## Important declarations

Please remove this info from manuscript text if it is also present there.

### Associated Data

**Data supplied by the author:**
The R script code and files needed to perform all simulation frameworks are available as Supplemental Information. All results are also available as Supplemental Information.

### Required Statements

**Competing Interest statement:**
The authors declare that they have no competing interests.

**Funding statement:**
This work was supported by National Council for Technological and Scientific Development, (CNPq) within Science without Borders scholarship scheme (Lélis Carlos-Júnior) and CNPq-305330/2010-1 (Joel Creed). Additional funding was provided by Brazilian Coordination for the Improvement of Higher Education Personnel (CAPES); CAPES- Ciências do Mar (Joel Creed 1137/2010); Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (Joel Creed., FAPERJ-E-26/111.574/2014 and E26/201.286/2014). Finally, Rob Lewis received funding from EU's Marie Skłodowska-Curie action (Grant 703258) and Rafael Feijó-Lima received funding from CAPES program Science Without Borders – project n. 166/2012.

# Generalized Linear Models outperform commonly used canonical analysis in estimating spatial structure of presence/absence data

**Lélis A Carlos-Júnior** Corresp., [1, 2, 3], **Joel C Creed** [4], **Rob Marrs** [2], **Rob J Lewis** [5], **Timothy P Moulton** [4], **Rafael Feijó-Lima** [1, 6], **Matthew Spencer** [2]

[1] Programa de Pós-Graduação em Ecologia e Evolução, Universidade do Estado do Rio do Janeiro, Rio de Janeiro, Brazil

[2] School of Environmental Sciences, University of Liverpool, Liverpool, United Kingdom

[3] Departamento de Biologia, Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, Brazil

[4] Departamento de Ecologia, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, Brazil

[5] Department of Bioscience, Norwegian Institute of Bioeconomy Research, Bergen, Norway

[6] Division of Biological Sciences, University of Montana, Missoula, Montana, United States

Corresponding Author: Lélis A Carlos-Júnior
Email address: lelisjr_cjr@puc-rio.br

**Background.** Ecological communities tend to be spatially structured due to environmental gradients and/or spatially contagious processes such as growth, dispersion and species interactions. Data transformation followed by usage of algorithms such as Redundancy Analysis (RDA) is a fairly common approach in studies searching for spatial structure in ecological communities, despite recent suggestions advocating the use of Generalized Linear Models (GLMs). Here, we compared the performance of GLMs and RDA in describing spatial structure in ecological community composition data. We simulated realistic presence/absence data typical of many β-diversity studies. For model selection we used standard methods commonly used in most studies involving RDA and GLMs.

**Methods.** We simulated communities with known spatial structure, based on three real spatial community presence/absence datasets (one terrestrial, one marine and one freshwater). We used spatial eigenvectors as explanatory variables. We varied the number of non-zero coefficients of the spatial variables, and the spatial scales with which these coefficients were associated and then compared the performance of GLMs and RDA frameworks to correctly retrieve the spatial patterns contained in the simulated communities. We used two different methods for model selection, Forward Selection (FW) for RDA and the Akaike Information Criterion (AIC) for GLMs. The performance of each method was assessed by scoring overall accuracy as the proportion of variables whose inclusion/exclusion status was correct, and by distinguishing which kind of error was observed for each method. We also assessed whether errors in variable selection could affect the interpretation of spatial structure.

**Results.** Overall GLM with AIC-based model selection (GLM/AIC) performed better than RDA/FW in selecting spatial explanatory variables, although under some simulations the methods performed similarly. In general, RDA/FW performed unpredictably, often retaining too many explanatory variables and selecting variables associated with incorrect spatial scales. The spatial scale of the pattern had a negligible effect on GLM/AIC performance but consistently affected RDA's error rates under almost all scenarios.

**Conclusion.** We encourage the use of GLM/AIC for studies searching for spatial drivers of species presence/absence patterns, since this framework outperformed RDA in situations most likely to be found

in natural communities. It is likely that such recommendations might extend to other types of explanatory variables.

# Generalized Linear Models outperform commonly used canonical analysis in estimating spatial structure of presence/absence data

Lélis A Carlos-Júnior[1,2,3], Joel C Creed[4], Rob Marrs[1], Rob J Lewis[5], Timothy P Moulton[4], Rafael Feijó-Lima[2,6], Matthew Spencer[1]

[1] School of Environmental Sciences, University of Liverpool. Liverpool, UK.
[2] Programa de Pós-Graduação em Ecologia e Evolução, Universidade do Estado do Rio de Janeiro. Rio de Janeiro, Brazil.
[3] Departamento de Biologia, Pontifícia Universidade Católica do Rio de Janeiro. Rio de Janeiro, Brazil.
[4] Departamento de Ecologia, Universidade do Estado do Rio de Janeiro. Rio de Janeiro, Brazil.
[5] Department of Bioscience, Norwegian Institute of Bioeconomy Research. Bergen, Norway.
[6] Division of Biological Sciences, University of Montana. Missoula, US.

Corresponding Author:
Lélis Carlos-Júnior[1,2,3]
Rua São Francisco Xavier, 524 – Maracanã, Rio de Janeiro, CEP: 20550-013, Brazil.
Email address: lelis_cjr@puc-rio.br / lelisufmg@gmail.com

## Abstract

**Background.** Ecological communities tend to be spatially structured due to environmental gradients and/or spatially contagious processes such as growth, dispersion and species interactions. Data transformation followed by usage of algorithms such as Redundancy Analysis (RDA) is a fairly common approach in studies searching for spatial structure in ecological communities, despite recent suggestions advocating the use of Generalized Linear Models (GLMs). Here, we compared the performance of GLMs and RDA in describing spatial structure in ecological community composition data. We simulated realistic presence/absence data typical of many β-diversity studies. For model selection we used standard methods commonly used in most studies involving RDA and GLMs.

**Methods.** We simulated communities with known spatial structure, based on three real spatial community presence/absence datasets (one terrestrial, one marine and one freshwater). We used spatial eigenvectors as explanatory variables. We varied the number of non-zero coefficients of the spatial variables, and the spatial scales with which these coefficients were associated and then compared the performance of GLMs and RDA frameworks to correctly retrieve the spatial patterns contained in the simulated communities. We used two different methods for model selection, Forward Selection (FW) for RDA and the Akaike Information

Criterion (AIC) for GLMs. The performance of each method was assessed by scoring overall accuracy as the proportion of variables whose inclusion/exclusion status was correct, and by distinguishing which kind of error was observed for each method. We also assessed whether errors in variable selection could affect the interpretation of spatial structure.

**Results.** Overall GLM with AIC-based model selection (GLM/AIC) performed better than RDA/FW in selecting spatial explanatory variables, although under some simulations the methods  performed similarly. In general, RDA/FW performed unpredictably, often retaining too many explanatory variables and selecting variables associated with incorrect spatial scales. The spatial scale of the pattern had a negligible effect on GLM/AIC performance but consistently affected RDA's error rates under almost all scenarios.

