Statistical Reports

Ecology, 99(10), 2018, pp. 2159–2166

Optimizing the choice of a spatial weighting matrix in eigenvector-based methods

David Bauman,[1,4] Thomas Drouet,[1] Marie-Josée Fortin,[2] and Stéphane Dray[3]

[1]*Laboratoire d'Écologie Végétale et Biogéochimie, Université Libre de Bruxelles, CP 244, 50 av. F. D. Roosevelt, Brussels 1050 Belgium*
[2]*Department of Ecology & Evolutionary Biology, University of Toronto, 25 Willcocks, Toronto, Ontario M5S 3B2 Canada*
[3]*Laboratoire de Biométrie et Biologie Evolutive, CNRS, Université Lyon, Université Claude Bernard Lyon 1, Villeurbanne F-69100 France*

Abstract. Eigenvector-mapping methods such as Moran's eigenvector maps (MEM) are derived from a spatial weighting matrix (SWM) that describes the relations among a set of sampled sites. The specification of the SWM is a crucial step, but the SWM is generally chosen arbitrarily, regardless of the sampling design characteristics. Here, we compare the statistical performances of different types of SWMs (distance-based or graph-based) in contrasted realistic simulation scenarios. Then, we present an optimization method and evaluate its performances compared to the arbitrary choice of the most-widely used distance-based SWM. Results showed that the distance-based SWMs generally had lower power and accuracy than other specifications, and strongly underestimated spatial signals. The optimization method, using a correction procedure for multiple tests, had a correct type I error rate, and had higher power and accuracy than an arbitrary choice of the SWM. Nevertheless, the power decreased when too many SWMs were compared, resulting in a trade-off between the gain of accuracy and the loss of power. We advocate that future studies should optimize the choice of the SWM using a small set of appropriate candidates. R functions to implement the optimization are available in the *adespatial* package and are detailed in a tutorial.

Key words: community ecology; community simulation; connection scheme; inference of ecological processes from spatial patterns; Moran's eigenvector maps (MEM); multiscale spatial patterns; optimization; principal coordinates of neighbor matrices (PCNM); spatial autocorrelation; spatial eigenvector mapping (SEVM); spatial weighting matrix; type I error rate inflation.

Introduction

Spatial (but also temporal or phylogenetic) autocorrelation can be seen either as a curse or as an opportunity for ecologists (Peres-Neto and Legendre 2010, Diniz-Filho et al. 2012, Bauman et al. 2018). Indeed, the lack of independence between observations (spatial autocorrelation, SAC) causes standard statistical procedures to be too liberal (inflated type I error rate; Legendre 1993, Diniz-Filho and Bini 2005). Space, therefore, hinders the correct assessment of a relation between the response variable(s) and a set of predictors (*spatial nuisance*, Peres-Neto and Legendre 2010). Yet, space was also shown to be a surrogate of the effect of ecological processes on living communities (McIntire and Fajardo 2009, Legendre and Legendre 2012, *spatial legacy*, Peres-Neto and Legendre 2010). Hence, several spatially explicit methods have been developed either to filter SAC from residuals or to depict multiscale spatial patterns and relate them to underlying ecological processes (Griffith 1996, 2004, Plotkin et al. 2000, Borcard and Legendre 2002, Diniz-Filho and Bini 2005, Dray et al. 2006, 2012, Wagner and Dray 2015). The advent of spatial eigenvector-based methods has brought a major advance in this field (Griffith 1996), especially with the development of Moran's eigenvector maps (MEM, Dray et al. 2006) that generalizes the ad hoc principal coordinates of neighbor matrices (PCNM; Borcard and Legendre 2002). MEM allow including multiscale spatial predictors in all kinds of univariate and multivariate models (e.g., generalized linear models, canonical analyses).

MEM variables (also further referred to as spatial predictors or spatial eigenvectors) are generated by the diagonalization of a doubly centered spatial weighting matrix (SWM) **W**. The matrix **W** is obtained as the Hadamard product (element-wise product) of a connectivity matrix (**B**) by a weighting matrix (**A**). The binary matrix **B** defines the pairs of connected and unconnected sites (binary matrix), while the matrix **A** allows weighting the connections, for instance to define that the strength of the connection between two sites decreases with the geographic distance (Dray et al. 2006). The matrix **B** can be distance based, when the connection status (i.e., 1 or 0) of each pair of sites depends on the distance between them with respect to a connection threshold distance (e.g., Euclidean distances; *db*-MEM), as in the original PCNM method, but it can also be based on geometrical connection schemes, such as the Delaunay triangulation, Gabriel's graph, relative neighborhood graph, or a minimum spanning tree (graph-based

Manuscript received 27 February 2018; revised 13 June 2018; accepted 2 July 2018. Corresponding Editor: Karen C. Abbott.
[4] E-mail: davbauman@gmail.com or dbauman@ulb.ac.be

MEM, hereafter *gb*-MEM; Dray et al. 2006, Legendre and Legendre 2012). Connections can also be built on the basis of landscape features (physical barriers, resistance to movement; Taylor et al. 1993, Fortin and Payette 2002, Spear et al. 2010).

