B. Gilbert and J. R. Bennett: Partitioning variation in ecological communities: do the numbers add up?

Supporting Information

Simulated Communities. All simulated communities had *Es* and *Er* determinants of species distributions, and two types of communities had an additional dispersal component. The contribution of these components in a high dispersal community is shown to illustrate these relationships (Fig. S1). The steepness and magnitude of the dispersal curves after weighting are illustrated in Fig. S2. The cells sampled in each of the four non-contiguous sampling designs are illustrated in Fig. S3.

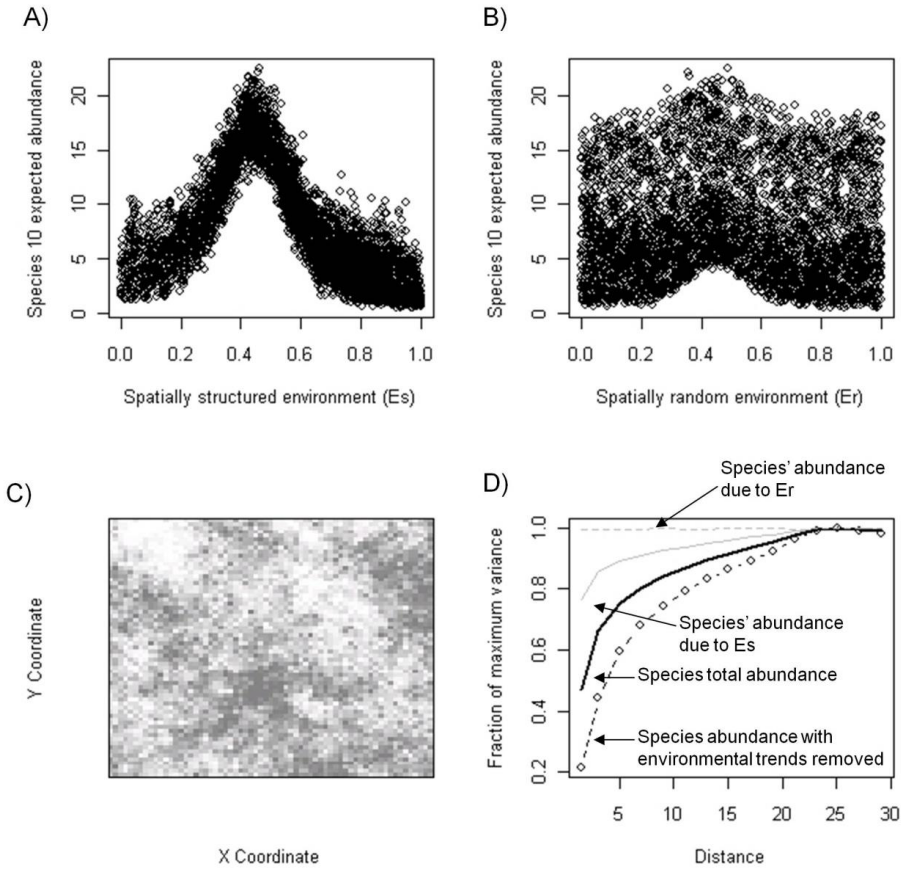


Fig. S1. An example of one species' distribution from a simulation of a community with a high space component. A) Expected abundance along the spatially structured environmental gradient (*Es*). B) Expected abundance along the spatially random environmental gradient (*Er*). C) Map of the spatially structured gradient, with a greyscale representation of the value of *Es* at each lattice cell (corresponding to one sample plot). D) Variance of each component of the species distribution as a function of distance. All values are divided by the maximum variance of the given driver to make them comparable. The dispersal function (i.e. the species' residual score when environmental gradients are removed) shows the largest and steepest change with distance, while the total expected change in the species' abundance has an intermediate change with distance.

