their second- and third-order polynomial terms. Although polynomial terms enable modelling more complex spatial patterns than simple linear trend surfaces, these are nonetheless restricted to broad-scale patterns (Borcard and Legendre 2002). Through the use of principal coordinates of neighbour matrices (PCNMs, Borcard and Legendre 2002; Borcard et al. 2004; Dray et al. 2006), complex spatial patterns can be modelled at different spatial scales, so a PCNM model may capture a larger proportion of the variation in community composition than the simpler polynomial model. How big the differences are has rarely been tested on real data (but see Borcard and Legendre 2002).

Similarly, the degree to which environmental effects on species composition can be discovered depends on which environmental variables are measured and on how these are modelled in the analysis. Many tropical forest studies have used only environmental data that are easy to obtain, such as topographic or forest structural variables, or coarse information on soils or geology (e.g. Clark et al. 1995; 1999; Harms et al. 2001; Balvanera et al. 2002; Dalle et al. 2002; Cannon and Leighton 2004; Valencia et al. 2004; Chust et al. 2006). Others have also included data from laboratory analyses of soil samples (e.g. Duque et al. 2002; Potts et al. 2002; Phillips et al. 2003; Tuomisto et al. 2003a, b; Arbeláez and Duivenvoorden 2004; Vormisto et al. 2004; Duque et al. 2005; John et al. 2007). Such methodological differences may have important consequences for the results, but this has been under-appreciated when different studies have been compared (e.g. Balvanera et al. 2002; Cottenie 2005; Chust et al. 2006). Moreover, Austin (2002) strongly criticised canonical ordination studies for failing to consider the realistic possibility that species responses to environmental gradients are non-linear.

In the present paper we document patterns in pteridophyte community composition at La Selva Biological Station, Costa Rica, and quantify the roles of environmental and spatial variables in explaining observed floristic patterns. We model the environmental component using a full set and different subsets of soil, topographic and forest structural variables (with and without their quadratic and cubic functions), and the spatial component both using the traditional polynomial model and a more flexible PCNM model. Through these comparisons, we examine the consequences of spatial and environmental model choice in terms of: (1) the total proportion of floristic variation explained, (2) the relative contributions of "space" and "environment", (3) how the different environmental subsets contribute to overall environmentally explained variation, and (4) the patterns of spatio-environmental structuring that can be identified in pteridophyte community composition and in the distributions of individual species.

## Materials and methods

### Study site

The study was carried out in c. 5 $km^2$ of old growth rain forest belonging to La Selva Biological Station of the Organization for Tropical Studies (OTS), in the Caribbean lowlands of Costa Rica. The site has a mean monthly tem-

perature of c. 26°C and receives an average of over 100 mm of rain each month and over 4,000 mm annually (OTS, unpublished data).

The study area is covered by a grid of 1,048 permanent intersection markers with a 50 × 100-m spacing. It encompasses a range of soil types, including alluvial terraces formed by recent or historical flooding, swamps, residual soils formed by in situ weathering of ancient lava flows, and stream valleys with infertile colluvial soils (Clark et al. 1999). Elevation increases by c. 100 m across the site in a south-west direction. Alluvial and swamp soils are restricted to lower elevations, which are also relatively flat, whereas higher elevations have a steeper topography and are dominated by residual soils.

### Floristic data

We inventoried pteridophytes (ferns and fern allies) in 1,042 circular sample plots (each 100 $m^2$) between July 2001 and July 2002. The plots were centred on 1,042 of the grid intersections within the study area. Within each plot we identified all individuals with at least one leaf longer than 10 cm; epiphytic and climbing individuals with no such leaves within 2 m of the ground were excluded. All apparently separate plants were counted as individuals, although in certain species some were probably clonal.

We collected voucher specimens of each species and of all individuals we were unable to identify in the field to a previously collected species. The specimens were cross-checked to obtain consistent identifications to morphospecies, and these were matched with named species using Flora Mesoamericana (Moran and Riba 1995) and comparisons with existing herbarium material. Our specimens are deposited in herbaria in Costa Rica [Herbario Nacional de Costa Rica (CR), Universidad de Costa Rica (USJ) and the on-site herbarium of La Selva Biological Station (LSCR); abbreviations according to Holmgren and Holmgren 1998] and Finland (University of Turku; TUR). Unicates are in CR.

Due to lack of access or accidental omission, we did not obtain plot data at six grid intersections. Three of the six missing plots overlapped with a parallel transect-based survey (Jones et al. 2006), so we used overlapping transect subunits of the same surface area (5 × 20 m) to estimate pteridophyte data for them. We did this because gaps in the sampling design result in irregular PCNM spatial descriptors, complicating the interpretation of the resulting spatial model (Borcard and Legendre 2002).

### Environmental data

We classified each of the 1,045 plots into one of five qualitative soil classes (old alluvium, recent alluvium, residual, stream valley or swamp). Soil chemical data on pH, total concentrations of C, N, and P, and exchangeable concentrations of K, Ca, Mg and Mn were also available for all plots. The soil samples (taken to 10 cm depth) were collected between March 1998 and May 1999 (D. B. Clark, unpublished data).

For each plot we also defined five topographic variables using data on slope, aspect, elevation and topographic position (Clark et al. 1999). Slope was measured in the steepest direction across the plot. Aspect was divided into sine (aspect), to distinguish plots on either side of a north–south axis, and cosine (aspect) to distinguish those on either side of an east–west axis. Elevation was based on optical ground surveys for 1,026 plots, and for 19 plots it was taken from a digital elevation model based on Light Detection and Ranging (LIDAR) data (from the University of Maryland and NASA Vegetation Canopy LIDAR Mission; cf. Hofton et al. 2002). Topographic position was defined as one of five ordered classes: flat high ground, upper slope, midslope, base of slope/flat low ground, riparian.

We measured canopy openness at 1,042 plots using the canopy-scope method (Brown et al. 2000), which estimates the size of the largest visible canopy gap on a scale of 0–25. We estimated missing data for three plots on the basis of average light levels measured at similar sites elsewhere in the study area (closed canopy, small canopy gap, medium-sized tree fall gap). Additional measures of forest structure were the number of tree stems ≥10 cm diameter at breast height and their basal area in each of the 1,045 plots (collected between 1993 and 1995, Clark et al. 1999).