Several works have investigated how different specifications of SWMs influence the results of spatial analyses (e.g., Stetzer 1982, Florax and Rey 1995, Kostov 2010, Griffith 2017). The choice of the SWM has been shown to greatly influence the accuracy of parameter estimations and the spatial patterns detected in different types of space-time forecasting models, such as Lagrange Multiplier tests, in econometrics (Stakhovych and Bijmolt 2008), spatial autoregressive models (Griffith and Lagona 1998), and in spatial eigenvector-based methods too, especially for irregular sampling designs (Dray et al. 2006, Patuelli et al. 2011, Griffith 2017). However, a thorough evaluation (in terms of type I error rate, power, and accuracy) is still lacking to understand how spatial eigenvector-based methods are affected by the specification of the SWM with respect to the type of sampling design, the strength of the SAC, and the scale of the pattern. A recent review revealed that most studies used a single – and seemingly arbitrarily chosen – SWM (Bauman et al. 2018). Bauman et al. (2018) also highlighted that only 58% of the studies describe clearly the specification of the SWM, and that over 60% of these studies used either *db*-MEM or the original PCNM approach without justification, even if the latter lacks mathematical formalism, is very sensitive to irregular sampling designs, and present a lower statistical power than its MEM counterpart (Dray et al. 2006). Although Dray et al. (2006) proposed a procedure to select an optimal SWM among a set of candidates, this approach was based on the computation of an Akaike information criterion extended to the case of multivariate response data (Godinez-Dominguez and Freire 2003) which suffers from poor theoretical bases and does not test the candidate matrices against a null model. Hence, this procedure returns an optimal SWM even if there is no spatial structure in the response data, so that its use has been discouraged (Bauman et al. 2018).

Here, we address the issue of the selection of the SWM. We first compare contrasted types of SWMs in terms of type I error rate, statistical power, and R^2 estimation accuracy for: (1) random and clustered sampling designs, (2) weak and strong degrees of SAC, and (3) broad and fine spatial scale patterns. Then, we propose a new procedure that optimizes the selection of the SWM. Finally, we evaluate the performances of this new procedure and compare it to the most common current practices.

Material and Methods

Comparing the performance of spatial weighting matrices

We defined a 90 × 90 grid and sampled 120 cells (sites) following a clustered (three clusters of 40 sites) or a random sampling design (right portion of Fig. 1 for an illustration and Appendix S2: Section 1 for methodological details). For each sampling type, we built 21 contrasting SWMs as a combination of connectivity and weighting matrices (see below) and compared their performance (i.e., type I error rate, statistical power, and R^2 estimation accuracy).

*Connectivity matrix (**B**).*—The connectivity matrix (**B**) was generated using four connection schemes (graph-based MEM, hereafter *gb*-MEM: Delaunay triangulation, *del*, Gabriel graph, *gab*, relative neighborhood graph, *rel*, and minimum spanning tree, *mst*), and one distance threshold (*db*-MEM; see Appendix S1: Fig. S1). The latter corresponded to the smallest distance that kept all sites connected (i.e., the connectivity criterion used in PCNM; *db* in Appendix S1: Fig. S1). The graph-based connection schemes are inclusive, so that all the links of *mst* are included in *rel*, included in *gab*, itself included in *del*. Hence, the number of connections increases along these graphs (Legendre and Legendre 2012).

*Weighting matrix (**A**).*—Different weighting matrices (**A**) were combined to each **B** matrix. We defined (1) a neutral weighting function (f_{bin}; that is, no weight added to the connections), (2) a linear function $f_{lin} = 1-(d/d_{max})$, (3) a concave-down function $f_{down} = 1-(d/d_{max})^{\alpha}$, and (4) a concave-up function $f_{up} = 1/d_{max}{}^{\alpha}$, where d is the Euclidean distance between two sites, d_{max} is the maximum distance between two sites, and $\alpha = 5$ and 0.5 in f_{down} and f_{up}, respectively (see plot of the weighting functions in Appendix S1: Fig. S2). The weighting function $f_{PCNM} = 1-(d/4t)^2$ was used with the *db*-**B** matrix, where t is the threshold distance below which two sites are connected, and 4 is an empirical value beyond which the eigenvectors remain stable (Borcard and Legendre 2002). This combination of *db* and f_{PCNM} corresponds to the PCNM criteria used in the framework of MEM (*db*-MEM_{PCNM}; see Dray et al. 2006).

For each SWM, we only considered the MEM variables associated to positive eigenvalues (hereafter "positive MEM variables"), as most studies focus on contagious ecological processes (i.e., displaying positive SAC). Using MEM variables associated to negative or to all eigenvalues yielded very similar results (not shown).

Type I error rate

A random univariate response variable **y** was generated in the sampled cells from four distributions: uniform, normal, exponential, and cubed exponential (Anderson and Legendre 1999, Manly 2007). We then generated MEM variables using the 21 above-described types of SWM. A global test of significance of **y** against each SWM was performed separately by 999 permutations (i.e., regressing **y** against the entire set of positive MEM variables). The above-described simulation procedure was repeated 1,000 times by resampling different sets of 120 cells within the 90 × 90 grid, and the type I error rate was the proportion of significant results (significance level of 0.05).