**Conclusion.** We encourage the use of GLM/AIC for studies searching for spatial drivers of species presence/absence patterns, since this framework outperformed RDA in situations most likely to be found in natural communities. It is likely that such recommendations might extend to other types of explanatory variables.

## Introduction

Ecological communities tend to be spatially structured in response to environmental gradients that are themselves organized in space, or to spatially contagious processes such as growth, dispersion, and species interactions (Legendre & Legendre, 2012; Peres-Neto & Legendre, 2010). Thus, disentangling the causes of spatial structure and identifying spatial variability and different scales of organization in natural communities is a central question in ecology (Legendre, 1993). Answering this question requires the construction of explanatory variables based on spatial relationships among sites (Dray et al., 2006). One approach extensively used to create spatial variables and/or control for spatial autocorrelation in residuals is an eigenvector-based method, called Moran's eigenvector maps  (MEMs, Dray et al., 2006). This method creates spatial explanatory variables representing structure on a range of spatial scales from the spatial relationships among sampling sites. These variables can be used for a broad range of goals, from controlling for phylogenetic autocorrelation in ecological data (Diniz-Filho et al., 2012) to searching for spatial structure in natural communities, even when irregularly sampled (*e.g.* Bauman et al., 2016; Neves et al., 2015).

In many studies the response variables for which ecologists seek to find spatial structure are community composition datasets containing either abundances or presence/absence information (here, we focus on the latter). For community ecology studies, Redundancy Analysis (RDA) is one of the most popular strategies due to its versatile framework, well-established literature and abundant toolkits available for implementation (see Blanchet, Legendre, Bergeron, & He, 2014; Borcard, Legendre, & Drapeau, 1992; Eisenlohr & Oliveira-Filho, 2015; Saiter, Eisenlohr, Barbosa, Thomas, & Oliveira-Filho, 2015). The RDA algorithm searches for optimal linear combinations (in the least-squares sense, see Legendre & Legendre, 2012) of the explanatory variables that best explain the variation in the transformed community composition data (Legendre & Gallagher, 2001; Borcard et al., 2011; Blanchet et al., 2014). The usual approach then consists of establishing the global significance of the relationship between the response matrix and all the explanatory variables, after which a subset of explanatory variables is usually selected by stepwise procedure such as Forward Selection (FW, *sensu* Blanchet et al., 2008). The most common approach uses two thresholds for variable selection: a significance level α and the adjusted R² (see below and Blanchet et al., 2008 for details). This whole framework will hereafter be called RDA/FW for brevity. A statistic related to the Akaike Information Criteria (AIC, Akaike, 1973) has also been suggested for RDA model selection (Godínez-Domínguez & Freire, 2003), but it has been shown to perform poorly and will not be further explored here (Bauman, Drouet, Dray, et al., 2018).

However, methods based on least-squares such as RDA are unlikely to perform well when applied to data that violate the assumption of constancy in the mean-variance relationship. This assumption is usually violated by datasets containing many zeros including abundance (count or semi-quantitative) and presence/absence (binary) data. Data transformation does not always solve such problems (O'Hara & Kotze, 2010; Warton, 2018), although least-squares can give reasonably robust tests of the significance of regression coefficients (Ives, 2015). In general, algorithmic methods such as RDA do not take into account the statistical properties of the response variable, such as the distribution of variances and how the response changes along spatial/environmental gradients (Ferrier et al., 2007; Warton et al., 2012, 2015, 2018). More recently, Generalized Linear Models (GLMs) have been proposed as an alternative model-based approach to the analysis of presence/absence or count data (Wang et al., 2012; Warton et al., 2015; Yee, 2006). The use of GLMs has long been established for univariate analyses and related approaches for  multivariate count data are now available (O'Hara & Kotze, 2010; Warton, 2018). The usual approach to selection of explanatory variables in this approach is Akaike's Information Criterion (AIC: Akaike, 1973; Wagenmakers & Farrell, 2004). This framework will hereafter be named GLM/AIC.

Here, we compared the performance of the RDA/FW and GLM/AIC approaches to selecting spatial explanatory variables for community presence/absence data by measuring the proportion of spatial patterns contained in simulated communities they could correctly retrieve. There have been some studies of simulated multivariate count data (Warton et al., 2012), but presence/absence data are particularly important in spatial studies because they are often the only data that can be collected consistently over large spatial extents. We therefore compare the performance of RDA/FW and GLM/AIC methods for the selection of MEM spatial variables (including one special case, the asymmetric eigenvector map or AEM) from realistic simulated presence/absence data. We used spatial variables as our predictors since we were interested in discovering whether varying the spatial scales in which communities were structured would affect model performance. We generated simulated data sets with predefined spatial structure based on three real data sets, under two different ecological interpretations of presence/absence data. First, we assumed that species are truly present at some sites and absent at others, and are detected if present (simulated presence method, SPM). Alternatively, absences may represent failure to detect species that are truly present. In this case, we simulated species abundances, followed by a simulated sampling step to obtain presence/absence data (simulated abundances method, SAM).

## Materials & Methods

### *Baseline Datasets*

We compared the two approaches to spatial variable selection using simulated community data based on three real community composition datasets with a range of properties:

A) Presence/Absence of 110 marine benthic macroalgae species from a Rapid Assessment Program for biodiversity of 42 sample sites spanning roughly 2000 km$^2$ at Ilha Grande Bay, Rio de Janeiro, Brazil (tropical southwest Atlantic) (Carlos-Júnior et al., 2019, permit number IBAMA/RJ:031/04);

B) Presence/Absence of 588 plant species from grassland covering 500 km$^2$ of Scotland's coast. Data were collected from 3639 5 × 5 m quadrats from 94 sites. We used sites as our sample units, treating species as present when they occurred in at least one quadrat at a site, and absent otherwise (see Lewis et al., 2014 for more information);

C) Presence/Absence of 47 freshwater aquatic insect species collected from 30 sample sites in five tributaries of the Guapiaçú River basin, Brazil which covers about 40 km² (Feijó-Lima in prep, permit number INEA-RJ: 019-2014).