### Spatial data

We generated two sets of continuous spatial variables from the $x$ and $y$ coordinates of each plot in the program SpaceMaker2 (Borcard and Legendre 2004). The first set consisted of the nine terms of a cubic trend surface polynomial (the centred site coordinates, $x$ and $y$, and $x^2$, $y^2$, $xy$, $x^3$, $y^3$, $x^2y$ and $xy^2$). The second set was created using the PCNM method (Borcard and Legendre 2002; Dray et al. 2006). The polynomial variables represent linear or curved structures at the extent of the entire study area, whereas PCNMs consist of orthogonal waves, whose wavelengths range across all scales encompassed by the sampling scheme. In our case, PCNM wavelengths ranged from c. 100 to 2,000 m. If sampling is unidimensional and regular, the PCNM variables are sine waves and their number is about two-thirds of the number of sampling sites (Borcard and Legendre 2002). If sampling is two-dimensional or irregular, the shape of the PCNMs is less regular and their number varies. To make our sampling grid and the resulting PCNMs more regular, we added three supplementary pairs of coordinates to fill holes in the grid, for the purpose of PCNM generation alone, where actual sample data were

unavailable (Borcard and Legendre 2002). The subsequent removal of these resulted in a small loss of orthogonality among the PCNMs (nonetheless, the largest correlation among any pair of PCNMs was just 0.0093). A total of 665 PCNMs was generated.

### Data analysis

Prior to analysis, we Hellinger-transformed the pteridophyte data (Legendre and Gallagher 2001) to express species abundances as square-root transformed proportionate abundances in each sampling site. This transformation reduces the weight of the most abundant species in the analysis. We also transformed the soil chemical data (except pH) by taking their natural logarithm. This was done because plants are likely to respond more strongly to a given absolute change in nutrient availability when the nutrient is scarce than when it is abundant. We coded each of the five soil type classes as a binary variable.

For comparison with the dataset comprising the original 20 environmental variables (simple environmental model), we generated a polynomial environmental dataset consisting of the original 20 variables and their quadratic and cubic functions. Additive combinations of the original variables and their higher order functions allow nonlinear relationships with variation in pteridophyte species composition to be modelled. Polynomial terms were calculated for all variables except the binary soil types and the sinusoid variables sine (aspect) and cosine (aspect). The polynomial environmental dataset thus included a total of 48 variables.

We ran forward selection on each set of environmental (simple or polynomial) and spatial (polynomial or PCNM) explanatory variables separately, to select those variables with a significant ($P \leq 0.05$ after 999 random permutations) contribution to explaining variation in floristic composition (following the procedure recommended by Blanchet et al. 2007). This was done using the R-language (R Development Core Team 2006) function forward.sel in the Packfor package (available at http://www.bio.umontreal.ca/legendre/). Only the selected variables were used in subsequent analyses.

We ran variation partitioning (Borcard et al. 1992) to quantify the proportion of the variation in community composition explained by variation in each of the four combinations of environmental and spatial explanatory variable sets. We adjusted the $R^2$-values to account for the number of sampling sites and explanatory variables, as unadjusted $R^2$-values are biased (Peres-Neto et al. 2006), and report the adjusted values ($R^2_a$) throughout. We recorded the proportion of variation explained ($R^2_a$) in RDA analyses by either the significant spatial (polynomial or PCNM) or the significant environmental (simple or polynomial) variables, or both simultaneously. Using these $R^2_a$ -values, we calculated the purely environmental (PE), purely spatial (PS), and spatially structured environmental (SSE) fractions of the total explained variation in floristic composition (Borcard et al. 1992). We tested the significance of the PS and PE fractions by means of 999 permutations under the reduced model. The R-language functions varpart, rda and anova.cca in the vegan library were used (Oksanen et al. 2007).

The remainder of the variation partitioning analyses focused on our most comprehensive environmental and spatial models, the full polynomial environmental model and PCNM spatial model. We first divided the significant explanatory variables into subsets. The environmental variables were divided into four subsets: topography, forest structure, soil type, and quantitative soil chemical data. The PCNMs were divided into three subsets of 34 PCNM variables each: broad-, medium- and fine-scale, with wavelengths c. 100–300, 300–650 and 650–2,000 m, respectively. RDAs followed by variation partitioning were run with each of the PCNM subsets combined with, in turn, each subset and the full set of environmental variables.

We then ran RDA using separately the polynomial spatial model, full PCNM spatial model or full polynomial environmental model as the explanatory dataset in CANOCO (ter Braak and Smilauer 1998). We extracted the fitted site scores for each of the first three canonical axes and mapped them, to visualize the major patterns of spatial and environmental structuring detected in the pteridophyte dataset. We ran multiple regression analyses to check which environmental variables contributed most to the site scores on each environmental ordination axis, as judged from their partial regression coefficients, in the program Permute (version 3.4, Casgrain 2001). We then used the same procedure to check which environmental variables were most strongly related to the spatial patterns in floristic composition represented by site scores on each PCNM and polynomial ordination axis. All variables were standardized prior to running the regression analyses.

Finally, we extracted species scores in CANOCO to check which pteridophyte species were best modelled by the PCNM spatial ordination axes and polynomial environmental ordination axes. We tabulated the species with the highest ten species scores (positive or negative) on each axis and interpreted their distribution patterns in light of the explanatory data.

## Results

### General

During the inventory, we encountered 89,708 pteridophyte individuals, the vast majority of which were terrestrial. They represented 128 morphospecies, of which 103 could

be identified to a named species, 24 were identified to genus level, and one (a single plant) remained unidentified. Two of the named species were confused in the field and are combined in the analyses, which hence include 127 species. Each 100-$m^2$ plot contained from three to 780 pteridophyte individuals (mean 86, median 72) and from one to 28 species (mean 10, median 9). The most abundant species was represented by 25,439 individuals in the entire dataset, and the least abundant 11 species were represented by a single individual each (mean 706, median 41). The five most abundant species, in descending order of abundance, were *Danaea wendlandii*, *Salpichlaena* sp. 1, *Polybotrya villosula*, *Lomariopsis vestita* and *Adiantum obliquum* (see Table 1 for authorities).

## Variation partitioning

Forward selection included all 20 original environmental variables as significant predictors of variation in community composition in the simple environmental model, and 39 of 48 terms in the polynomial environmental model. The simple environmental model explained $R^2_a$ =21.0% of community variation and the polynomial environmental model explained 25.8%. In the latter case, total variation explainable using soil chemistry was 19.1%, followed by topography (14.3%), soil type (8.9%) and forest structure (3.5%). These subsets of environmental data were partly redundant (Fig. 1), but each also made a unique contribution to explained variation. The unique contribution of soil chemistry was the largest (7.1%), followed by topography (4.1%), forest structure (1.1%), and soil type (1.0%).