Statistical power and R^2 *estimation accuracy*

The SWMs were then evaluated on the basis of their statistical power and R^2 estimation accuracy in a set of scenarios where the response variable **y** was spatially structured. We considered different sampling designs (clustered or random, see previous section), degrees of SAC (low or high), and spatial scales (broad or fine) in these simulations (Fig. 1; Appendix S2: Section 3 for details). Note that we

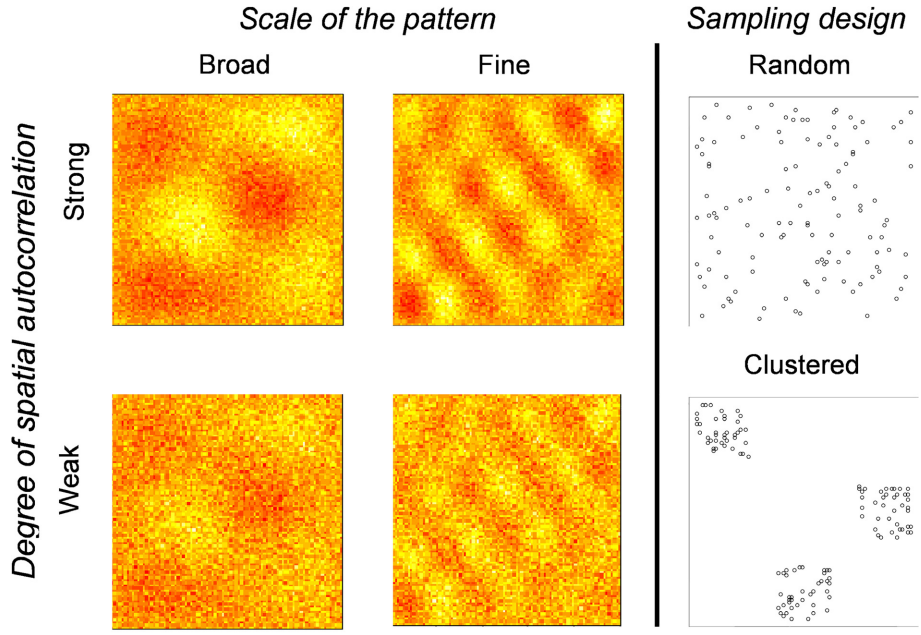

FIG. 1. Schematic definition of the simulation design used for evaluating power and accuracy. The response variables were generated with spatial patterns structured either at broad or fine spatial scales, with a strong or a weak degree of SAC on a grid of 8,100 cells (90 × 90). Then, 120 cells were sampled either randomly distributed, or following a clustered sampling design. The sampled values were then considered as the response variable (**y**) to assess the statistical performances of the different types of SWMs and of the optimization method (see Appendix S2: Section 3). The type I error rate evaluation used the same design but considered a random **y**.

did not consider regular sampling designs (e.g., equally spaced sites on a transect or grid), as spatial eigenvectors built with different SWMs detect roughly the same structures in this context (see *Discussion*). The response variable (**y**) was regressed on the global set of positive MEM variables generated from the same 21 SWMs, and a forward selection with double stopping criterion (Blanchet et al. 2008) was performed separately on the significant SWMs to select a suitable subset of spatial predictors. Then, **y** was regressed on the forward selected MEM variables of each significant SWM and the spatial contribution (R^2) to the overall variability of **y** was computed for the significant SWMs. The R^2 estimation accuracy of each SWM (hereafter ΔR^2) was defined as the difference between the true R^2 value (R^2_{ref}) and the R^2 estimated by the forward selected MEM variables of a given SWM (R^2_{sim}; that is, $\Delta R^2 = R^2_{\text{sim}} - R^2_{\text{ref}}$), so that negative and positive values indicated underestimation and overestimation of the true spatial signal, respectively. The complete simulation procedure was repeated 1,000 times by resampling different sets of 120 cells in the 90 × 90 grid, and the power was computed for each SWM as the proportion of simulations returning significant global R^2. The R^2 estimation accuracy was the mean ΔR^2 of the significant simulations. The complete simulation procedure is detailed in Appendix S2: Section 3.

Optimizing the selection of the spatial weighting matrix

Optimization method.—When a number of potential candidate SWMs is considered, it is expected that the probability to accidentally detect a spatial signal for a given response variable will increase with the number of candidates. As a consequence, we proposed a procedure to optimize the selection of a SWM while maintaining a correct type I error rate. After defining a set of potential SWMs, our method consists in (1) performing a global test (based on the R^2 of the model considering all MEM variables as explanatory variables) on each candidate matrix with a P-value correction for multiple tests (corrected by the number of SWMs compared), (2) running a forward selection with double stopping criterion (Blanchet et al. 2008) on the significant SWMs to define the best subset of eigenvectors for each one of them, and (3) selecting the optimal SWM as the one for which the best subset of eigenvectors yields the highest adjusted R^2. In this paper, the P-value is corrected by the Šidak correction (Šidák 1967), where $P_S = 1-(1-P)^k$, with P_S = the corrected P-value, P = the uncorrected P-value, and k = the number of tests (i.e., the number of SWMs), but other correction methods can be considered.