For each of the datasets we used the geographical coordinates (maps and sampling sites in Supplemental Figure S1) to calculate spatial explanatory variables for regression (Fig. 1). We chose MEMs as our spatial variables since they are commonly used to describe spatial structure in ecological studies. Moreover, in contrast to coarser methods such as trend-surface analysis, MEMs are a flexible method, capable of describing all spatial scales provided by the sampling design (Borcard et al., 2011). They are also more flexible and powerful than the method of principal coordinates of neighbor matrices (PCNMs, a special case of distance-based MEMs) (Bauman, Drouet, Dray, et al., 2018; Bauman, Drouet, Fortin, et al., 2018; Borcard & Legendre, 2002; Dray et al., 2006). One needs two matrices to build the MEM variables for a given set of site coordinates: matrix **B** describing the connectivity among the geographical sampling sites and matrix **A** describing the weights of such connections. The Hadamard product of these two matrices generates the spatial weighting matrix (matrix **W**), which is then doubly centred and diagonalized, yielding eigenvectors to be used as spatial variables. For ecological studies, the processes of interest are usually those generating positive autocorrelation, and it is therefore common to use only MEMs associated with positive eigenvalues (as in this study). For studies in which negative spatial autocorrelation is also of interest (*e.g.* where negative interactions such as competitive exclusion, predation, etc are suspected), the eigenvectors associated with negative eigenvalues can also be separately used (Bauman, Drouet, Dray, et al., 2018). We made decisions about **B** and **A** for each dataset based on our ecological knowledge of the spatial structure of these regions, since our goal was to simulate communities with ecologically sensible spatial structures. Therefore, for dataset A we chose the minimum spanning tree (**B**) with Euclidian linear distances as weights (**A**). Our decision was based on the shape of the bay and the fact that the main water movements make the sampling sites geographically compartmentalised in subregions where sites are likely to be minimally connected (Carlos-Júnior et al., 2019). Similarly, spatial organisation in dataset B could be sensibly described in terms of Delaunay triangulation (**B**) with Euclidian weights (**A**). Despite some degree of connectivity among all sites, pairs of sites could be mostly associated not to their immediate neighbours but rather as a function of their distances. This is due to cultural differences in land management. For example, northern and western islands share cultural histories, which is reflected in species composition (Lewis et al., 2014). Directional spatial processes in ecological data, such as those observed in rivers, are well described by a special case of MEMs called asymmetric eigenvector maps (AEMs, Blanchet, Legendre, & Borcard, 2008), which were used for constructing variables for dataset C. In MEMs, larger eigenvalues are associated with broader-scale spatial structures while smaller eigenvalues represent fine-scale spatial structures. This allowed us to control the spatial scale of variation in community structure. Dataset A had 16 positive MEMs from 42 sites, dataset B had 30, and dataset C had 12 AEMs with positive autocorrelation. For computation of the MEMs for the three datasets we used the packages *adespatial* (version 0.3-7, Dray et al., 2019) and *spdep* (version 0.7-4, Bivand & Piras, 2015; Bivand, Hauke, & Kossowski, 2013).

*Simulating communities with chosen spatial drivers*

We simulated realistic communities with known spatial structure, based on the three datasets. We used spatial eigenvectors as explanatory variables. We varied the number of MEMs with non-zero coefficients and created new binary (presence/absence) communities (with the same number of sites and same expected number of species as the real ones) using two different modelling scenarios. These simulated communities reflected the effect of those MEMs with non-zero coefficients. By varying the number and ordering of the non-zero coefficients, we could therefore control the spatial structure and scale of the simulated community data (see scheme in Fig. 1 and Table 1).

In order to simulate new binary communities under the simulated presence method (SPM, in which species are always detected if present), we first estimated a coefficient matrix $\mathbf{C}$ of size (*m* variables + 1 (first) row with intercepts) × *p* species from each real data set. This was achieved using the `manyglm` function with binomial errors in R package *mvabund* (version 3.11.9, Wang et al., 2012), with explanatory matrix **X** (*n* sites × *m* positive MEMs + an initial column of 1's ). The matrix $\mathbf{C}$ gives the effect of each explanatory variable on the logit-transformed probabilities of presence. The *mvabund* package provides a GLM framework for multivariate response data.

We then created new hypothetical scenarios by generating a new coefficient matrix $\mathbf{C}^{*}$, of the same size as $\mathbf{C}$, whose elements $c_{kj}^{*}$ are given by

$$\begin{cases} c_{kj}^{*} = c_{1j},\ if\ k=1,\ j=1,2,\ldots,p,\ \text{(intercepts)} \\ c_{kj}^{*} \sim \grave{F}_b,\ if\ k-1 \in K,\ j=1,2,\ldots,\ p, \\ c_{kj}^{*} = 0,\ \text{otherwise,} \end{cases} \qquad \text{eqn 1}$$

where $\grave{F}_b$ is the empirical distribution function of $c_{kj}$ (*k*=2, 3, …, *m*+1, *j*= 1, 2, …, *p*) (Evans et al., 2000), and the $b_{kj}^{*}$ are sampled with replacement. The set *K* defines to which rows of $\mathbf{C}^{*}$ the non-zero coefficients were allocated: we studied 14 such sets (see below and Table 1 a-c). In other words, we used the originally-estimated intercepts in each simulation (first row of eqn 1), and drew those coefficients assigned to non-zero values (second row of eqn 1) from the empirical distribution of all the originally-estimated explanatory variable coefficients. We sampled the values of the non-zero coefficients from the empirical distribution in order to simulate plausible but not fixed spatial structures. Table 1 depicts for each dataset how the non-zero coefficients were assigned for each dataset and simulation scenario (see below).

We then calculated predicted probabilities of presence $\grave{p}_{ij}$ for the *j*th species at the *i*th site. Given the matrix $\grave{\mathbf{Y}} = \mathbf{X}\mathbf{C}^{*}$ (*n* sites × *p* species) of predicted logit probabilities of presence, the predicted probability of presence is

$$\grave{p}_{ij} = \frac{\exp(\grave{y}_{ij})}{1+\exp(\grave{y}_{ij})}. \qquad \text{eqn 2}$$

The simulated presence/absence value for species *j* at site *i* was sampled from a Bernoulli distribution with success probability $\grave{p}_{ij}$. The result is a community matrix with the same number of sites and the same expected number of species as the real community, and with realistic coefficients for spatial eigenvectors. As in the maximum likelihood estimation done by `manyglm` (Wang et al., 2012), species and sites were assumed conditionally independent when generating simulated presence/absence data, given the values of the explanatory variables. Our simulated communities correspond to the simple case in which presence/absence patterns are affected by environmental variables but not interspecific interactions. Nevertheless, interspecific interactions could be well relevant to real world systems and other models (Godsoe & Harmon 2012; Anderson, 2017).

Since GLMs are specified correctly for presence/absence data generated this way, we would expect them to perform well. We therefore devised a second ecologically meaningful simulation method in which absences arise from the sampling protocol, called the simulated abundance method (SAM). The two simulation methods differ in whether they assume we have true absences or sampling-related absences. Note that it is not possible to simulate binary data directly using RDA, because RDA does not generate predicted probabilities of presence. Instead, we treated $\grave{\mathbf{Y}}$ as log expected abundances and exponentiated each element to get expected abundances $\lambda$. Then we calculated the probability of detecting the species under Poisson sampling (*i.e.* the probability of drawing a value of at least 1 from a Poisson distribution with parameter $\lambda$), which is

$$\grave{p}_{ij} = 1 - e^{-\lambda} \quad \text{eqn 3}$$

Finally, we generated a Bernoulli random variable with success probability $\grave{p}_{ij}$ to produce a simulated presence-absence observation. Both GLM and RDA are mis-specified for data generated in this way. Codes for both the SPM and SAM simulation frameworks and all the datasets used in our simulations are available as supplemental information (Data S1, S2 and S3).