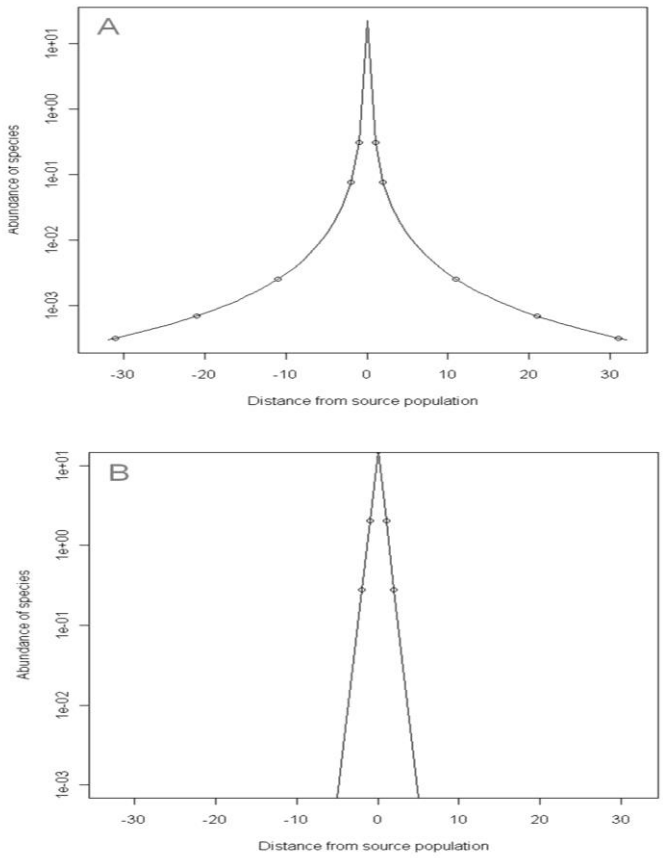


Fig. S2. The effect of the dispersal kernel after weighting. In both figures, the source population is at the center and represents the maximum source population possible for the community. Open circles show the dispersal effect at select cells, with the two closest showing the nearest neighbors (distance=1) and the next placed at distance=2. A) communities with a minor dispersal component ($w_s = 0.25, f \propto 1/\text{distance}^2$, w_e=0.75). B) communities with a larger dispersal component ($w_s = 0.5, f \propto e^{-\text{distance}/0.5}$, w_e=0.5). Note the magnitude of the effect in neighboring cells.

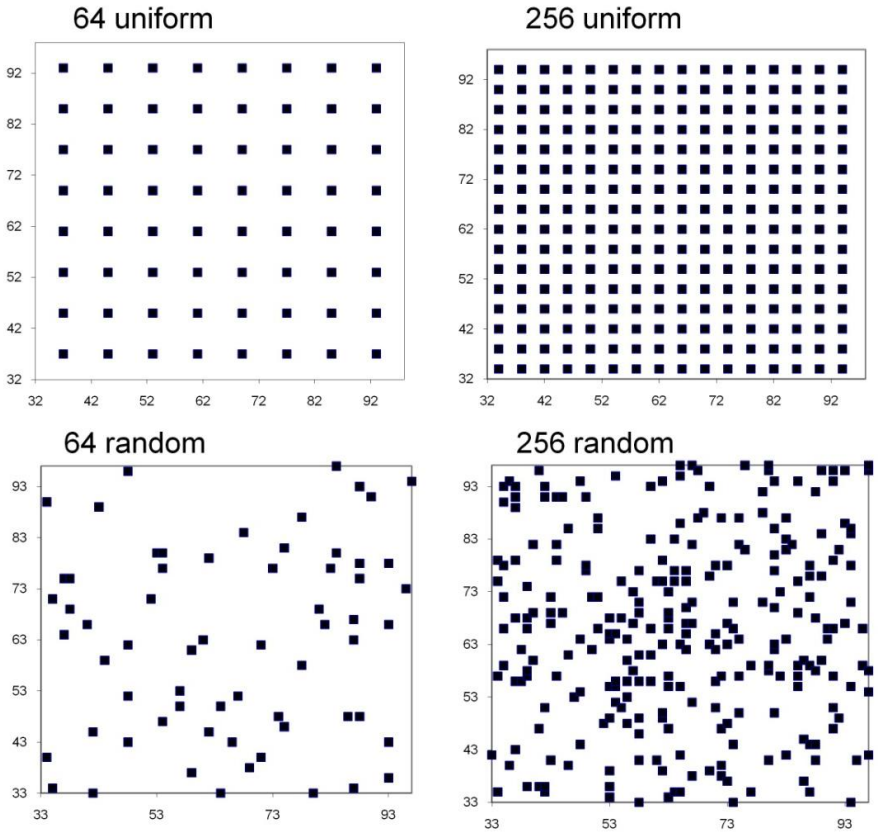
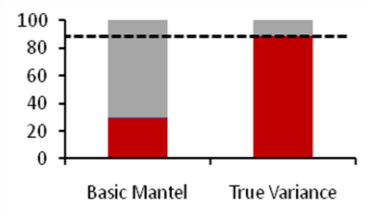
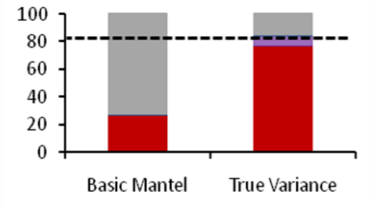


Fig. S3. Uniform and random sampling configurations used for analyses. Contiguous sampling (not shown) sampled either 64 (8×8) or 256 (16×16) adjacent cells from the centre of the simulated communities.

A - Community 1, no independent space component



B - Community 2, small independent space component



C - Community 3, large independent space component

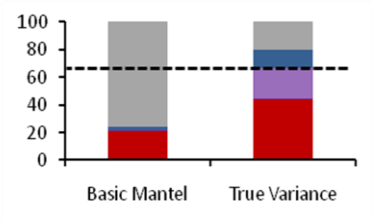


Fig. S4. Variation explained by the basic Mantel test, for all three simulated community types. Figure format as per Fig. 1 in the main article. The Mantel test underestimates the contributions of all components, and exhibits poor differentiation among the different community types.