The optimization method has been implemented in R functions available in the *adespatial* package (Dray et al. 2018) (details and R tutorial in Appendix S3). These functions provide also alternate optimization procedures (e.g., minimizing residual SAC instead of maximizing adjusted R^2) that can be more suitable depending on the objective of the analysis (see details in Appendix S2: Section 4 and illustration in Appendix S3).

Performance of the optimization method.—The type I error rate, power, and accuracy of this optimization method were

calculated through 1000 repetitions using the same simulation design and scenarios as before. To assess the effect of the *P*-value correction, and because optimizing the selection of the SWM has so far been done without control of false discovery rate (Bauman et al. 2018), the type I error rate of the optimization method was computed with and without the *P*-value correction.

Five contrasting candidate SWMs were used in the optimization procedure: *gab* and *mst* (**B** matrices) associated with the f_{lin} and f_{down} functions (**A** matrices), and the *db*-MEM$_{\text{PCNM}}$. The power and accuracy of our optimization procedure were compared to those of the arbitrary choice of *db*-MEM$_{\text{PCNM}}$ (i.e., the most common current practice), and to those of the random selection of a SWM among a set of 57 SWMs (see details in Appendix S2: Section 5). This allowed assessing the benefits of optimizing the choice of the SWM with respect to a randomly chosen SWM or the common choice of *db*-MEM$_{\text{PCNM}}$.

All analyses were conducted in the R environment (version 3.4.3., R Core Team 2017) using the packages *vegan* (Oksanen et al. 2017), *spdep* (Bivand 2006), and *adespatial* (Dray et al. 2018). The R code of the study is provided in Data S1.

Results

Comparing the performance of spatial weighting matrices

The four random distributions yielded similar results. Hence, we only present the results of the uniform distribution (the other results are available in Appendix S4: Table S1).

Fig. 2a displays the type I error rate for each combination of **B** and **A** matrices, and shows that all the tested SWMs presented a correct type I error rate (between 0.04 and 0.06), regardless of the sampling design considered.

Fig. 2c, e show the R^2 estimation accuracy and statistical power of the SWMs tested with a strong degree of SAC, respectively. Regardless of the degree of SAC, the spatial scale, or the type of sampling design, the *gb*-MEM (*del*, *gab*, *rel*, and *mst*) systematically yielded a higher accuracy of R^2 estimation than the *db*-MEM (except for the strong degree of autocorrelation at broad scale, for the random sampling design). Among the *db*-MEM, the PCNM and binary weighting functions were nearly always associated to the strongest model underestimations. These underestimations were maximal when $\boldsymbol{y}$ displayed a fine-scale pattern. Overall, the *db*-MEM always performed poorly compared with at least one type of *gb*-MEM.

High degree of SAC.—With a clustered sampling design, the *gb*-MEM slightly underestimated the real R^2 (ΔR^2 down to -0.07 with mst_{bin}), while the *db*-MEM led to more severe underestimations (ΔR^2 down to -0.36 with *db*-MEM$_{\text{PCNM}}$; Fig. 2c). The results were very similar using the random sampling design, with a slight underestimation for the *gb*-MEM (ΔR^2 down to -0.09), except for the *del*-**B** matrix that led to strong underestimations when considering a fine-scaled pattern, regardless of the **A** matrix (mean ΔR^2 of -0.34). The *db*-MEM, again, yielded strong R^2 underestimations (ΔR^2 down to -0.37).

All the SWMs displayed high statistical power except for the *db*-MEM at fine scale and for the *del*-**B** matrix at fine scale for a random sampling design (i.e., for the above-mentioned cases of strong R^2 underestimation; Fig. 2e).

Low degree of SAC.—The results with a low degree of SAC were very similar, except for a general tendency toward a lower statistical power for all SWMs, and a more accurate R^2 estimation for both the *gb*-MEM and *db*-MEM. Yet, the *db*-MEM still underestimated the R^2, regardless of the spatial scale or sampling design considered (see details for the low degree of SAC in Appendix S1: Fig. S3a, c).

Optimizing the selection of the spatial weighting matrix

Fig. 2b shows the type I error rates of the optimizing method with and without *P*-value correction. Without correcting the *P*-value for multiple tests, optimizing the choice of the SWM among the five candidates tested inflated the type I error rate (0.18 for both sampling designs), while the method presented a correct type I error rate when correcting the *P*-value for the number of SWMs tested (0.01 and 0.02 for the clustered and random sampling designs, respectively). As expected, without *P*-value correction, the type I error rate inflation increased with the number of SWMs tested (results not shown).

In all simulation scenarios, the optimization method had a higher or equal power (Fig. 2f) and was more accurate (Fig. 2d) than the random choice of a SWM and the arbitrary choice of *db*-MEM (*db*-MEM$_{\text{PCNM}}$). Indeed, while the mean ΔR^2 of the optimization was always close to 0, the mean ΔR^2 of the random choice and the *db*-MEM went down to -0.33 and -0.37, respectively, hence causing severe underestimations of the spatial signal.

db-MEM performed the worst in most cases, mostly when the sampling design was clustered and for detecting fine-scaled patterns. The statistical power of this practice was particularly low for detecting fine-scaled patterns, and so was the power of the random choice of a SWM (although less markedly).