We compared GLM and RDA variable selection under up to 14 different scenarios, differing in the number of non-zero coefficients (*nVar*) and whether these coefficients were associated with fine or broad spatial scales. We simulated up to six different choices of the number of MEM variables creating the spatial structure in the data (*i.e.* having non-zero coefficients): none, approximately one sixth, approximately one third, approximately half, approximately three-quarters, and all (Table 1 a-c, rows). We also simulated three different spatial scales of the patterns. As mentioned above, MEMs associated with larger eigenvalues represent broader spatial scales. We ordered the MEMs in descending order of eigenvalues and arranged the non-zero coefficients within matrix $\mathbf{C}^*$ in three different ways (Table I a-c, columns): only broad-scale MEMs with non-zero coefficients (scaling 1); only fine-scale MEMs with non-zero coefficients (scaling 2); half broad-scale, half fine-scale (scaling 3). Because not every combination of number of non-zero coefficients and spatial scaling is possible (*e.g.* it is not possible to assign one non-zero coefficient in scaling 3), there were 14 possible combinations overall for each dataset (Table 1). The main steps of the simulation scheme are summarized in Fig. 1.

### *RDA and GLM*

We used the default RDA function from the R package *vegan* (version 2.4-1, Oksanen et al*.,* 2016), with simulated community composition as the response variable, and MEMs associated with positive eigenvalues generated from geographical coordinates of the sample sites as explanatory variables. In order to perform a transformation-based RDA (Borcard et al., 2011; Blanchet et al., 2014) we used the Ochiai coefficient, which is the Hellinger transformation analogue for binary data, as recommended by Legendre & Gallagher (2001) and  Borcard et al. (2011).

Binomial GLMs were fitted to the same data using the `manyglm` function in R package *mvabund* (Wang et al., 2012). We fitted our models using a logistic regression (logit link function for binomial response), with species compositional data as the multivariate response variable and MEMs as predictors. No interaction terms were included, following common practice in spatial modelling of community data.

### *Comparing model selection between RDA and GLM frameworks*

We compared the results of model selection between the approach usually taken in the RDA and a somewhat-similar approach for GLMs. For RDA, we used the forward selection with double stopping criterion following Blanchet et al. (2008), beginning with a global test of significance (model with all spatial predictors) and carrying on with the variable selection if the global model was significant. The forward selection itself consists of a stepwise procedure including in the model the variable contributing the most to the adjusted $R^2$. The procedure stops either when the next variable with the highest contribution is not significant (first stopping criterion) or causes the adjusted $R^2$ to be bigger than that of the global model (*i.e.* containing all variables; second criterion). This is implemented in the function `ordiR2step` in the *vegan* package (Oksanen et al., 2016). For GLM, we used  forward selection with a stopping rule based on minimum Akaike Information Criterion (AIC) (Akaike, 1973; Wagenmakers & Farrell, 2004). The selection procedure started from a model with intercept only and added one explanatory variable at a time, until no further improvement in the sum of AIC over each of the response variables was possible. We used this approach because the usually large number of MEMs makes it difficult to compare the AIC sum over all possible GLMs.

The performance of each method on simulated data was mainly assessed by two criteria. First, we assessed how many MEMs with zero coefficients were incorrectly included in the final model. Second, we assessed how many MEMs with non-zero coefficients were incorrectly excluded from the final model. Also, we assessed overall accuracy (score) as the percentage of MEMs whose inclusion/exclusion status was correct. The goals of ecological studies are usually not directly related to the inclusion/exclusion of individual MEM variables, but instead to identify spatial pattern, represented by a linear combination of MEMs. However, since the MEMs form a basis for the space spanned by the transformed spatial weighting matrix, such a linear combination is unique (Fraleigh & Beauregard, 1995, pages 197-198). Furthermore, the MEMs are orthogonal, so that each represents a qualitatively distinct aspect of spatial pattern. Therefore, if an individual MEM is incorrectly included or excluded, the estimated spatial pattern is qualitatively wrong.

We further explored the ability of each method to capture spatial pattern using a graphical approach (Article S1). For each real dataset and each method, we haphazardly picked one simulated data set. We plotted the MEM decompositions of both the true and estimated spatial patterns. We chose the scenarios in which each method had the worst performance in terms of correctly including/excluding variables, in order to determine whether in such cases, overall spatial pattern would still be captured.

Finally, we calculated how much of the variation in response variables was explained by each method using the adjusted $R^2$ for the linear model in RDA and its analogue for GLMs, the *D*-value (Tjur, 2009). These two values cannot be directly compared since they are not exactly equivalent, but their results could yield interesting insights and are made available as supplemental information (see table results in Data S4).

For each of the combinations of conditions in Table I, 1000 simulated data sets were generated under each of SPM and SAM. For each simulated data set, spatial explanatory variables were selected using both GLM/AIC and RDA/FW.

## Results

Overall, GLM/AIC outperformed RDA/FW in selecting spatial explanatory variables when data were simulated under either SPM or SAM in all three scaling patterns (Fig.2).

In general, GLM/AIC had fairly predictable performance: it performed nearly perfectly when few or none of the available variables had non-zero true coefficients (*i.e.* $nVar = 0$, $^{m}/6$, $^{m}/3$ or $m/2$), but was less accurate when many or all the variables had non-zero true coefficients ( $nVar = 3^{m}/4$ or $nVar = m$) (blue lines in Fig.2 A-E). There was also some discernible pattern in RDA/FW's scores: it performed best at $nVar = 0$ and $nVar = m$, with intermediate values showing a considerable decrease in selection success. The loss of accuracy for intermediate values of $nVar$ (drop in red lines across different $nVar$ values in Fig.2 A-E) varied substantially among datasets, making general inferences about results more difficult. There was little difference between the results from the SPM and SAM simulations (Fig. 2B, D, F).

It is also noteworthy that when the model had a smaller number of variables to select from (River dataset C with 12 MEMs),  scores in GLM/AIC were higher, with virtually no incorrect inclusion of variables, and incorrect exclusion of variables occurring on average in only approximately 6% of all 14000 simulations over the whole set of replicates (Figure 3E). Under the same conditions, RDA/FW's rate of success was approximately 81%, incorrectly including variables at a rate of 18% (incorrect exclusions represented less than 1%) as depicted in Figure 3E.

Under both the SPM and SAM simulation methods, GLM/AIC differed substantially from the RDA/FW framework in regard to the type of errors it most often produced. GLM/AIC had virtually no incorrect inclusion of variables (Fig. 3, blue). However, when $nVar = 3^{m}/4$ or $nVar = m$ some variables that should be included in the final model were left out. Nevertheless, GLM/AIC never had less than around 90% accuracy over all three datasets (overall mean = 96 ± 1.3% against 71 ± 1.7% from RDA/FW). On the other hand, RDA/FW often included more variables than it should in the model (Fig. 3, red). Such errors especially occurred when $0 < nVar \leq 3^{m}/4$. Under some conditions, up to one third of the variables selected by RDA/FW had zero coefficients.