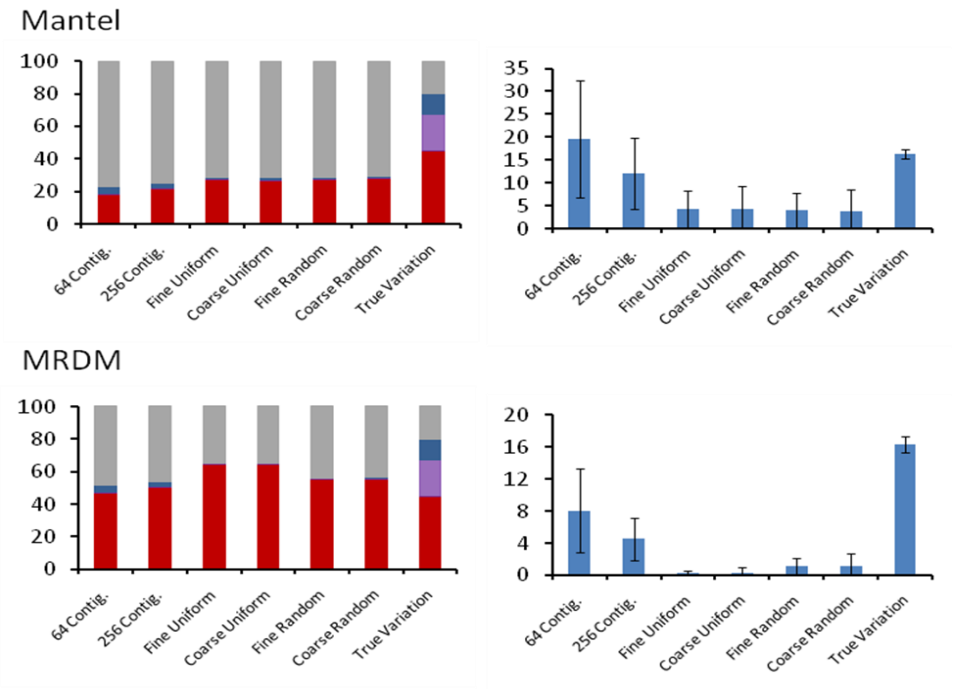


Fig. S5. The effect of different sampling methods on Mantel-type analyses: basic Mantel test and multiple regression on distance matrices (MRDM), for Community 3 (large independent spatial signal). Format as per Fig. 2 in the main article. Left hand graphs present absolute variation explained, while right hand graphs present the independent spatial signal as a proportion of explained variation: $S/(E+S+ES)$.

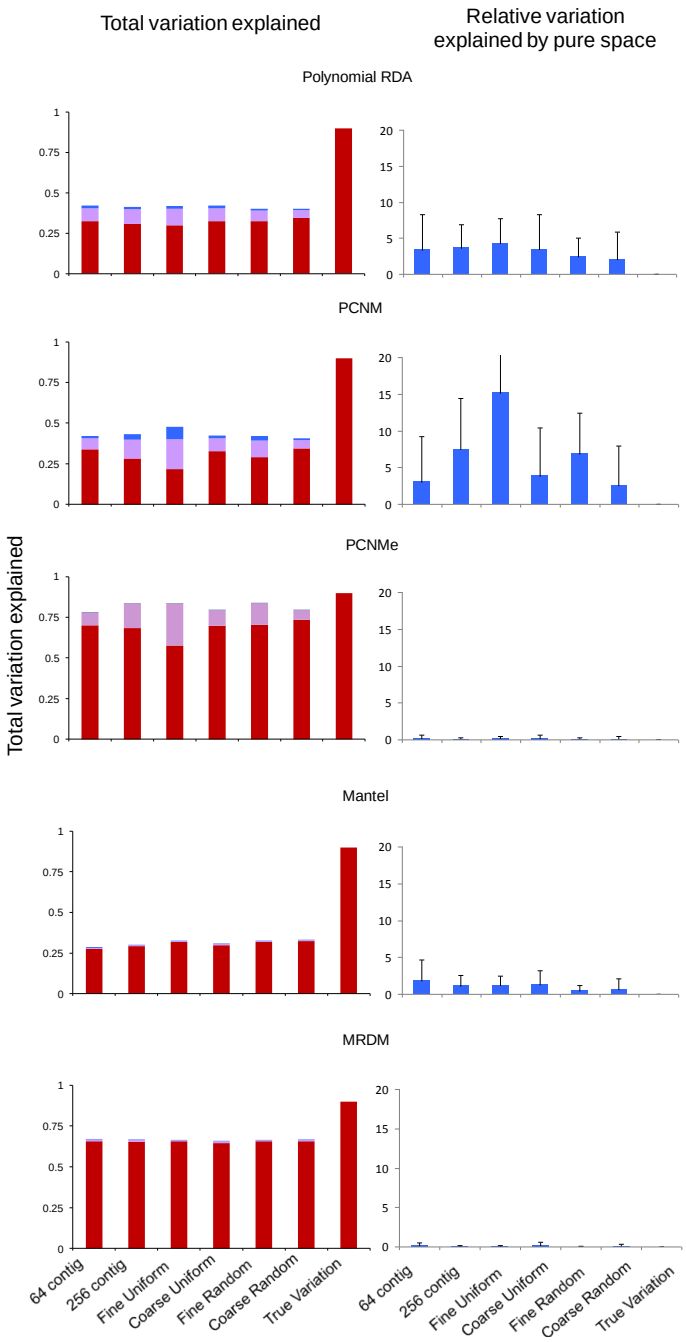


Fig. S6. The effect of the different sampling methods on all analyses for the simulated community with no true independent spatial signal. Format and segment colors as per Figs. 1 and 2 in the main article. Left hand graphs present absolute variation explained, while right hand graphs present the independent spatial signal as a proportion of explained variation: $S/(E+S+ES)$.