Forward selection of the spatial variables included all nine terms of the polynomial spatial model, and 102 of 665 PCNMs in the PCNM spatial model. The polynomial spatial model explained 4.3% of the variation in pteridophyte community composition and the PCNM spatial model explained 15.9%.

Depending on the combination of spatial and environmental models used, the variation partitioning results varied greatly (Fig. 2). The lowest total variation explained (TVE), 6.9%, resulted when the polynomial spatial model was combined with the simple model of forest structure. The highest TVE (32.3%) resulted when the PCNM spatial model was combined with the polynomial model of all environmental data. Similarly, the PE fraction varied from 1.8 to 23.7%, the SSE fraction from 0.3 to 9.4%, and the PS fraction from 2.2 to 14.8%. When the polynomial spatial model was used, PE > PS resulted in all cases except when the environmental model comprised forest structural data alone. When the PCNM spatial model was used, the situation was usually reversed, but PE > PS resulted both when all environmental data types were used, and when soil chemical variables were used together with their polynomials (Fig. 2).

We further decomposed the PS and the SSE fractions of variation explained by our most comprehensive model into broad, medium and fine-scale fractions. Most of the spatial structure in community composition was found at broad

**Table 1** Pteridophyte species with the clearest relationships (ten highest species scores) with each of the first three constrained RDA axes. The explanatory variables consisted of either a spatial model based on principal coordinates of neighbour matrices (PCNM) or an environmental model based on all available environmental variables and their polynomials. Positive (+) and negative (−) species scores are shown separately; within each category, the species are listed in order of decreasing absolute value of their scores

| PCNM axis 1 | PCNM axis 2 | PCNM axis 3 | Env. axis 1 | Env. axis 2 | Env. axis 3 |
|---|---|---|---|---|---|
| + Trich coll | + Thely nica | + Lomar vest | + Thely nica | + Polyb vill | + Lomar vest |
| + Polyb osmu | + Trich coll | + Polyb vill | + Trich coll | + Thely nica | + Polyt feei |
| + Thely nica | + Campy sphe | + Asple cirr | + Tecta athy/riva | + Sacco inae | + Trich coll |
| + Lomar vest | + Tecta athy/riva | | + Polyp lori | + Salpi sp1 | |
| + Selag arth | + Selag arth | | + Bolbi nico | + Dipla stri | |
| + Tecta athy/riva | + Dipla stri | | + Dipla stri | | |
| | + Bolbi nico | | | | |
| | + Tecta sp7 | | | | |
| | + Pteris sp2 | | | | |
| − Polyb vill | − Danae wend | − Polyb alfr | − Adian obli | − Danae wend | − Polyp lori |
| − Salpi sp1 | | − Polyp lori | − Polyb vill | − Lomar vest | − Trich eleg |
| − Adian obli | | − Polyb osmu | − Sacco inae | − Asple cirr | − Danae medi |
| − Sacco inae | | − Elaph sp9 | − Salpi sp1 | − Trich eleg | − Cyath ursi |
| | | − Salpi sp1 | | − Alsop cusp | − Elaph sp9 |
| | | − Trich eleg | | | − Alsop cusp |
| | | − Tecta plan | | | − Polyb alfr |

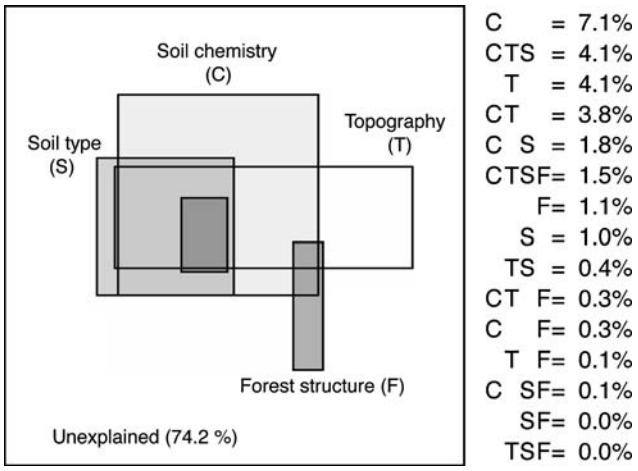


**Fig. 1** Partitioning of the variation in pteridophyte community composition using four subsets of environmental data: soil type (*S*), soil chemistry (*C*), forest structure (*F*) and topography (*T*). The *enclosing box* indicates total variation in composition, of which 25.8% was explained by the environmental datasets. The *rectangles* within the box approximately indicate the fraction of explained variation attributable to each environmental dataset (forest structure is divided into two separate rectangles to allow its illustration). The exact sizes of the unique contributions of each dataset, as well as their intersections, are listed to the *right* of the figure. All the testable model fractions (i.e. the unique contributions) were significant with $P = 0.001$ after 999 permutations