MEM decompositions of true and estimated spatial structure provided a visual assessment of the extent of the misspecification yielded by each method (Article S1). In all three datasets, the worst performance of GLM/AIC corresponded to those models in which it should have included all MEM variables (Fig. 2). Those scenarios represented communities structured at all spatial scales (broad, intermediate and fine). Despite incorrectly excluding several individual variables, GLM/AIC was capable of selecting subsets of variables that corresponded to all those scaling categories (Article S1.2-S1.7). In contrast, RDA/FW performed worse when there were few spatial variables (nVar = 5, nVar = 10 and nVar=2 for datasets A, B and C, respectively). Under those conditions, incorrect inclusion of variables also resulted in the inclusion of incorrect spatial scales. For example, in one simulation from dataset A (Article S1.8) the true spatial structure contained only five MEMs describing finer spatial scale patterns (scaling 2 = MEMs 12-16). However, the final model selected by RDA/FW included 13 variables describing both broad (MEMs 1-6) and intermediate spatial scales (MEMs 9, 11), along with the correct ones (Article S1.9). Similar results were found in all three datasets (Article S1.10-S1.13). Moreover, these incorrect inclusions of individual variables by RDA/FW resulted in the inclusion of MEM variables associated to eigenvalues substantially different from the correct ones, representing spatial scales much larger than those actually present in the data (Article S1.14). For matters of space, we only plotted one failure example from each dataset for both GLM/AIC and RDA/FW. However, the correct spatial structures within simulated communities and those structures retrieved by both methods in all our simulations scenarios are available as supplemental data (Data S5).

Under SPM simulations, the scale of spatial pattern (fine, broad or mixed: scaling 1, 2 and 3, respectively) had negligible effect on GLM/AIC performance (Fig.4A, C, E). A slight difference in variable selection scores between scaling 1 to 2 and 3 was only found in one modelling condition (Fig. 4, $nVar = 3^{m}/4$). On the other hand, scaling often affected the performance of RDA/ FW, although there was no obvious general pattern across different conditions and datasets (Fig.4A, C, E). Under SAM simulations, both frameworks performed similarly to what was observed under SPM (Fig.4B, D, F).

## Discussion

Here, we showed that a GLM/AIC-based method for finding spatial structure in communities outperformed an RDA/FW-based method, for presence-absence data simulated under two different ecologically plausible scenarios about how absences arise. We based our simulated datasets on real datasets from marine, terrestrial and freshwater data. Notably, differences in assumptions about how absences arise made little difference to performance. This might be due to the structure of our community presence/absence datasets, which (like most ecological datasets) had many rare species and, therefore, many expected abundances close to zero. In such cases, the relationship between the community data and explanatory variables could be approximated by a binomial GLM with a logit link function, even if this was not the correct model (as in the SAM simulations). We therefore focus below on general patterns that apply equally to both assumptions about absences, rather than on the details of these assumptions.

In selecting spatial explanatory variables, GLM followed by AIC-based model selection (GLM/AIC) performed better than the widely-used approach of RDA followed by forward selection (RDA/FW). Not only did GLM/AIC have better performance overall, but its performance varied little between simulation conditions (Fig. 2). In contrast, RDA/FW performed unpredictably, but often retained too many explanatory variables (Fig. 3).

The problems arising from data with non-Gaussian error distributions, such as classic community presence and absence data, in a linear modelling framework are not new to science ( Legendre & Gallagher, 2001; McCullagh & Nelder, 1989; Wolda, 1981). Classical linear models such as RDA ( Legendre & Anderson, 1999; Legendre & Legendre, 2012) make assumptions regarding constancy of variance in the data (ter Braak & Prentice, 1988) that cannot be true for presence-absence data, even after data transformation (O'Hara & Kotze, 2010; Warton, 2018; Warton et al., 2012). The problem may be negligible in some hypothesis testing situations (Ives, 2015). Regardless, incorrectly assuming linearity (and constant variance) may lead to serious problems. Unfortunately, RDA is an algorithmic method that makes implicit decisions about the distribution of variances (ter Braak & Prentice, 1988; Warton et al., 2012) and does not provide the flexibility to separate systematic variation from random variation in the way that statistical models such as GLMs do (Warton et al., 2015; and see O'Neil & Schutt, 2013 for differences between algorithms and statistical models). New frameworks, such as using GLMs with spatially-structured random effects (followed by variation partitioning to find environmental and spatial components) have also been specifically proposed as a model-based alternative to MEMs (Ovaskainen et al., 2017). Despite recent advances showing that better estimates could be obtained by using sensible selection procedures, manipulating the data appropriately and/or by splitting the analysis of the response data over shorter spatial/environmental gradients (Bauman, Drouet, Dray, et al., 2018; Ives, 2015; Vieira et al., 2019), employing statistical models that match the distribution of the response data is better practice in most cases (Ferrier et al., 2007; Warton, 2018; Warton et al., 2015).

Another relevant aspect of the general performances of the two methods concerns the peaks of performance in detecting spatial structure. The scores in the GLM/AIC framework were close to ideal across datasets when the number of variables that should be selected was none or was small relative to the number of variables available. The performance only decayed when many or all of the available variables should have been retained in the final model. Thus, if a few variables are responsible for most of the spatial structure in community composition, GLM/AIC will usually outperform RDA/FW (Fig. 2). Considering that the majority of effects could be derived from a small number of causes (Sullivan, 2019) in many biological systems, GLM/AIC could presumably perform well on many real systems. On the other hand, RDA/FW worked best precisely in situations thought unlikely in real systems, when no spatial structure is present among communities (where GLM/AIC also performed equally well), or when composition is structured at all possible spatial scales (i.e., $nVar = 0$ and $nVar = m$, respectively). Moreover, when the model had a small number of variables to select from (River dataset, Fig.3E-F), performance of RDA/FW was very variable (Fig. 3E-F).

The two approaches also differed in the ways they failed. GLM/more often included too few variables, while RDA/FW more often included too many. This was consistent among all three datasets under SPM and SAM simulations (Fig. 3) and is in contrast with results from previous studies where GLMs produced higher Type I error rates compared to a linear model (Ives, 2015). For beta diversity studies, where the aim is to identify the most important variables associated with differences in community composition, leaving out a few variables that affect composition is better, in our opinion, than including many variables whose effects are not important. On the contrary, in other scenarios such as when one tries to select pivotal attributes that could be important for the conservation of a population or community, it might be better to accept a higher risk of including spurious variables. Furthermore, model selection problems involve a trade-off between bias and variance, with inclusion of unnecessary variables inflating the uncertainty in parameter estimates (Miller, 1990). Using AIC is often a good way to deal with this trade-off ( Anderson et al., 2000), and in our simulations, an AIC-based approach worked well. Thus, we suggest that GLM/AIC will usually outperform RDA/FW in selecting spatial explanatory variables for presence/absence community composition data. Unfortunately, AIC-like statistics are not recommended for constrained ordination methods such as RDA, and therefore its use cannot be trusted (see below and Bauman et al., 2018 for details). When different RDA-based procedures were systematically compared, the commonly (mis)used combination of RDA and AIC model selection produced the worst results, yielding inflated Type I errors rates (Bauman, Drouet, Dray, et al., 2018). Therefore, the benefits from AIC in dealing with the bias and variance trade-off do not apply to RDA or related ordination methods. Despite our interest in some attributes of the MEMs for our simulations, such as differences in model performance under varying spatial scales, we hypothesize that the results demonstrated here hold true for other types of explanatory variables (*e.g*. environmental) not tested here.