Further Evaluation of Eigenvector Techniques

Results from the PCNM, MEM and PCNMe analyses indicated that, when power was relatively low, the number of axes selected was bimodal (Fig. 3c inset in the main article). These results suggested that the forward selection of axes created a positive feedback whereby it was more likely to select subsequent axes once one had already been selected. To explore this possibility, we analyzed a simple linear gradient, similar to the unimodal analysis performed in Borcard and Legendre (2002). In particular, we created a linear transect with 200 sample locations (labeled 1 to 200) and a single species with a unimodal distribution, centered on the 100th site. The abundance of the species was a Poisson random variate with an expected mean (λ) given as:

$$\lambda = 20e^{-(100-location)/\sigma^2} + c \qquad \text{eqn S1}$$

Where σ is the standard deviation of the species distribution (i.e. determines the spread) and *c* is a constant. Because a Poisson distribution has a variance equal to the mean, *c* can be considered the known 'noise'. The spread of the species (σ) in this first simulation was 20 and *c* was set to zero. The "species matrix" was the ranged (λ/λ_{max}) abundance value, and a PCNM predictor matrix was created using the same procedure as for the main analyses (using QuickPCNM v. 7.7.1). An MEM predictor matrix was created in the same manner as for the main analyis, by selecting only those PCNM axes that had a significant, positive spatial structure (using QuickPCNM and spacemakeR). An example of simulated species abundances and predictor axes is illustrated in Fig. S7.

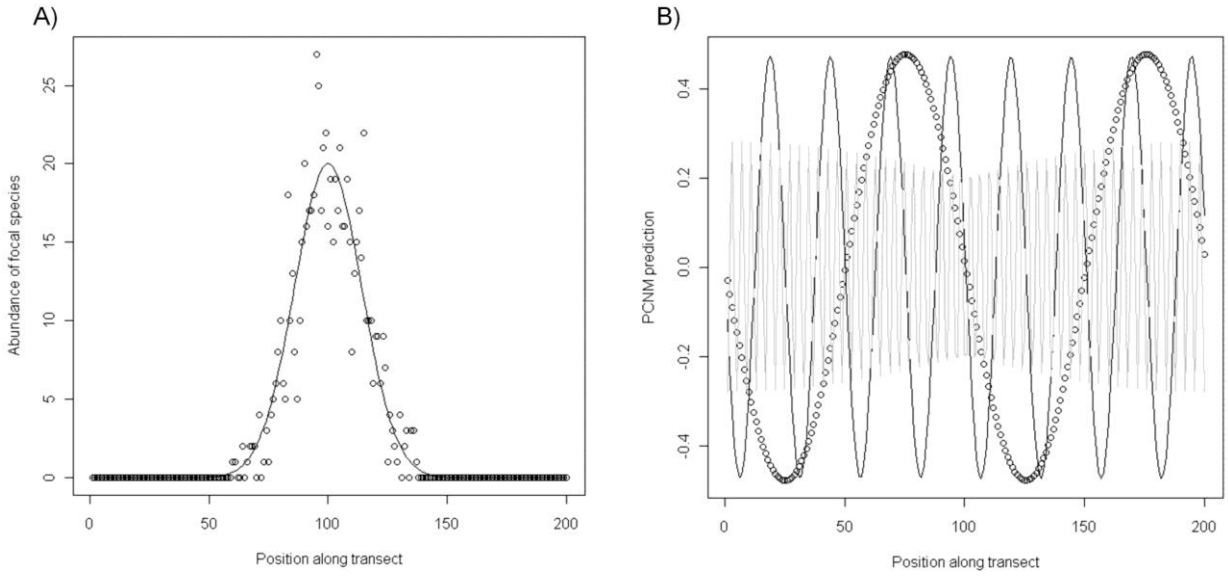


Fig. S7. Simple linear transect simulation to explore PCNM forward selection. A) The abundance of the species along the linear transect (points) with the expected abundance given by the solid line. B) Three PCNM axes (#3: open circles, #15: black line, #100: grey line).

To explore the behavior of the forward selection of eigenvector axes, we began by considering PCNM axes. We used the "packfor" library forward-selection procedure designed by

Blanchet, Legendre & Borcard (2008) to correct for inflated type 1 error in forward selection. We first calculated all axes that were significant using the Blanchet *et al.* procedure. We then tested the significance of each of these axes individually against the species matrix, and each axis that was significant *without* the influence of the forward-selection procedure was included as 'independently significant axes'. As an example, for the first simulation run (Fig. S7), 25 axes were significant through the forward-selection procedure. However, when these 25 axes were tested against the species matrix one at a time, only three were statistically significant (Fig. S8), indicating that 22 axes were 'selection-dependent'.