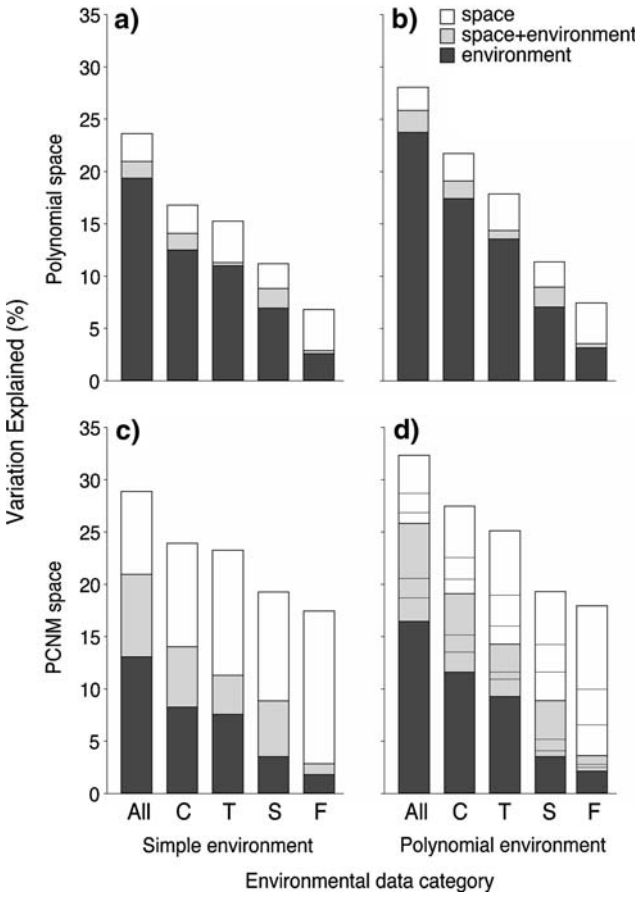


**Fig. 2** Variation in pteridophyte community composition explained using two different spatial models based on the *x* and *y* coordinates of the plots [a third-order polynomial (**a**, **b**) or principal coordinates of neighbour matrices (*PCNM*) variables (**c**, **d**)], and ten different environmental models (five categories of environmental data and two levels of model complexity). The environmental data categories were all data (*All*), soil chemistry alone (*C*), topography alone (*T*), soil type alone (*S*) and forest structure alone (*F*). Environmental model complexity refers to either a simple model of the selected environmental variables (**a**, **c**) or to a polynomial model including cubic and quadratic functions of the selected variables as well (**b**, **d**). For every spatial and environmental model combination, explained variation is partitioned into three fractions: purely spatial (*space*), spatially structured environmental (*space + environment*) and purely environmental (*environment*). For the most comprehensive model (**d**), the spatial and spatially structured environmental fractions of explained variation are further partitioned by spatial scale (from top to bottom: broad, medium, fine). All the testable model fractions (i.e. purely spatial or purely environmental fractions) were significant with $P = 0.001$ after 999 permutations

spatial scales (>650 m) in both the PS and SSE fractions. Environmental variables contributed slightly more at fine (100–300 m) than at medium scales, with the exception of soil type (Fig. 2d).

## Spatio-environmental patterns in pteridophyte community composition

In the RDA analysis where polynomial environmental data were used as explanatory variables, the five environmental variables with the largest partial contributions to axis 1 (henceforth environmental axis 1) were (soil Mg)$^2$, soil Ca, slope, topographic position, and soil pH. This axis can be interpreted as a floristic gradient from flat, relatively fertile swamp and other poorly drained sites at low topographic positions to sloping, relatively poor, well-drained sites at higher topographic positions (Fig. 3). Environmental axes 2 and 3, in contrast, reflect floristic variation independent of this swamp–upland gradient. Environmental axis 2 is interpretable as floristic responses to relatively fine-scale variation in exchangeable cations and topography, and environmental axis 3 reflects responses to variation in soil organic matter, especially between stream valleys and other sites. The five variables with the largest partial contributions to environmental axis 2 were (soil Ca)$^2$, (soil Ca)$^3$, (soil Mg)$^2$, (soil pH)$^2$ and (soil P)$^3$. Topographic position also made a sizeable contribution to this axis, ranking sixth. The five variables with the largest partial contributions to environmental axis 3 were soil C, soil N, (soil C)$^2$, slope and soil Ca. Canopy openness also made a relatively large contribution to this axis, ranking seventh. Generally, however, the contributions of the soil type and forest structural variables to all three environmental axes were minor.

In the RDA analysis where PCNM spatial data were used as explanatory variables, axis 1 (henceforth PCNM axis 1) clearly separated the largest swamp with its