The spatial scale represented by the MEMs had a negligible effect on GLM/AIC's performance, with only one condition in one dataset slightly differing in results between different scales (see Fig.4 when the number of non-zeros is $\lfloor 3m/4 \rfloor$). In contrast, RDA/FW's performance was strongly affected by spatial scale (Fig.4). In real systems, where the spatial scale at which community composition varies is not known *a priori*, the performance of RDA/FW could therefore be unpredictable. The uncertainty around RDA/FW performance over differing spatial scales could be especially troublesome for analyses involving processes that may not be constant along spatial/environmental gradients, as commonly observed for rates of species turnover, for example (Ferrier et al., 2007; Fitzpatrick et al., 2013).

## Conclusions

We discourage the use of traditional RDA/FW to search for spatial descriptors of variation in multivariate presence/absence data sets of moderate size, although large datasets could potentially overcome the issues found here. Instead, we recommend the GLM/AIC framework, in which the relationship between the response and its predictors is modelled in a way that respects the nature of the response. Similar recommendations are likely to apply to other forms of community abundance data with non-normal error distributions (e.g. count data with many zeros or proportional data, Bolker et al., 2009; Warton et al., 2012, 2016).

## Acknowledgements

We thank the James Hutton Institute, Aberdeen, for providing data. We are also grateful for Dr Petr Šmilauer for valuable suggestions given at BES 2015 and Dr Ian Smith for technical support.

## References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Caski (Eds.), *Proceedings of the Second International Symposium on*

*Information Theory* (pp. 267–281). Akademiai Kiado.

Anderson, D. R., Burnham, K. P., & Thompson, W. L. (2000). Null hypothesis testing problems prevalence and an alternative. *The Journal of Wildlife Management*, *64*(4), 912–923.

Anderson, R. P. (2017). When and how should biotic interactions be considered in models of species niches and distributions? *Journal of Biogeography*, *44*(1), 8–17. https://doi.org/10.1111/jbi.12825

Bauman, D., Drouet, T., Dray, S., & Vleminckx, J. (2018). Disentangling good from bad practices in the selection of spatial or phylogenetic eigenvectors. *Ecography*, *41*(10), 1638–1649. https://doi.org/10.1111/ecog.03380

Bauman, D., Drouet, T., Fortin, M.-J., & Dray, S. (2018). Optimizing the choice of a spatial weighting matrix in eigenvector-based methods. *Ecology*, *99*(10), 2159–2166. https://doi.org/10.1002/ecy.2469

Bauman, D., Raspé, O., Meerts, P., Degreef, J., Ilunga Muledi, J., & Drouet, T. (2016). Multiscale assemblage of an ectomycorrhizal fungal community: the influence of host functional traits and soil properties in a 10-ha miombo forest. *FEMS Microbiology Ecology*, *92*(10). https://doi.org/10.1093/femsec/fiw151

Bivand, R., Hauke, J., & Kossowski, T. (2013). Computing the Jacobian in Gaussian Spatial Autoregressive Models: An Illustrated Comparison of Available Methods. *Geographical Analysis*, *45*(2), 150–179. https://doi.org/10.1111/gean.12008

Bivand, R., & Piras, G. (2015). Comparing Implementations of Estimation Methods for Spatial Econometrics. *Journal of Statistical Software*, *63*(18). https://doi.org/10.18637/jss.v063.i18

Blanchet, F. G., Legendre, P., & Borcard, D. (2008). Forward selection of spatial explanatory variables. *Ecology*, *89*(9), 2623–2632. https://doi.org/10.1890/07-0986.1

Blanchet, F. G, Legendre, P., & Borcard, D. (2008). Modelling directional spatial processes in ecological data. *Ecological Modelling*, *215*(4), 325–336. https://doi.org/10.1016/j.ecolmodel.2008.04.001

Blanchet, F. G., Legendre, P., Bergeron, J. A. C., & He, F. (2014). Consensus RDA across dissimilarity coefficients for canonical ordination of community composition data. *Ecological Monographs*, *84*(3), 491–511. https://doi.org/10.1890/13-0648.1

Blanchet, F. G., Legendre, P., Bergeron, J. A., & He, F. (2014). Consensus RDA across dissimilarity coefficients for canonical\nordination of community composition data. *Ecological Monographs*, *84*(3), 491–511. https://doi.org/10.1890/13-0648.1

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., & White, J. S. S. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology and Evolution*, *24*(3), 127–135. https://doi.org/10.1016/j.tree.2008.10.008

Borcard, D., Gillet, F., & Legendre, P. (2011). *Numerical Ecology with R*. Springer-Verlag New York. https://doi.org/10.1007/978-1-4419-7976-6

Borcard, D., & Legendre, P. (2002). All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices. *Ecological Modelling*, *153*(1–2), 51–68. https://doi.org/10.1016/S0304-3800(01)00501-4

Borcard, D., Legendre, P., & Drapeau, P. (1992). Partialling out the Spatial Component of Ecological Variation Author ( s ): Daniel Borcard , Pierre Legendre and Pierre Drapeau Published by : Ecological Society of America PARTIALLING OUT THE SPATIAL COMPONENT OF ECOLOGICAL VARIATION1. *Ecology*, *73*(3), 1045–1055. https://doi.org/10.2307/1940179

Carlos-Júnior, L. A., Spencer, M., Neves, D. M., Moulton, T. P., Pires, D. de O., e Castro, C. B., Ventura, C. R. R., Ferreira, C. E. L., Serejo, C. S., Oigman-Pszczol, S., Casares, F. A., Mantelatto, M. C., & Creed, J. C. (2019). Rarity and beta diversity assessment as tools for guiding conservation strategies in marine tropical subtidal communities. *Diversity and Distributions*. https://doi.org/10.1111/ddi.12896

Diniz-Filho, J. A. F., Bini, L. M., Rangel, T. F., Morales-Castilla, I., Olalla-Tárraga, M. Á., Rodríguez, M. Á., & Hawkins, B. A. (2012). On the selection of phylogenetic eigenvectors for ecological analyses. *Ecography*, *35*(3), 239–249. https://doi.org/10.1111/j.1600-0587.2011.06949.x

Dray, S., Bauman, D., Blanchet, F. G., Borcard, D., Clappe, S., Guenard, G., Jombart, T., Larocque, G., Legendre, P., Madi, N., & Wagner, H. H. (2019). *adespatial: Multivariate Multiscale Spatial Analysis* (0.3-7). https://cran.r-project.org/package=adespatial

Dray, S., Legendre, P., & Peres-Neto, P. R. (2006). Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbour matrices (PCNM). *Ecological Modelling*, *196*(3–4), 483–493. https://doi.org/10.1016/j.ecolmodel.2006.02.015

Eisenlohr, P. V., & Oliveira-Filho, A. T. de. (2015). Revisiting patterns of tree species composition and their driving forces in the Atlantic Forests of Southeastern Brazil. *Biotropica*, *47*(6), 689–701. https://doi.org/10.1111/btp.12254

Evans, M., Hastings, N., & Peacock, B. (2000). Statistical Distributions. In *New York* (Vol. 2, Issue 4). Willey. https://doi.org/10.1002/9780470627242

Ferrier, S., Manion, G., Elith, J., & Richardson, K. (2007). Using generalized dissimilarity modelling to analyse and predict patterns of beta diversity in regional biodiversity assessment. *Diversity and Distributions*, *13*(3), 252–264. https://doi.org/10.1111/j.1472-4642.2007.00341.x

Fitzpatrick, M. C., Sanders, N. J., Normand, S., Svenning, J. C., Ferrier, S., Gove, A. D., & Dunn, R. R. (2013). Environmental and historical imprints on beta diversity: Insights from variation in rates of species turnover along gradients. *Proceedings of the Royal Society B: Biological Sciences*, *280*(1768). https://doi.org/10.1098/rspb.2013.1201

Fraleigh, J., & Beauregard, R. (1995). *Linear algebra* (3rd ed.). Addison Wesley.