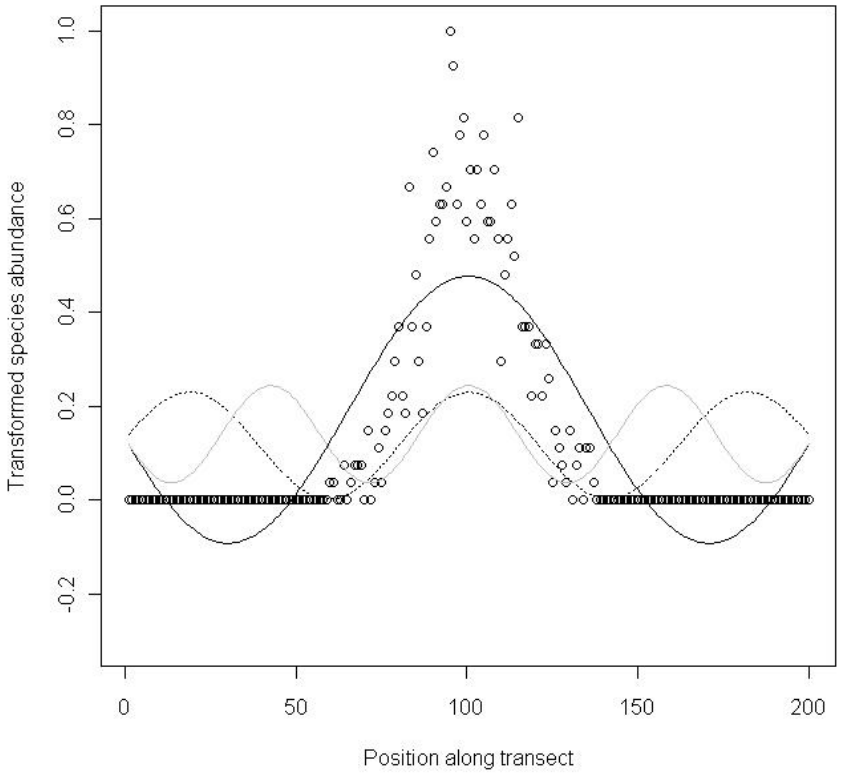


Fig. S8. Predicted fits for the three independently-significant PCNM axes (axis 2: black, axis 4: dotted, axis 6: grey) and the species abundance (open circles). The subsequent 22 PCNM axes selected in forward selection were not significant before fitting the three axes shown here.

The 22 selection-dependent PCNM axes that were selected in the forward-selection procedure, but were not independently significant, could be caused either by masking or a statistical artifact. It is not clear why, in this simple scenario, eigenvector axes would mask each other. However, when we looked at the residuals once the first three (independently) significant axes had been fit, it was apparent that the axes had created distinct patterns in the residuals (Fig. S9). In particular, the residuals at sites where the species was present showed a standard scatter around zero. However, residuals at sites where the species was absent showed a distinct sinusoidal pattern that was clearly an artifact of fitting the independently significant axes (Fig. S9). Given that the vast majority of species in most communities are relatively infrequent, and

thus will have zero abundance in many sample plots, this problem could be quite significant in real ecological studies.

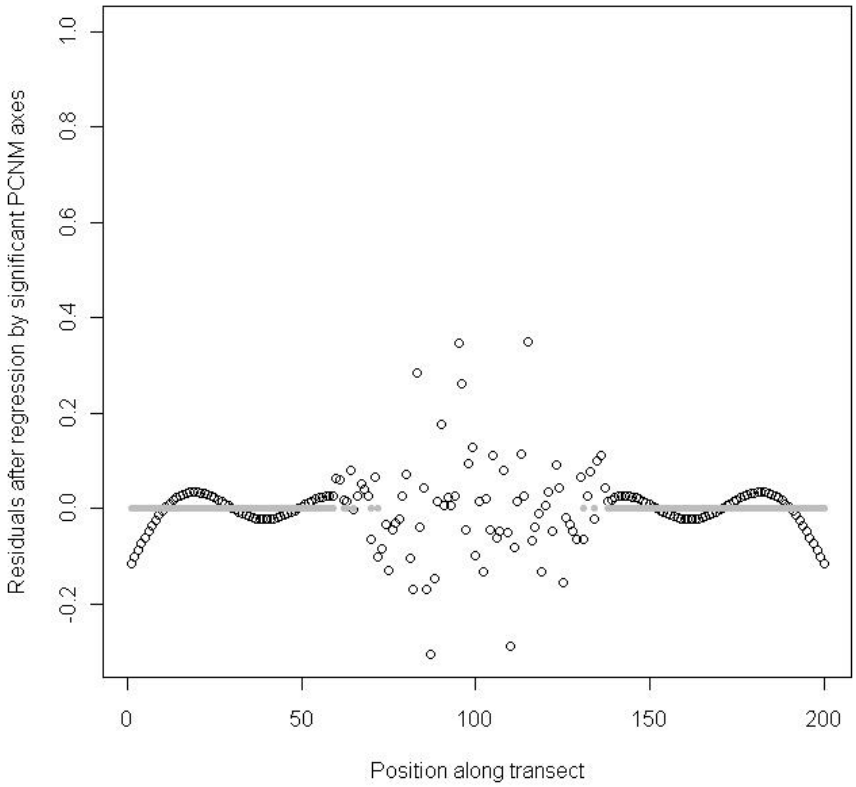


Fig. S9. The distribution of residuals (open circles) after the first three independently significant PCNM axes had been fit. The points in grey (flat line at zero) indicate points that originally had no individuals.