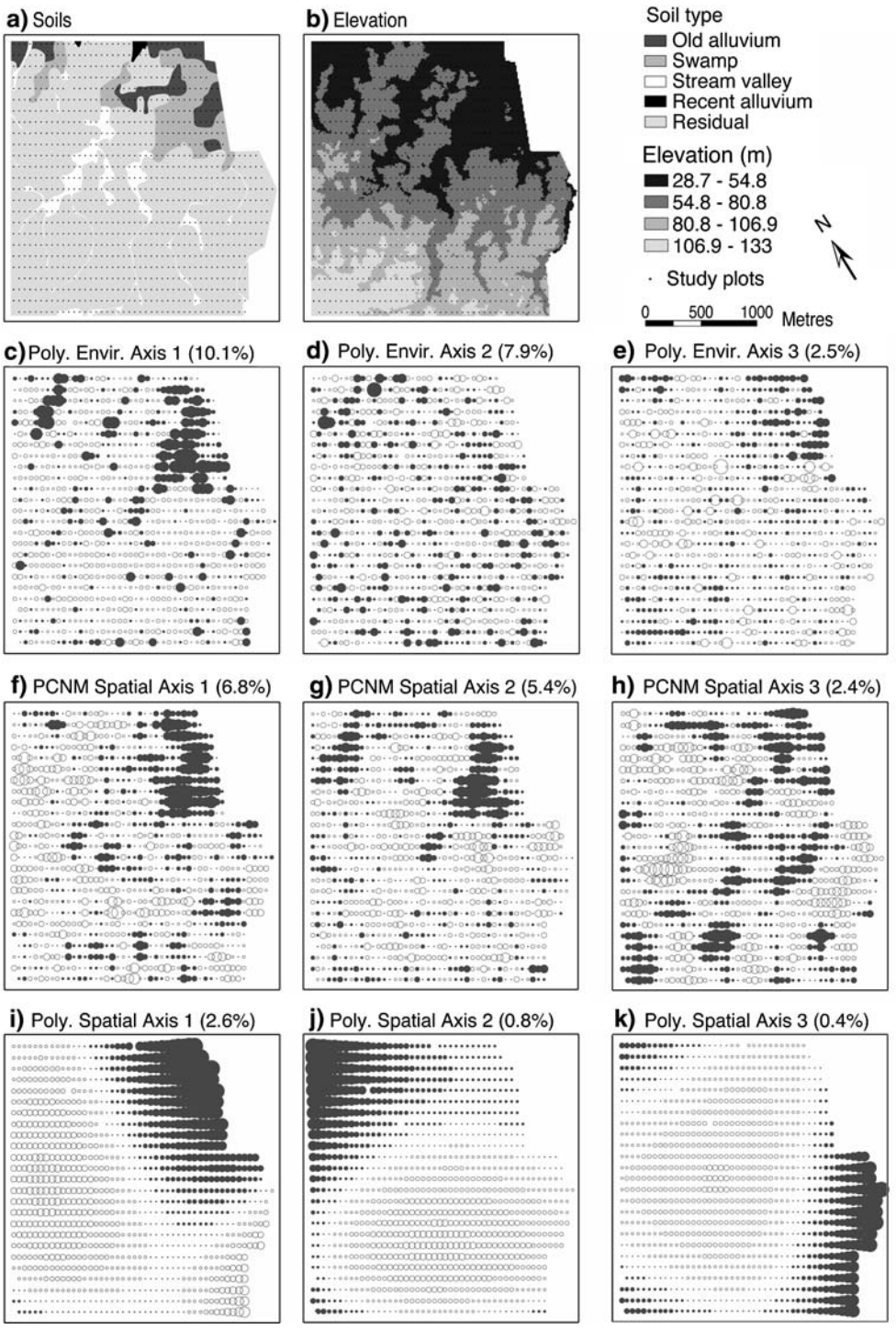


**Fig. 3** Distributions across the study site at La Selva Biological Station of **a** soil types, **b** elevation classes and **c**–**k** site scores of 100 $m^2$ pteridophyte sampling plots on the ordination axes 1–3 obtained in redundancy analyses. Site scores were obtained using as the explanatory dataset either the polynomial environmental (*Poly. Envir.*) model (**c**–**e**), the PCNM spatial model (**f**–**h**) or the polynomial spatial (*Poly. Spatial*) model (**i**–**k**). The *filled circles* indicate positive values, and the *open circles* negative values. The site scores represent the main gradients detected in species composition, as predicted by a linear combination of the explanatory variables. The proportion of variation in species composition explained by each axis ($R^2_a$) is given *in parentheses*

surroundings and some stream valleys both from many upland areas and from the two smaller swamps, one of which was largely treeless (Fig. 3). PCNM axis 2 was visibly similar to environmental axis 1, and also mainly reflected the gradient from poorly drained soils to uplands, whereas PCNM axis 3 was less environmentally interpretable, although it showed some correspondence with environmental axis 3 (Fig. 3). Although environmental variables were not included in the analysis, there was a strong positive relationship between site scores on PCNM axes 1 and 3 and the swamp soil type and soil Ca concentration, respectively, and a strong negative relationship between site scores on PCNM axis 2 and the residual soil type (all $P < 0.001$ in multiple regression analysis).

In the RDA analysis where polynomial spatial data were used as explanatory variables, site scores on axis 1 (hence-

forth polynomial axis 1) showed a much coarser spatial pattern that resembled that of environmental axis 1 and PCNM axes 1 and 2 (Fig. 3). Among the environmental variables, soil Ca was most strongly and positively related to site scores on polynomial axis 1, the residual soil type negatively to scores on polynomial axis 2, and soil P positively to scores on polynomial axis 3 (all $P < 0.001$).

Recording the ten species with the highest species scores on each of the environmental and PCNM axes produced a list of 26 species, of which 19 had high scores along more than one axis (Table 1). For eight species (e.g. *Polybotrya villosula*, *Danaea wendlandii* and *Thelypteris nicaraguensis*; Fig. 4), observed spatio-environmental distribution patterns seem to represent a primary relationship with the swamp to non-swamp gradient (environmental axis 1 or PCNM axis 2), and a secondary relationship with edaphic–topographic variation in non-swamp areas (environmental axis 2 or 3; Table 1). Four species (e.g. *Trichomanes collariatum*; Fig. 4), had spatial distribution patterns strongly reflected in both PCNM axes 1 and 2 (Table 1), which indicates a main distributional bias either towards or away from the largest, forested swamp and its surroundings, and a secondary bias related to swamp soils more generally. The spatial distributions of four species (e.g. *Polybotrya villosula, Polybotrya osmundacea* and *Lomariopsis vestita*; Fig. 4), were strongly reflected in PCNM axis 1 and secondarily in PCNM axis 3 (Table 1). The distributions of a further suite of species did not indicate any strong bias along the swamp–upland gradient, but these were instead associated with stream and other humid valleys (e.g. *Polypodium loriciforme*, with a relatively high species score on both environmental axis 3 and PCNM axis 3; Fig. 4, Table 1).

Many species pairs had complementary distribution patterns, often resulting from biases either towards or away from swamp-like conditions (e.g. *Thelypteris nicaraguensis* vs. *Danaea wendlandii*; Fig. 4). The three *Polybotrya* species also had notably contrasting distributions. *Polybotrya villosula* contrasted with *Polybotrya osmundacea* along PCNM axis 1 (highly negative vs. positive species scores, respectively), and with *Polybotrya alfredii* along PCNM axis 3 (highly positive vs. negative species scores, respectively). Whereas *Polybotrya villosula* had high species scores on environmental axes 1 and 2, *Polybotrya alfredii* had a high species score on environmental axis 3 (Fig. 4, Table 1).

## Discussion

### Ecological interpretation of community variability

With the variables at hand, we were able to explain up to 32% of pteridophyte community variation (after correcting for sample size and the number of explanatory variables; Peres-Neto et al. 2006). The main axes of floristic differentiation could roughly be characterized as differences between swamps and uplands, between open and forested swamps, between ridge tops and valleys, and between sites varying in their organic matter deposition and proximity to streams. Especially soil pH, soil concentrations of Ca, Mg, C and N, and slope angle and relative topographic position were strongly related to these major axes of floristic variation. The distributions of numerous pteridophyte species reflected more than one of these gradients. Soil Ca and Mg contents have been identified as important in several earlier studies of pteridophyte communities in Amazonian forests, at spatial scales ranging from metres to kilometres (e.g. Tuomisto et al. 2003a; Costa et al. 2005; Poulsen et al. 2006). Topographic variation was another major factor both in these studies and in a recent temperate forest study at a similar scale to ours (Karst et al. 2005). Responses to soil