Godínez-Domínguez, E., & Freire, J. (2003). Information-theoretic approach for selection of spatial and temporal models of community organization. *Marine Ecology Progress Series*, *253*, 17–24.

Godsoe, W., & Harmon, L. J. (2012). How do species interactions affect species distribution models? *Ecography*, *35*(9), 811–820. https://doi.org/10.1111/j.1600-0587.2011.07103.x

Ives, A. R. (2015). For testing the significance of regression coefficients, go ahead and log-transform count data. *Methods in Ecology and Evolution*, *6*(7), 828–835. https://doi.org/10.1111/2041-210X.12386

Legendre, P., & Anderson, M. J. (1999). Distance-based redundancy analysis: Testing multispecies responses in multifactorial ecological experiments. *Ecological Monographs*, *69*(1), 1–24.

Legendre, P. (1993). Spatial Autocorrelation : Trouble or New Paradigm ? *Ecology*, *74*(6), 1659–1673.

Legendre, P., & Gallagher, E. D. (2001). Ecologically meaningful transformations for ordination of species data. *Oecologia*, *129*(2), 271–280. https://doi.org/10.1007/s004420100716

Legendre, P., & Legendre, L. (2012). *Numerical Ecology* (Third Engl). Elsevier Ltd.

Lewis, R. J., Pakeman, R. J., & Marrs, R. H. (2014). Identifying the multi-scale spatial structure of plant community determinants of an important national resource. *Journal of Vegetation Science*, *25*(1), 184–197.

McCullagh, P., & Nelder, J. A. (1989). *Generalized Linear Models* (2° Edition). Chapman and Hall/CRC.

Miller, A. (1990). *Subset Selection in Regression*. Chapman and Hall.

Neves, D. M., Dexter, K. G., Pennington, R. T., Bueno, M. L., & Oliveira Filho, A. T. (2015). Environmental and historical controls of floristic composition across the South American Dry Diagonal. *Journal of Biogeography*, *42*(8), 1566–1576. https://doi.org/10.1111/jbi.12529

O'Hara, R. B., & Kotze, D. J. (2010). Do not log-transform count data. *Methods in Ecology and Evolution*, *1*(2), 118–122. https://doi.org/10.1111/j.2041-210X.2010.00021.x

O'Neil, C., & Schutt, R. (2013). *Doing Data Science* (First). O'Reilly.

Oksanen, A. J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., Mcglinn, D., Minchin, P. R., Hara, R. B. O., Simpson, G. L., Solymos, P., Stevens, M. H. H., & Szoecs, E. (2016). *Package ' vegan .'*

Ovaskainen, O., Tikhonov, G., Norberg, A., Guillaume Blanchet, F., Duan, L., Dunson, D., Roslin, T., & Abrego, N. (2017). How to make more out of community data? A conceptual framework and its implementation as models and software. *Ecology Letters*, *20*(5), 561–576. https://doi.org/10.1111/ele.12757

Peres-Neto, P. R., & Legendre, P. (2010). Estimating and controlling for spatial structure in the study of ecological communities. *Global Ecology and Biogeography*, *19*(2), 174–184. https://doi.org/10.1111/j.1466-8238.2009.00506.x

Saiter, F. Z., Eisenlohr, P. V., Barbosa, M. R. V., Thomas, W. W., & Oliveira-Filho, A. T. de. (2015). From evergreen to deciduous tropical forests: how energy–water balance, temperature, and space influence the tree species composition in a high diversity region. *Plant Ecology & Diversity*, *9*(October), 1–10. https://doi.org/10.1080/17550874.2015.1075623

Sullivan, W. (2019). Rockets, gauges, and pendulums: applying engineering principles to cell biology. *Molecular Biology of the Cell*, *30*(14), 1635–1640. https://doi.org/10.1091/mbc.E19-02-0100

Ter Braak, C. J. F., & Prentice, I. C. (1988). A Theory of Gradient Analysis. *Advances in Ecological Research*, *18*(C), 271–317. https://doi.org/10.1016/S0065-2504(08)60183-X

Tjur, T. (2009). Coefficients of Determination in Logistic Regression Models—A New Proposal: The Coefficient of Discrimination. *American Statistician*, *63*(4), 366–372. https://doi.org/10.1198/tast.2009.08210

Vieira, D. C., Brustolin, M. C., Ferreira, F. C., & Fonseca, G. (2019). segRDA: An R package for performing piecewise redundancy analysis. *Methods in Ecology and Evolution*, *1*(1), 2041-210X.13300. https://doi.org/10.1111/2041-210X.13300

Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, *11*(1), 192–196. https://doi.org/10.3758/BF03206482

Wang, Y., Naumann, U., Wright, S. T., & Warton, D. I. (2012). Mvabund- an R package for model-based analysis of multivariate abundance data. *Methods in Ecology and Evolution*, *3*(3), 471–474. https://doi.org/10.1111/j.2041-210X.2012.00190.x

Warton, D. I. (2018). Why you cannot transform your way out of trouble for small counts. *Biometrics*, *74*(1), 362–368. https://doi.org/10.1111/biom.12728

Warton, D. I., Foster, S. D., De'ath, G., Stoklosa, J., & Dunstan, P. K. (2015). Model-based thinking for community ecology. *Plant Ecology*, *216*(5), 669–682. https://doi.org/10.1007/s11258-014-0366-3

Warton, D. I., Lyonsy, M., Stoklosa, J., & Ivesz, A. R. (2016). Three points to consider when choosing a LM or GLM test for count data. *Methods in Ecology and Evolution*, *7*, 882–890. https://doi.org/10.1111/2041-210X.12552

Warton, D. I., Wright, S. T., & Wang, Y. (2012). Distance-based multivariate analyses confound location and dispersion effects. *Methods in Ecology and Evolution*, *3*(1), 89–101. https://doi.org/10.1111/j.2041-210X.2011.00127.x

Wolda, H. (1981). Similarity indices, sample size and diversity. *Oecologia*, *50*(3), 296–302. https://doi.org/10.1007/BF00344966

Yee, T. W. (2006). Constrained additive ordination. *Ecology*, *87*(1), 203–213. http://www.ncbi.nlm.nih.gov/pubmed/16634311

## Table 1(on next page)

Simulation scenarios for the three datasets as described in main text.