To test the effects of these additional axes on R^2 values, we considered scenarios in which we independently varied the background noise (*c* in eqn S1) and the abundances of the species (by varying σ in eqn S1). Our simulations showed that the patterns reported here were consistent; all simulations tended to have fewer independently significant PCNM axes than the number found in the forward-selection procedure. Moreover, both the abundance of the species and the level of unexplained variation determined the degree of overfitting by the PCNM (Figs. 4, S10). Overfitting was higher when species were more abundant (less sites with no occupancy). In addition, the effect of selection-dependent axes became more important as the unexplained variation increased (Fig. 4). Generally, the average number of selection-dependent PCNM axes was sufficient to cause the PCNM R^2 to be greater than the true R^2 (Fig. 4, S10). Species with a narrow band of occurrence had a large number of 'selection-influenced' axes, although these species tended to have less inflated R^2 than their more abundant counterparts, except when unexplained variation was high. The impact of these factors on both the inference of scale-

dependent effects and variation explained warrants further study in the more complex situations that are typical of ecological communities.

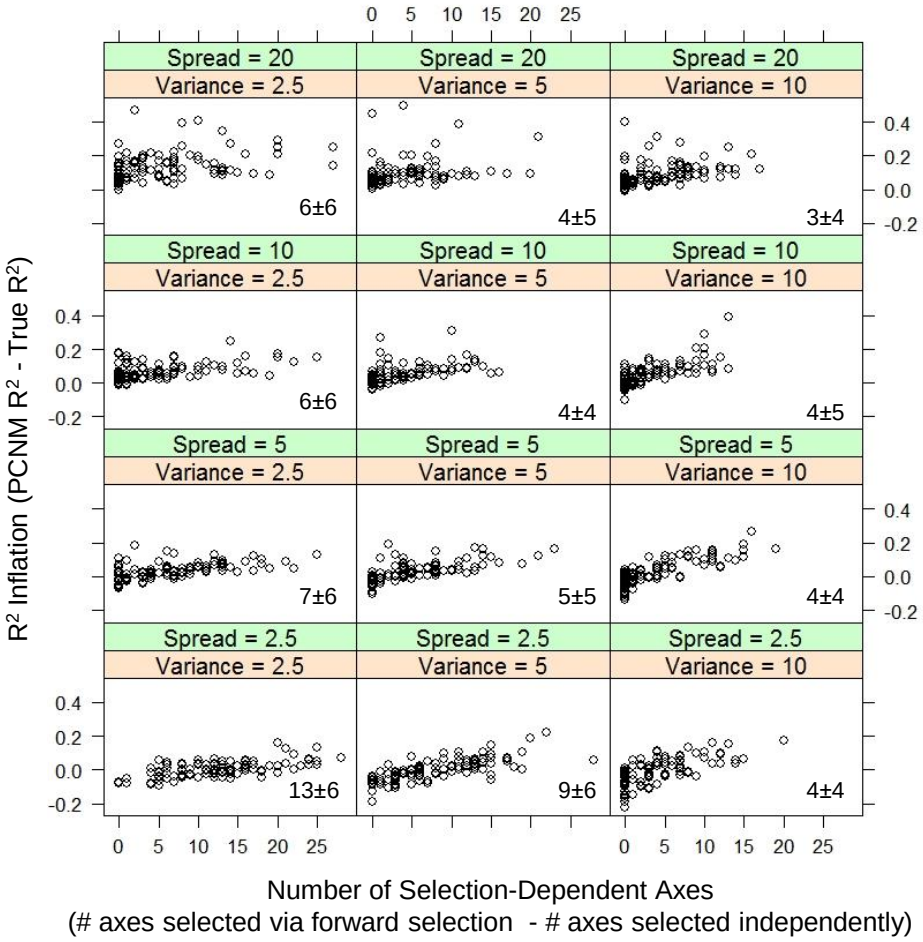


Fig. S10. Inflated R^2 from PCNM axes (PCNM R^2 – True R^2) with a value of zero indicating where the PCNM R^2 is equal to the true R^2. All results are from the single-species simulations along a linear transect, with each point representing a single simulation. The x-axis indicates the number of PCNM axes that were selected in the forward-selection procedure, but were not significant when tested individually. Each panel represents a different degree of species' spread (σ in eqn S1) and noise (labeled variance, *c* in eqn S1). Each panel shows points from 100 simulations, with the numbers in the bottom indicating the mean number of selection dependent axes (± standard deviation).

When these analyses were repeated with MEM axes, we found that the results from the PCNM analyses were generally consistent (Fig. S11). In particular, the MEM technique also tended to select more axes than were independently significant (Fig. S11). Although the MEM technique tended to be slightly more conservative than the PCNM technique, both tended to inflate the variance explained under similar conditions (compare Fig. S10 and S11).