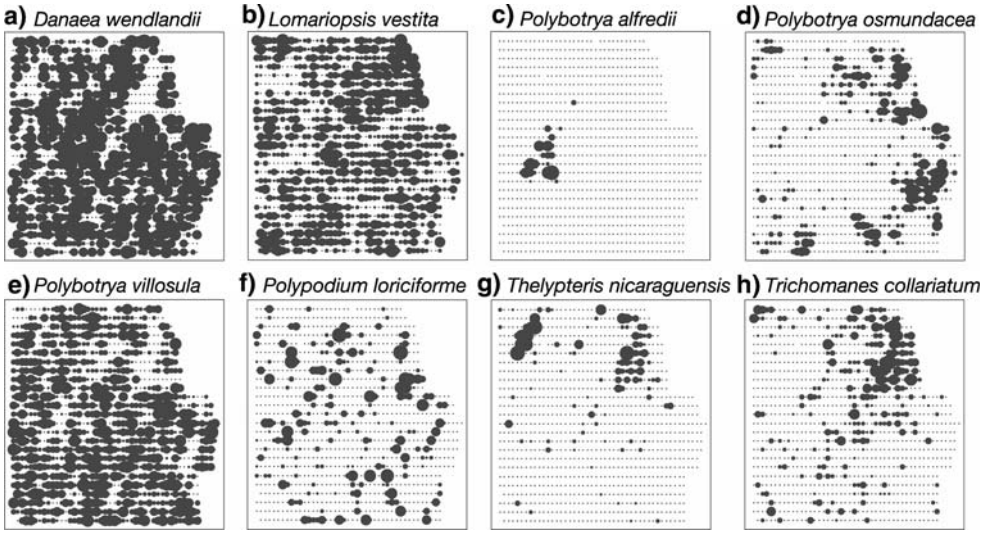


**Fig. 4** **a**–**h** Distribution maps of the eight pteridophyte species discussed in the text. *Small dots* represent the locations of 100-$m^2$ study plots at La Selva Biological Station. *Larger circles* indicate presence of the species in question, with the size of the circle proportional to the Hellinger-transformed abundance of the species

pH have often been less apparent (except in Karst et al. 2005) than at our site, where pH variation is strongly linked with the main swamp–upland gradient and exchangeable cation contents. When soil C and N (or $NO_3^-$) concentrations have been investigated, they have also been found important for explaining fern distributions (Costa et al. 2005; Karst et al. 2005).

The pteridophyte species whose distributions were most strongly related to the RDA axes differed widely in their relationships with environmental variables. Many species had contrasting distributions, which in some cases clearly reflected specific environmental variables (e.g. *Thelypteris nicaraguensis* and *Danaea wendlandii* showed opposite associations with swamp and upland soils). In other cases the relationships were less obvious. For example, the rare *Polybotrya alfredii* was restricted to a single valley, whereas congeneric *Polybotrya villosula* and *Polybotrya osmundacea* were abundant but mutually negatively associated elsewhere. The distribution of *Polybotrya villosula* showed strong environmental structuring, but that of *Polybotrya osmundacea* did not. Without information about the distributions of these species over time, and over broader spatial scales and longer ecological gradients (cf. Tuomisto 2006), it is difficult to draw conclusions about the relative roles of niche differentiation and other factors in determining these patterns.

Spatial structure in community composition was evident at all scales encompassed by our study design. The environmental component to spatially explained variation in community composition was strongest at broad scales (650–2,000 m), but was stronger at fine (100–350 m) than at intermediate scales. This pattern probably reflects the spatial configuration of environmental conditions at our site. Much of the broad-scale spatial variability is attributable to floristic differences between the largest swamp and other areas. The detected fine-scale spatial variability is more likely related to differences in soil fertility, drainage, and air humidity along topographic gradients.

The complex and varied patterns of floristic variation identified here, and the fact that these were strongly associated with quantitative variation in soil chemistry, suggest that habitat specificity in this community would be severely underestimated if habitat were characterized by soil type or topographic position alone, as has often been done in rain forest studies (Harms et al. 2001; Cannon and Leighton 2004; Valencia et al. 2004; but see Hall et al. 2004; John et al. 2007). By extension, fewer rain forest species may be habitat generalists than earlier studies have proposed.

The purely spatial fraction of explained variation has sometimes been interpreted as predominantly a dispersal effect (e.g. Gilbert and Lechowicz 2004; Cottenie 2005; Karst et al. 2005), but we do not believe this to be the case in our study. We suspect that this fraction had a considerable environmental component, which was not detected because some relevant environmental variables were omitted even from our most comprehensive dataset. For example, the main spatial pattern detected in our floristic data corresponded to the distinction between forested swamp conditions as opposed to open swamp and uplands, but this was not well captured by our environmental data. Temporal variation in environmental conditions, caused by gap dynamics or climatic variability, may also produce a spatial pattern that a snap shot environmental dataset cannot represent.

A relatively large proportion (at least 68%) of community variation in our dataset was unexplained by either environmental or spatial data. Undoubtedly, this is partly due to random spore dispersal and mortality, but it may also include deterministic variation caused by unmeasured environmental variables. Moreover, the role of processes operating at finer scales than those covered by a study's sampling design cannot be quantified. These may be very important at our site, as strong turnover in pteridophyte species composition has been identified at distances less below 100 m (Jones et al. 2006). Some local turnover is visibly related to environmental variation, but distance-limited spore dispersal and interspecific interactions are also likely to be strongest at short distances.

## Data quality and variation partitioning

We obtained very different variation partitioning results depending on which of 20 alternative combinations of environmental and spatial data we used as explanatory variables. The total proportion of explained floristic variation varied more than fourfold, as did the proportion explained by space, and the proportion explained by the environment varied ninefold. The unique contribution of the environment varied 12-fold, and the unique contribution of space fivefold. When the results are interpreted in terms of the relative importance of space versus environment, the ratio of the PS to PE fractions is of particular interest. This ratio ranged from 1:11 to 8:1 (or excluding the forest structural model, which was our most poorly performing environmental model, from 1:11 to 3:1). This shows that the main result of a study can easily be reversed by model choice.

Given that the variation partitioning method is an extension of multiple regression, different explanatory models can be expected to give somewhat different results. However, the magnitude of this effect has been underestimated or overlooked in recent comparisons (e.g. Balvanera et al. 2002; Cottenie 2005; Chust et al. 2006).

We found that over two-thirds of the spatial floristic variation detected by the PCNM model was undetected by the polynomial model. Of the four environmental data types, soil chemical data had the highest power to explain floristic

variation, followed by topographic, soil type and forest structural descriptors. Although these environmental data types were partly redundant, total environmentally explainable variation would have been reduced by a third if had soil chemistry been omitted from our study.

How do these results compare with earlier floristic variation partitioning studies? Most studies where space explained more variation than environment used no data on soil chemistry (Borcard et al. 1992; Svenning et al. 2004 for trees; Chust et al. 2006). Svenning and Skov (2005) provide an exception, but their data were derived from coarse-scale maps rather than actual soil sampling. In contrast, in those studies where environment explained more variation than space, either soil chemical data were included (Duivenvoorden 1995; Gilbert and Lechowicz 2004; Duque et al. 2005; Karst et al. 2005; possibly Cottenie 2005 in some cases), disturbed landscapes were included (Dalle et al. 2002), or the spatial model was especially coarse (Duivenvoorden 1995; Balvanera et al. 2002). In these cases, model choice seems to be a strong predictor of the analysis results.