The benefit of the gain of precision of the optimization method was more marked for high degrees of SAC (see results for the low degree of SAC in Appendix S1: Fig. S3b, d), and more specifically for the fine-scaled patterns, both for clustered and random sampling designs (Fig. 2d). Moreover, at fine scales, the power of the optimization method was ~1, while the powers of the random choice and the *db*-MEM both went down to ~0.5 (Fig. 2f). Increasing the number of candidate SWMs in the optimization procedure enhanced the R^2 estimation accuracy but also decreased the statistical power, as the corrected significance threshold became more severe (i.e., smaller; results not shown).

Discussion

Properly defining the SWM to be used in spatially explicit analyses of ecological data is important to avoid biases, accurately capture and study the multiscale distribution patterns of living organisms. To do so, it is crucial to evaluate the practices related to the most crucial step of these methods, that is, the selection of a SWM. Bauman et al. (2018)

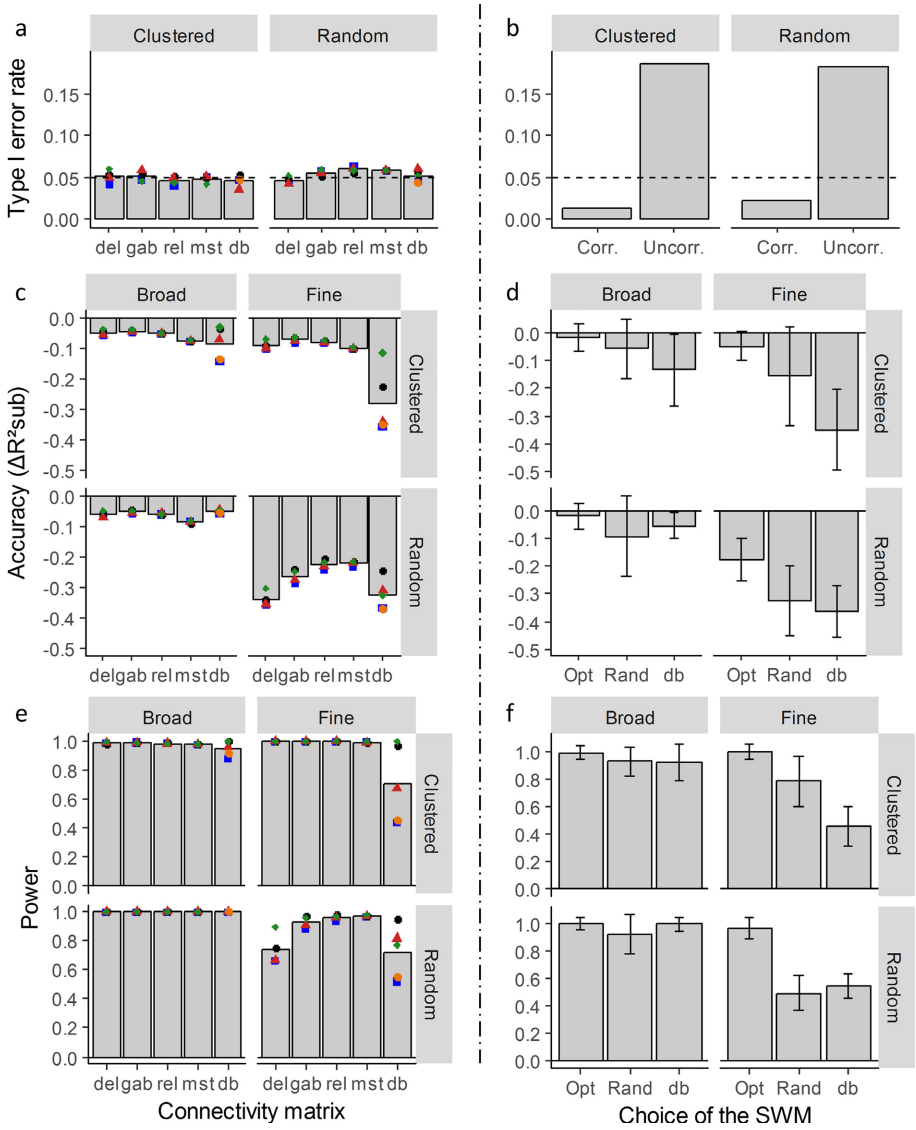

FIG. 2. Type I error rate, R^2 estimation accuracy, and statistical power of the different SWMs (a, c, e), the optimization method used with a forward selection criterion, the random choice of a SWM, and the arbitrary choice of *db*-MEM$_{PCNM}$ (b, d, f), at broad and fine spatial scales and for different types of sampling design ("Clustered" and "Random"). These are the results for the high degree of SAC (see results with a low degree of SAC in Appendix S1: Fig. S3). a, c, e: The gray vertical bars correspond to the mean of the type I error rate (a), mean ΔR^2 (i.e., $R^2_{sim}-R^2_{ref}$) (c), and statistical power (e) of the different **A** matrices within each **B** matrix (*x*-axis). The symbols give the detailed values for the combinations of the matrices **B** and **A**. Squares: f_{bin}, black circles: f_{lin}, triangles: f_{down}, diamonds: f_{up}, orange circles: f_{PCNM}. b: Type I error rate of the procedure with ("Corr.") and without ("Uncorr.") the Sidak correction of the global *P*-value for multiple tests. d, f: R^2 estimation accuracy (d) and statistical power (f) of the optimization procedure with *P*-value correction ("Opt"), the random choice of a SWM among 57 candidates ("Rand"), and the *db*-MEM$_{PCNM}$ ("db"). a, b: The dashed line is the correct type I error rate (i.e., 0.05). c, d: Negative and positive values of ΔR^2 correspond to underestimations and overestimations of the actual R^2 (i.e., R^2_{ref}), respectively.