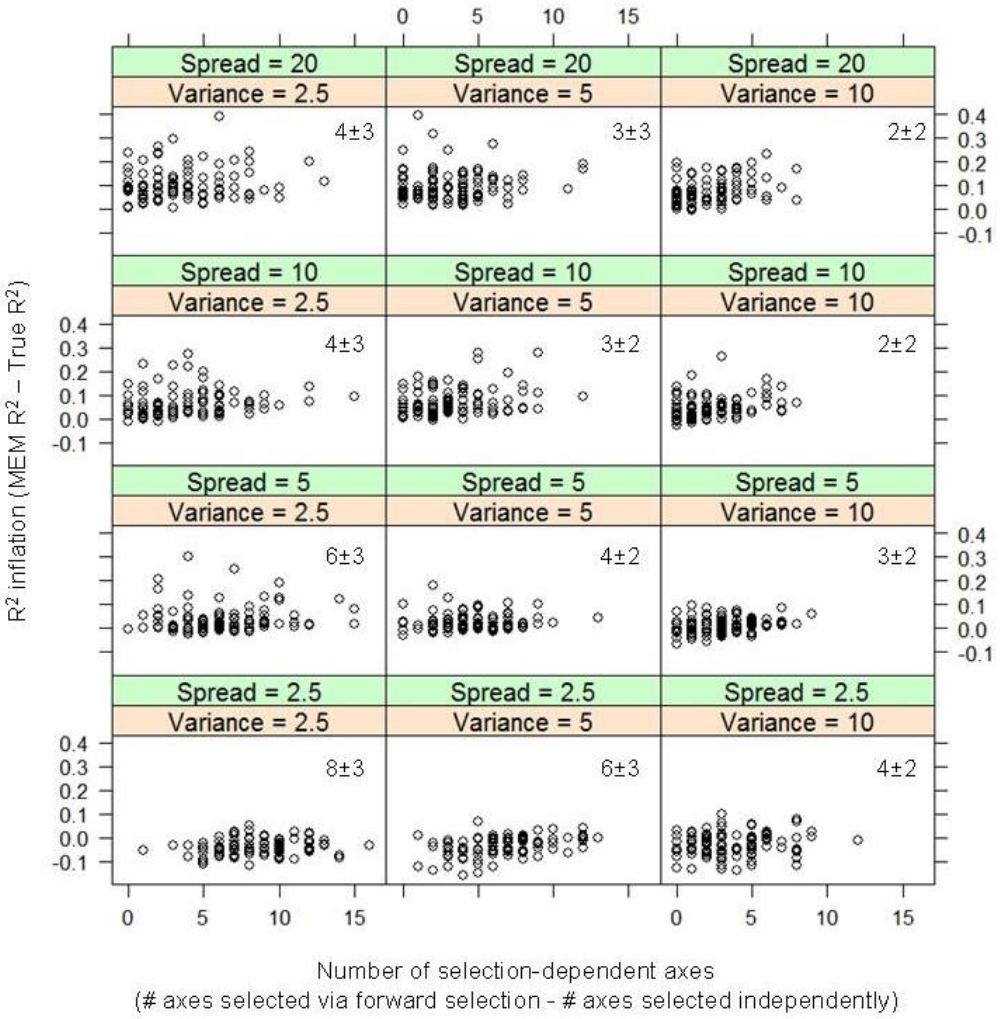


Fig. S11. Inflated R^2 from MEM axes (MEM R^2 – True R^2) with a value of zero indicating where the MEM R^2 is equal to the true R^2. Only MEM axes representing significant positive autocorrelation were included in the initial MEM matrix. All results are from the single-species simulations along a linear transect, with each point representing a single simulation. The x-axis indicates the number of MEM axes that were selected in the forward-selection procedure, but were not significant when tested individually. Each panel represents a different degree of species' spread (σ in eqn S1) and noise (labeled variance, *c* in eqn S1). Each panel shows points from 100 simulations.

Evaluation of Species-by-Species Forward-Selection Technique

Peres-Neto and Legendre (2010) noted that standard forward-selection procedures in canonical ordination may miss variables that are significant for only one or a few species in a dataset, a problem analogous to analysis of variance where a large number of sample means are compared, and the power to detect a difference among means is low if only one or a few sample means are different. They recommended a new forward-selection procedure: Once global significance is determined using all spatial (or environmental) variables, forward selection of

variables is conducted for each species, and the retained matrix of predictors includes all variables which have been forward selected for at least one species.

We evaluated this technique using the MEM approach to model spatial components. We conducted our analyses on the first 50 simulated communities of each community type for all sampling regimes, and compared our results to those using MEM with the standard (Blanchet, Legendre & Borcard 2008) forward-selection procedure. In every test scenario, the variance explained by the spatial component using the new forward-selection technique was inflated. Indeed, the inflation of the spatial component was consistently worse than other methods (Fig. S12). The problem of anomalous selection of eigenvectors was also exacerbated (Fig. S13): In scenarios where global significance was found, many eigenvectors were invariably selected. The exacerbation of the problems exhibited in other eigenvector techniques may have been due in part to the promulgation of spurious associations between individual species and individual eigenvectors, which quickly inflates the number of individual statistical tests. For example, a community with 20 species and a 50-axis MEM matrix, the total number of tests increases from 50 (with the Blanchet *et al.* method) to 1000. For our simulated communities, the number of tests run when we used the species-by-species forward-selection technique varied with the sampling regime and ranged from 255 to 1700.