Austin (2002) suggested that canonical ordination analyses would yield ecologically more meaningful results by enabling non-linear functions of environmental variables. In the present study, which covered a limited range of ecological variation, including polynomials of the environmental variables increased total environmentally explained variation by between one-quarter and one-third. In broader scale studies that encompass a wider range of environmental conditions, the results of linear and non-linear methods can be expected to diverge much more.

In addition to differing in their explanatory variables, community studies have also differed in their response variable (species presence–absence or abundance data, focal taxa), in the ordination method applied (RDA or CCA), and in spatial extent, spatial resolution, and the environmental gradients they cover. The overall proportion of explained variation should be adjusted for the number of sampling sites and explanatory variables (Legendre et al. 2005; Peres-Neto et al. 2006), but this adjustment has yet to be commonly implemented. Consequently, it is almost impossible to evaluate to what degree differences in the results among studies are methodological, and to what degree they reflect real differences among focal plant groups or geographical areas. Studies applying consistent methods in cross-site and cross-taxon analyses would be of particular value for resolving these issues.

## Conclusion

We found evidence of strong environmental control of beta-diversity in Costa Rican rain forest pteridophytes. However, the explanatory power of environmental and spatial variables together varied between 7 and 32%, and the relative importance of "space" and "environment" could be reversed by model choice. This leads us to the following conclusions about the interpretation of variation partitioning results, and recommendations for future studies:

1. Ecological background knowledge is needed when selecting environmental variables to avoid omitting key factors. Plant growth is known to depend on the availability of various nutrients, water and light, so quantitative descriptors of these should be included in floristic studies. Incorporating non-linear relationships between floristic and environmental variation may also be needed, especially if the sampled environmental gradient is long. Results cannot be assumed to reflect the effect of "the environment" in general, unless all potentially relevant environmental variables have been adequately modelled.
2. For the adequate modelling of "space", a sufficiently flexible spatial model is needed. A simple spatial model, such as one based on $x$ and $y$ coordinates, or polynomial functions of these, will only be able to represent broad-scale spatial patterns. The ability to detect spatial pattern will also depend on the sampling setup, such as interplot distances and the spatial arrangement of the plots.
3. $R^2$ adjustment needs to be applied to eliminate the influence of sample size and the number of explanatory variables on the proportion of variation explained.
4. Great care needs to be taken in interpreting the results, as these are subject to constraints imposed by the dataset and the methods applied. Both generalisations from a particular study and comparisons across studies need to carefully consider these constraints. Biological meaning can only be separated from methodological artefacts if the impact of data quality on the results is recognised.

**Acknowledgements** We thank Rigoberto Gonzalez for assisting during the fern inventory, and Jens Mackensen and Edzo Veldkamp for soil data collection. Diana and Milton Lieberman kindly allowed us access to their study areas. La Selva Biological Station of the Organization for Tropical Studies provided logistic support. Jérôme Chave, Michael Kessler and three anonymous reviewers provided helpful comments on the manuscript. The work was funded by grants from the Academy of Finland (to H. Tuomisto), the Andrew W. Mellon foundation (to D. B. and D. A. Clark), and NSERC (grant number OGP0007738 to P. Legendre). Inventories and specimen collection complied with Costa Rican law. Research permits were kindly granted by SINAC-MINAE.

## References

Arbeláez MV, Duivenvoorden JF (2004) Patterns of plant species composition on Amazonian sandstone outcrops in Colombia. J Veg Sci 15:181–188