showed that few studies considered this issue and that around half of the published works did not describe precisely how they defined the SWM in their study. It was also highlighted that an arbitrary choice of *db*-MEM – often the original PCNM method – was made in the great majority of the studies that specified their SWM. The PCNM method has, however, long been shown to lack mathematical formalism, to generate less spatial predictors (therefore displaying a lower statistical power), and to be particularly sensitive to irregular sampling designs (Dray et al. 2006). It is, therefore, fundamental to define good practices to wisely choose the SWM for spatial data analysis based on eigenvector-based methods.

Our results showed that the *gb*-MEM always displayed a higher accuracy in R^2 estimation and nearly always had a higher statistical power than the *db*-MEM (the power and accuracy of *db*-MEM were particularly low when weighted by the neutral or PCNM functions). This result can easily be understood, as distance-based connectivity criterions connect all sites aside from a distance corresponding either to the minimum distance necessary to keep all sites connected (i.e., the largest edge of a minimum spanning tree) or

to any threshold distance greater than that. With irregular and clustered sampling designs, this threshold distance is likely to connect too many sites, therefore, avoiding a proper detection of fine to medium-scaled spatial patterns. This will cause distant sites, potentially belonging to different clusters of sites (from a biological or ecological standpoint), to be artificially connected, hence leading to misspecifications of the SWM and poor detection of spatial patterns. The *db*-MEM definition is therefore unsuitable for a proper detection of spatial patterns in a set of irregularly distributed sites. Unlike *db*-MEM, *gb*-MEM does not present the constraint of building connectivity with respect to a distance threshold potentially defined by a single pair of distant sites. Therefore, this family of MEM yields more realistic connections, regardless of the regularity/clustering of the sampled sites, and provided higher R^2 estimation accuracy and statistical power in our simulation study.

The connections based on the Delaunay triangulation, however, performed poorly in different scenarios. This was most likely caused by long-distance connections known to be generated at the edges of the sampling design by the Delaunay criterion (Kenkel et al. 1989). These edge effects artificially connect distant sites, hence misspecifying the SWM and causing the observed underestimations. The Delaunay triangulation should, therefore, be avoided, unless used with an edge effect correction (e.g., Lane et al. 1994). A solution could be to use minimum planar graphs (MPG) (Fall et al. 2007), a generalization of Delaunay triangulation that accounts for the resistance to connect sites, hence providing least-cost paths and avoiding excessively long links as those obtained with *del* in this study.

The optimization procedure that we proposed to select the SWM achieved a higher R^2 estimation accuracy than the random choice of a SWM, hence highlighting the importance of the SWM selection. However, the optimization had a high false discovery rate when not associated to a *P*-value correction for multiple tests, which confirmed our expectation. Previous studies following Dray et al. (2006) and Borcard et al. (2011) to select a SWM (based on the AIC), therefore, not only probably had an inflated type I error rate when selecting spatial eigenvectors (see Bauman et al. 2018) but likely also for the selection of the best SWM. The latter inflation was caused by the lack of an adapted control for the potentially very high number of candidate SWMs tested (up to ~100 in Borcard et al. 2011). This type I error rate inflation issue also occurred in additional simulations with varying parameters values (e.g., different values used as α exponent in the concave-down or concave-up functions, or different threshold distances in *db*-MEM, Borcard et al. 2011). This issue can be solved by correcting the global *P*-value according to the number of SWMs tested before selecting a subset of predictors (see details of these additional simulations in Appendix S2: Section 2).

The *P*-value correction can quickly become very severe, however, as the number of tests not only increases with the **B** matrices compared but also with the number of connectivity distance thresholds of *db*-MEM or the number of parameters within each **A** matrix, the latter being generally used to weight each of the tested **B** matrices. In our simulations, the cost of the optimization procedure was nearly inexistent in most cases. However, we only performed the optimization on the basis of five SWMs. Comparing a higher number of candidates rapidly lowered the statistical power of the procedure (results not shown). A trade-off is thus necessary between the benefit and the cost of the optimization, that is, the gain of accuracy against the loss of statistical power.