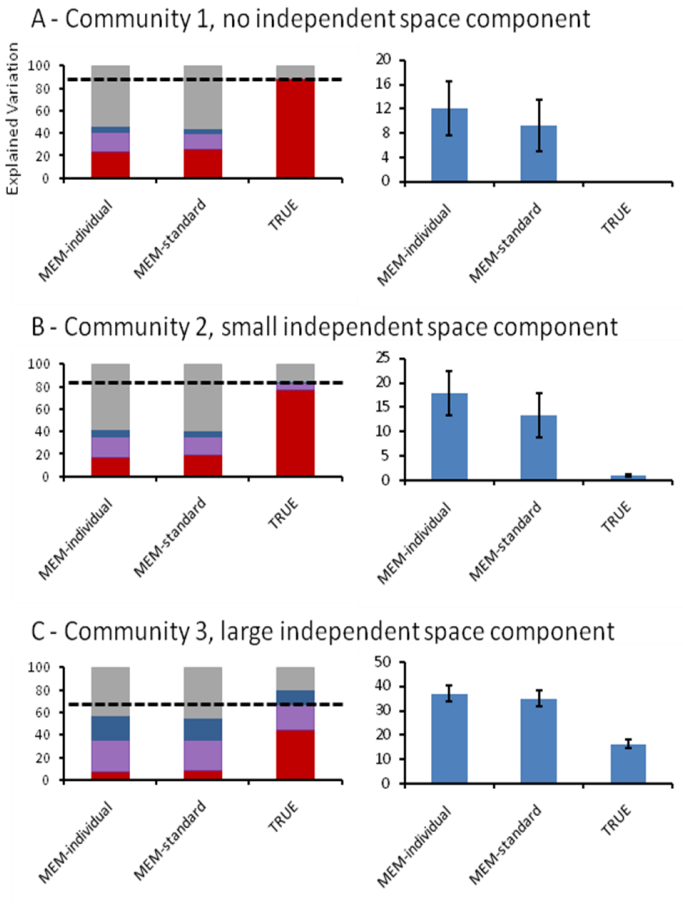


Fig. S12. Comparison across community types of MEM technique using species-by-species forward selection, as per Peres-Neto and Legendre (2010; MEM-individual), versus forward selection as per Blanchet, Legendre & Borcard (2008; MEM-standard). Sampling regime is 256 Contiguous; figure format as per Fig. 1 in the main article. The MEM-individual method appears to exacerbate the problems exhibited by the MEM-standard method, when compared with the true variation explained: The explained environmental variation (red) is lower, and both covariation (purple) and the spatial signal (blue) are higher than the true variation. The independent spatial variation explained as a proportion of explained variation (*S*/(*E*+*S*+*ES*); right-hand graphs), shows similar inflation of the spatial signal, which is slightly worse in the MEM-individual technique.

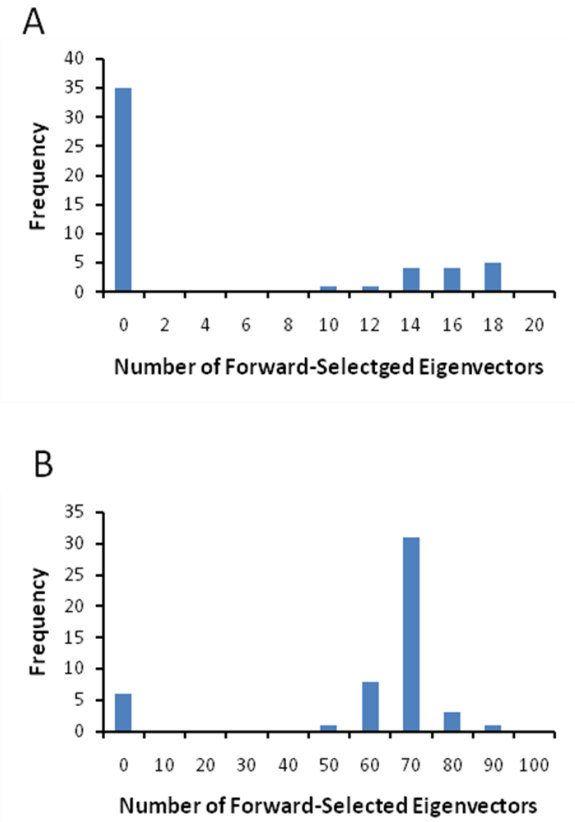


Fig. S13. Histograms showing examples of bimodal selection of MEM spatial eigenvectors using the species-by-species forward-selection technique proposed by Peres-Neto and Legendre (2010). A – Coarse random sample configuration, Community 2 (small spatial component); B – Fine random sample configuration, Community 2.

Literature Cited

Blanchet, F.G., Legendre, P & Borcard, D. (2008) Forward selection of explanatory variables. Ecology, **89**, 2623-2632.

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.

Peres-Neto, P.R. & Legendre, P. (2010) Estimating and controlling for spatial structure in the study of ecological communities. Global Ecology and Biogeography **19**, 174-184.