Austin MP (2002) Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. Ecol Modell 157:101–118
Balvanera P, Lott E, Segura G, Siebe C, Islas A (2002) Patterns of $\beta$-diversity in a Mexican tropical dry forest. J Veg Sci 13:145–158
Blanchet G, Legendre P, Borcard D (2007) Forward selection of explanatory variables. Ecology (in press)
Borcard D, Legendre P (2002) All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices. Ecol Modell 153:51–68
Borcard D, Legendre P (2004) SpaceMaker2—user's guide. Département de sciences biologiques, Université de Montréal. http://www.bio.umontreal.ca/legendre/
Borcard D, Legendre P, Avois-Jacquet C, Tuomisto H (2004) Dissecting the spatial structure of ecological data at multiple scales. Ecology 85:1826–1832
Borcard D, Legendre P, Drapeau P (1992) Partialling out the spatial component of ecological variation. Ecology 73:1045–1055
ter Braak CJF, Smilauer P (1998) CANOCO reference manual and user's guide to CANOCO for Windows. Software for canonical community ordination. Version 4. Centre for Biometry, Wageningen
Brown N, Jennings S, Wheeler P, Nabe-Nielsen J (2000) An improved method for the rapid assessment of forest understorey light environments. J Appl Ecol 37:1044–1053
Cannon CH, Leighton M (2004) Tree species distributions across five habitats in a Bornean rain forest. J Veg Sci 15:257–266
Casgrain P (2001) Permute! Version 3.4 user's manual. Département de sciences biologiques, Université de Montréal, Montreal
Chust G, Chave J, Condit R, Aguilar S, Lao S, Pérez R (2006) Determinants and spatial modeling of tree $\beta$-diversity in a tropical forest landscape in Panama. J Veg Sci 17:83–92
Clark DA, Clark DB, Sandoval M R, Castro C MV (1995) Edaphic and human effects on landscape-scale distributions of tropical rain forest palms. Ecology 76:2581–2594
Clark DB, Palmer MW, Clark DA (1999) Edaphic factors and the landscape-scale distributions of tropical rain forest trees. Ecology 80:2662–2675
Corney PM, Le Duc MG, Smart SM, Kirby KJ, Bunce RGH, Marrs RH (2006) Relationships between the species composition of forest field-layer vegetation and environmental drivers assessed using a national scale survey. J Ecol 94:383–401
Cottenie K (2005) Integrating environmental and spatial processes in ecological community dynamics. Ecol Lett 8:1175–1182
Dalle SP, López H, Díaz D, Legendre P, Potvin C (2002) Spatial distribution and habitats of useful plants:an initial assessment for conservation on an indigenous territory Panama. Biodivers Conserv 11:637–667
Dalling JW, Muller-Landau HC, Wright SJ, Hubbell SP (2002) Role of dispersal in the recruitment limitation of neotropical pioneer species. J Ecol 90:714–727
Denslow JS (1987) Tropical rainforest gaps and tree species diversity. Annu Rev Ecol Syst 18:431–451
Dirzo R, Horvitz CC, Quevedo H, López MA (1992) The effects of gap size and age on the understorey herb community of a tropical Mexican rain forest. J Ecol 80:809–822
Dray S, Legendre P, Peres-Neto PR (2006) Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbour matrices (PCNM). Ecol Modell 196:483–493
Duivenvoorden JF (1995) Tree species composition and rain forest-environment relationships in the middle Caquetá area Colombia NW Amazonia. Vegetatio 120:91–113
Duque A, Sánchez M, Cavallier J, Duivenvoorden JF (2002) Different floristic patterns of woody understorey and canopy plants in Colombian Amazonia. J Trop Ecol 18:499–525
Duque AJ, Duivenvoorden JF, Cavelier J, Sanchez M, Polanía C, León A (2005) Ferns and Melastomataceae as indicators of vascular plant composition in rain forests of Colombian Amazonia. Plant Ecol 178:1–13
Fine PVA, Mesones I, Coley PD (2004) Herbivores promote habitat specialization by trees in Amazonian forests. Science 305:663–665
Gilbert B, Lechowicz MJ (2004) Neutrality niches and dispersal in a temperate forest understory. Proc Natl Acad Sci USA 101:7651–7656
Hall JS, McKenna JJ, Ashton PMS, Gregoire TG (2004) Habitat characterizations underestimate the role of edaphic factors controlling the distribution of *Entandrophragma*. Ecology 85:2171–2183
Harms KE, Condit R, Hubbell SP, Foster RB (2001) Habitat associations of trees and shrubs in a 50-ha neotropical forest plot. J Ecol 89:947–959
Hofton MA, Rocchio LE, Blair JB, Dubayah R (2002) Validation of Vegetation Canopy Lidar sub-canopy topography measurements for a dense tropical forest. J Geodyn 34:491–502
Holmgren PK, Holmgren NH (1998 onwards) [continuously updated]. Index Herbariorum. New York Botanical Garden. http://sciweb.nybg.org/science2/IndexHerbariorum.asp
Hubbell SP (2001) The unified neutral theory of biodiversity and biogeography. Princeton University Press, Princeton
John R, Dalling JW, Harms KE, Yavitt JB, Stallard RF, Mirabello M, Hubbell SP, Valencia R, Navarrete H, Vallejo M, Foster RB (2007) Soil nutrients influence spatial distributions of tropical tree species. Proc Natl Acad Sci USA 104:864–869
Jones MM, Tuomisto H, Clark DB, Olivas P (2006) Effects of mesoscale environmental heterogeneity and dispersal limitation on floristic variation in rain forest ferns. J Ecol 94:181–195
Karst J, Gilbert B, Lechowicz MJ (2005) Fern community assembly: the roles of chance and the environment at local and intermediate scales. Ecology 86:2473–2486
Legendre P, Gallagher ED (2001) Ecologically meaningful transformations for ordination of species data. Oecologia 129:271–280
Legendre P, Borcard D, Peres-Neto PR (2005) Analyzing beta diversity: partitioning the spatial variation of community composition data. Ecol Monogr 74:435–450
Moran RC, Riba R (eds) (1995) Flora Mesoamericana, vol 1. Psilotaceae a Salviniaceae. Universidad Nacional Autónoma de México, Mexico
Oksanen J, Kindt R, Legendre P, O'Hara RB (2007) vegan: community ecology package version 1.8–5. http://cc.oulu.fi/~jarioksa/
Peres-Neto PR, Legendre P, Dray S, Borcard D (2006) Variation partitioning of species data matrices: estimation and comparison of fractions. Ecology 87:2614–2625
Phillips OL, Núñez Vargas P, Lorenzo Monteagudo A, Peña Cruz A, Chuspe Zans M-E, Galiano Sánchez W, Yli-Halla M, Rose S (2003) Habitat association among Amazonian tree species: a landscape-scale approach. J Ecol 91:757–775
Potts MD, Ashton PS, Kaufman LS, Plotkin JB (2002) Habitat patterns in tropical rain forests: a comparison of 105 plots in northwest Borneo. Ecology 83:2782–2797
Poulsen AD, Tuomisto H, Balslev H (2006) Edaphic and floristic variation within a 1-ha plot of lowland Amazonian rain forest. Biotropica 38:468–478
R Development Core Team (2006) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, ISBN 3–900051-07-0, http://www.R-project.org
Svenning J-C, Kinner DA, Stallard RF, Engelbrecht BMJ, Wright SJ (2004) Ecological determinism in plant community structure across a tropical forest landscape. Ecology 85:2526–2538
Svenning J-C, Skov F (2005) The relative roles of environment and history as controls of tree species composition and richness in Europe. J Biogeogr 32:1019–1033
Thomsen RP, Svenning J-C, Balslev H (2005) Overstorey control of understorey species composition in a near-natural temperate broadleaved forest in Denmark. Plant Ecol 181:113–126

Tuomisto H (2006) Edaphic niche differentiation among *Polybotrya* ferns in Western Amazonia: implications for coexistence and speciation. Ecography 29:273–284

Tuomisto H, Poulsen AD (2000) Pteridophyte diversity and species composition in four Amazonian rain forests. J Veg Sci 11:383–396

Tuomisto H, Ruokolainen K, Kalliola R, Linna A, Danjoy W, Rodriguez Z (1995) Dissecting Amazonian biodiversity. Science 269:63–66

Tuomisto H, Poulsen AD, Ruokolainen K, Moran RC, Quintana C, Celi J, Cañas G (2003a) Linking floristic patterns with soil heterogeneity and satellite imagery in Ecuadorian Amazonia. Ecol Appl 13:352–371

Tuomisto H, Ruokolainen K, Yli-Halla M (2003b) Dispersal environment and floristic variation of Western Amazonian forests. Science 299:241–244

Valencia R, Foster RB, Villa G, Condit R, Svenning J-C, Hernández C, Romoleroux K, Losos E, Magård E, Balslev H (2004) Tree species distributions and local habitat variation in the Amazon: large forest plot in eastern Ecuador. J Ecol 92:214–229

Vormisto J, Svenning J-C, Hall P, Balslev H (2004) Diversity and dominance in palm (Arecaceae) communities in terra firme forests in the western Amazon basin. J Ecol 92:577–588

Wyatt JL, Silman MR (2004) Distance-dependence in two Amazonian palms: effects of spatial and temporal variation in seed predator communities. Oecologia 140:26–35