In this study, the optimization was performed using a criterion associated to the fit of spatial predictors to a response variable (adjusted R^2). This procedure is relevant for any framework focusing on capturing all the spatial patterns of **y** (e.g., variation partitioning, Borcard et al. 1992, Peres-Neto and Legendre 2010). However, the questions regarding the selection of the best SWM and an optimal subset of spatial predictors are also relevant when the objective is to remove residual SAC in a model considering additional explanatory variables (e.g., environmental; see spatial eigenvector mapping, Diniz-Filho and Bini 2005). In this case, the most adapted eigenvector-selection method should aim to minimize the residual SAC with a small number of spatial predictors (MIR method in Bauman et al. 2018) (Griffith and Peres-Neto 2006, Bauman et al. 2018), and we recommend to perform the optimization of the SWM with the same criterion (see details in Appendix S2: Section 4 and illustration in Appendix S3). In both previous cases, the selection of the SWM is based on the selection of a subset of spatial eigenvectors (maximizing the fit or minimizing the residual SAC), as subsequent analyses will consider only this subset of predictors. However, some other methods require the complete set of spatial eigenvectors (e.g., Moran spectral randomizations, Wagner and Dray 2015, or smoothed MEM, Munoz 2009). In this case, the optimization of the SWM should be performed on the basis of the whole set of MEM variables without considering any procedure of selection of a subset of eigenvectors (details in Appendix S2: Section 4 and Appendix S3).

It is worth mentioning that optimizing the choice of the SWM when the sampling design is roughly regular is less interesting, as the MEM variables originating from different SWMs detect roughly the same patterns (Dray et al. 2006). In those cases, rook (shared edge) or queen (shared edge or vertex) neighbor definitions or *db*-MEM can be used, for instance (see Appendix S3). For irregular sampling designs, visualizing the different connection schemes would help identifying **B** matrices worth being tested and compared. In addition, considering the landscape ecology (e.g., natural barriers) should help improving the definition of the connectivity among sites in matrix **B**. Regarding the **A** matrix, plotting connectivity against distance and visualizing the curve of the different functions with several values of parameter (Appendix S1: Fig. S2) should also help choosing appropriate weighting functions and parameter values. Note that our study is not exhaustive and other functions (e.g., negative exponential functions) and connectivity schemes (e.g., *k* nearest neighbors) may be relevant. Visualizing the **B** and **A** matrices bearing in mind the above-mentioned trade-off between power and accuracy should be a fundamental step to reduce the number of candidates and improve the performance of the optimization.

Finally, it has been shown that explicitly integrating the resistance to movement/dispersal in spatial weighting (or

connectivity) matrices allowed obtaining much more precise results than simple distance-based criteria (e.g., Rayfield et al. 2010, Hanks and Hooten 2013, Saura et al. 2014, Ver Hoef et al. 2018). Incorporating the cumulative effects of landscape fragmentation and land use change into connectivity matrices makes the distance between locations more ecologically realistic and the integration of landscape connectivity in spatial eigenvector methods is an exciting challenge.

Acknowledgment

This research was supported by the Belgian National Fund for Scientific Research (F.R.S.-FNRS) to DB, and was performed using the Shared ICT Services Centre, Université Libre de Bruxelles, and the computing facilities of the CC LBBE/PRABI. We are thankful to two anonymous reviewers and to Dr. Karen Abbott for their constructive suggestions.

Author Contributions

DB conceived the ideas of the study; DB and SD designed methodology; DB analyzed the data, led the writing, and designed the R functions; SD supervised the integration of the functions to the package *adespatial*; MJF contributed to the discussion; TD made constructive revisions of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

Literature Cited

Anderson, M. J., and P. Legendre. 1999. An empirical comparison of permutation methods for tests of partial regression coefficients in a linear model. Journal of Statistical Computation and Simulation 62:271–303.

Bauman, D., T. Drouet, S. Dray, and J. Vleminckx. 2018. Disentangling good from bad practices in the selection of spatial or phylogenetic eigenvectors. Ecography 41:1–12.

Bivand, R. 2006. spdep: spatial dependence: weighting schemes, statistics and models. R package (version 0.6-13).

Blanchet, F. G., P. Legendre, and D. Borcard. 2008. Forward selection of explanatory variables. Ecology 89:2623–2632.

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

Borcard, D., P. Legendre, and P. Drapeau. 1992. Partialling out the spatial component of ecological variation. Ecology 73:1045–1055.

Borcard, D., F. Gillet, and P. Legendre. 2011. Numerical ecology with R. Springer New-York, New York, New York, USA.

Diniz-Filho, J. A. F., and L. M. Bini. 2005. Modelling geographical patterns in species richness using eigenvector-based spatial filters. Global Ecology and Biogeography 14:177–185.

Diniz-Filho, J. A. F., L. M. Bini, T. F. Rangel, I. Morales-Castilla, M. Á. Olalla-Tárraga, M. Á. Rodríguez, and B. A. Hawkins. 2012. On the selection of phylogenetic eigenvectors for ecological analyses. Ecography 35:239–249.

Dray, S., P. Legendre, and P. R. Peres-Neto. 2006. Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbour matrices (PCNM). Ecological Modelling 196:483–493.

Dray, S., et al. 2012. Community ecology in the age of multivariate multiscale spatial analysis. Ecological Monographs 82:257–275.

Dray, S., et al. 2018. adespatial: Multivariate multiscale spatial analysis. R package version 0.2-0.

Fall, A., M. J. Fortin, M. Manseau, and D. O'Brien. 2007. Spatial graphs: principles and applications for habitat connectivity. Ecosystems 10:448–461.

Florax, R. J., and S. Rey. 1995. The impacts of misspecified spatial interaction in linear regression models. Pages 111–135 *in* L. Anselin, and R. J. G. M. Florax, editors. New directions in spatial econometrics. Springer, Berlin, Heidelberg.

Fortin, M. J., and S. Payette. 2002. How to test the significance of the relation between spatially autocorrelated data at the landscape scale: a case study using fire and forest maps. Ecoscience 9:213–218.

Godinez-Dominguez, E., and J. Freire. 2003. Information-theoretic approach for selection of spatial and temporal models of community organization. Marine Ecology Progress Series 253:17–24.

Griffith, D. A. 1996. Spatial autocorrelation and eigenfunctions of the geographic weights matrix accompanying geo-referenced data. Canadian Geographer 40:351–367.

Griffith, D. A. 2004. A spatial filtering specification for the autologistic model. Environment and Planning A 36:1791–1812.

Griffith, D. A. 2017. Some robustness assessments of Moran eigenvector spatial filtering. Spatial Statistics 22:155–179.

Griffith, D. A., and F. Lagona. 1998. On the quality of likelihood-based estimators in spatial autoregressive models when the data dependence structure is misspecified. Journal of Statistical Planning and Inference 69:153–174.

Griffith, D. A., and P. R. Peres-Neto. 2006. Spatial modeling in Ecology: the flexibility of eigenfunction spatial analyses. Ecology 87:2603–2613.

Hanks, E. M., and M. B. Hooten. 2013. Circuit theory and model-based inference for landscape connectivity. Journal of the American Statistical Association 108:22–33.

Kenkel, N. C., J. A. Hoskins, and W. D. Hoskins. 1989. Edge effects in the use of area polygons to study competition. Ecology 70:272–274.

Kostov, P. 2010. Model boosting for spatial weighting matrix selection in spatial lag models. Environment and Planning B: Planning and Design 37:533–549.

Lane, S. N., K. S. Richards, and J. H. Chandler. 1994. Developments in monitoring and modelling small-scale river bed topography. Earth Surface Processes and Landforms 19:349–368.

Legendre, P. 1993. Spatial autocorrelation: Trouble or new paradigm? Ecology 74:1659–1673.

Legendre, P., and L. Legendre. 2012. Numerical ecology. Elsevier, Amsterdam, Netherlands.

Manly, B. F. J. 2007. Randomization, bootstrap and Monte Carlo methods in biology. Chapman & Hall/CRC, London, UK.

McIntire, E. J. B., and A. Fajardo. 2009. Beyond description: the active and effective way to infer processes from spatial patterns. Ecology 90:46–56.

Munoz, F. 2009. Distance-based eigenvector maps (DBEM) to analyse metapopulation structure with irregular sampling. Ecological Modelling 220:2683–2689.

Oksanen, J., et al. 2017. Package 'vegan': Community ecology package (version 2.4-3).

Patuelli, R., D. A. Griffith, M. Tiefelsdorf, and P. Nijkamp. 2011. The use of spatial filtering techniques: the spatial and space-time structure of German unemployment data. International Regional Science Review 34:253–280.

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

Plotkin, J. B., M. D. Potts, N. Leslie, N. Manokaran, J. LaFrankie, and P. S. Ashton. 2000. Species-area curves, spatial aggregation, and habitat specialization in tropical forests. Journal of Theoretical Biology 207:81–99.

R Core Team. 2017. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Rayfield, B., M. J. Fortin, and A. Fall. 2010. The sensitivity of least-cost habitat graphs to relative cost surface values. Landscape Ecology 25:519–532.

Saura, S., Ö. Bodin, and M. J. Fortin. 2014. EDITOR'S CHOICE: Stepping stones are crucial for species' long-distance dispersal and range expansion through habitat networks. Journal of Applied Ecology 51:171–182.

Šidák, Z. 1967. Rectangular confidence regions for the means of multivariate normal distributions. Journal of the American Statistical Association 62:626–633.

Spear, S. F., N. Balkenhol, M. J. Fortin, B. H. McRae, and K. Scribner. 2010. Use of resistance surfaces for landscape genetic studies: considerations for parameterization and analysis. Molecular Ecology 19:3576–3591.

Stakhovych, S., and T. H. A. Bijmolt. 2008. Specification of spatial models: a simulation study on weights matrices. Papers in Regional Science 88:389–408.

Stetzer, F. 1982. Specifying weights in spatial forecasting models: the results of some experiments. Environment and Planning A 14:571–584.

Taylor, P. D., L. Fahrig, K. Henein, and G. Merriam. 1993. Connectivity is a vital element of landscape structure. Oikos 68:571–573.

Ver Hoef, J. M., E. E. Peterson, M. B. Hooten, E. M. Hanks, and M.-J. Fortin. 2018. Spatial autoregressive models for statistical inference from ecological data. Ecological Monographs 88: 36–59.

Wagner, H. H., and S. Dray. 2015. Generating spatially constrained null models for irregularly spaced data using Moran spectral randomization methods. Methods in Ecology and Evolution 6:1169–1178.

Supporting Information

Additional supporting information may be found in the online version of this article at http://onlinelibrary.wiley.com/doi/10.1002/ecy.2469/suppinfo

Statistical Reports