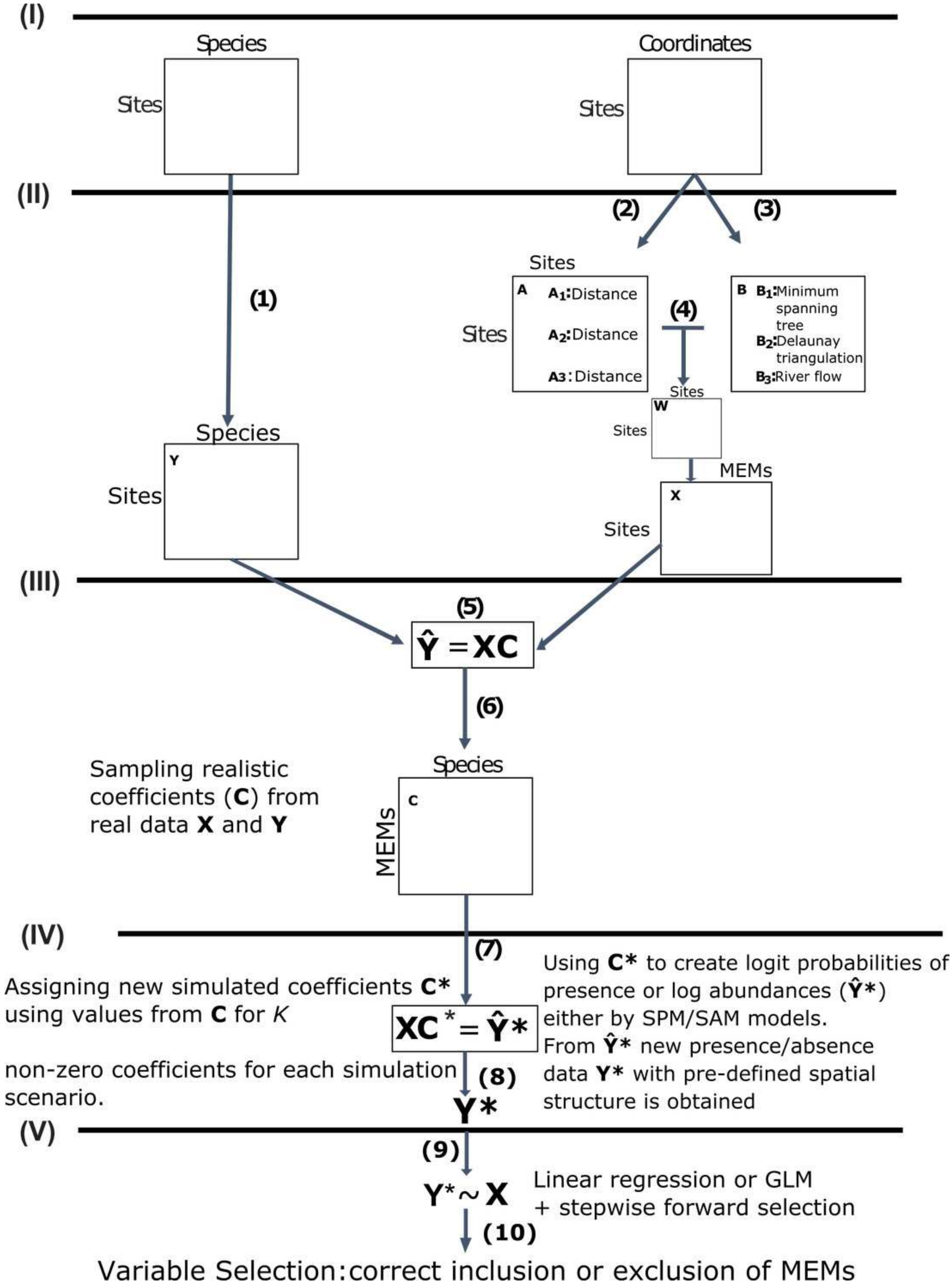
Distribution of MEM variables with non-zero coefficient under each simulation scenario in all three datasets (A = marine algae from Ilha Grande Bay, *m* = 16; B= Scotland grasslands, *m* = 30; C= freshwater insects, *m* = 12). Rows and columns define all simulation scenarios regarding the number of variables to be used and their position. Rows represent the number of non-zero variables to be included based on set *K* (see main text), whereas columns define the scaling of these non-zero variables, *i.e.* position to which those non-zero variables would be assigned. Scaling 1 assigned non-zero coefficients only to MEMs associated with larger eigenvalues representing broader spatial scales. Scaling 2 assigned non-zero coefficients only to MEMs associated with smaller eigenvalues, representing finer spatial scales. Scaling 3 assigned non-zero coefficients to MEMs representing a range of spatial scales. Cells contain sets of indices of explanatory variables. When nVar=0, none of the variables had non-zero coefficients.

| | | | Scaling | |
|---|---|---|---|---|
| | | 1 (only broad) | 2 (only fine) | 3 (mixed) |
| (A) | 0 | None | - | - |
| | $\lfloor m/6 \rfloor$ | {1,2} | {15,16} | {1,16} |
| | $\lfloor m/3 \rfloor$ | {1,2,3,4,5} | {12,13,14,15,16} | {1,2,3,15,16} |
| | $\lfloor m/2 \rfloor$ | {1,2,...,8} | {9,11,...,16} | {1,2,3,4,13,14,15,16} |
| | $\lfloor 3m/4 \rfloor$ | {1,2,...,12} | {5,7,...,16} | {1,2,...,6,11,12,...,16} |
| | $m$ | {1,2,...,16} | - | - |
| (B) | 0 | None | - | - |
| | $\lfloor m/6 \rfloor$ | {1,2,3,4,5} | {26,27,28,29,30} | {1,2,3,29,30} |
| | $\lfloor m/3 \rfloor$ | {1,2,...,10} | {21,22,...,30} | {1,2,...,10,21,22,...,30} |
| | $\lfloor m/2 \rfloor$ | {1,2,...,15} | {16,17,...,30} | {1,2,...,8,24,25,...,30} |
| | $\lfloor 3m/4 \rfloor$ | {1,2,...,22} | {6,7,...,30} | {1,2,...,11,21,22,...,30} |
| | $m$ | {1,2,...,30} | - | - |
| (C) | | | | |
| | 0 | None | - | - |
| | $\lfloor m/6 \rfloor$ | {1,2} | {11,12} | {1,12} |
| | $\lfloor m/3 \rfloor$ | {1,2,3,4} | {9,10,11,12} | {1,2,11,12} |
| | $\lfloor m/2 \rfloor$ | {1,2,…,6} | {7,8,…,12} | {1,2,3,10,11,12} |
| | $\lfloor 3m/4 \rfloor$ | {1,2,…,9} | {4,5,…,12} | {1,2, 3,4, 5,9,10,11,12} |
| | $m$ | {1,2,…,12} | - | - |

# Figure 1

Schematic diagram of the main steps used in this study to simulate community presence/absence data with pre-defined spatial structure.

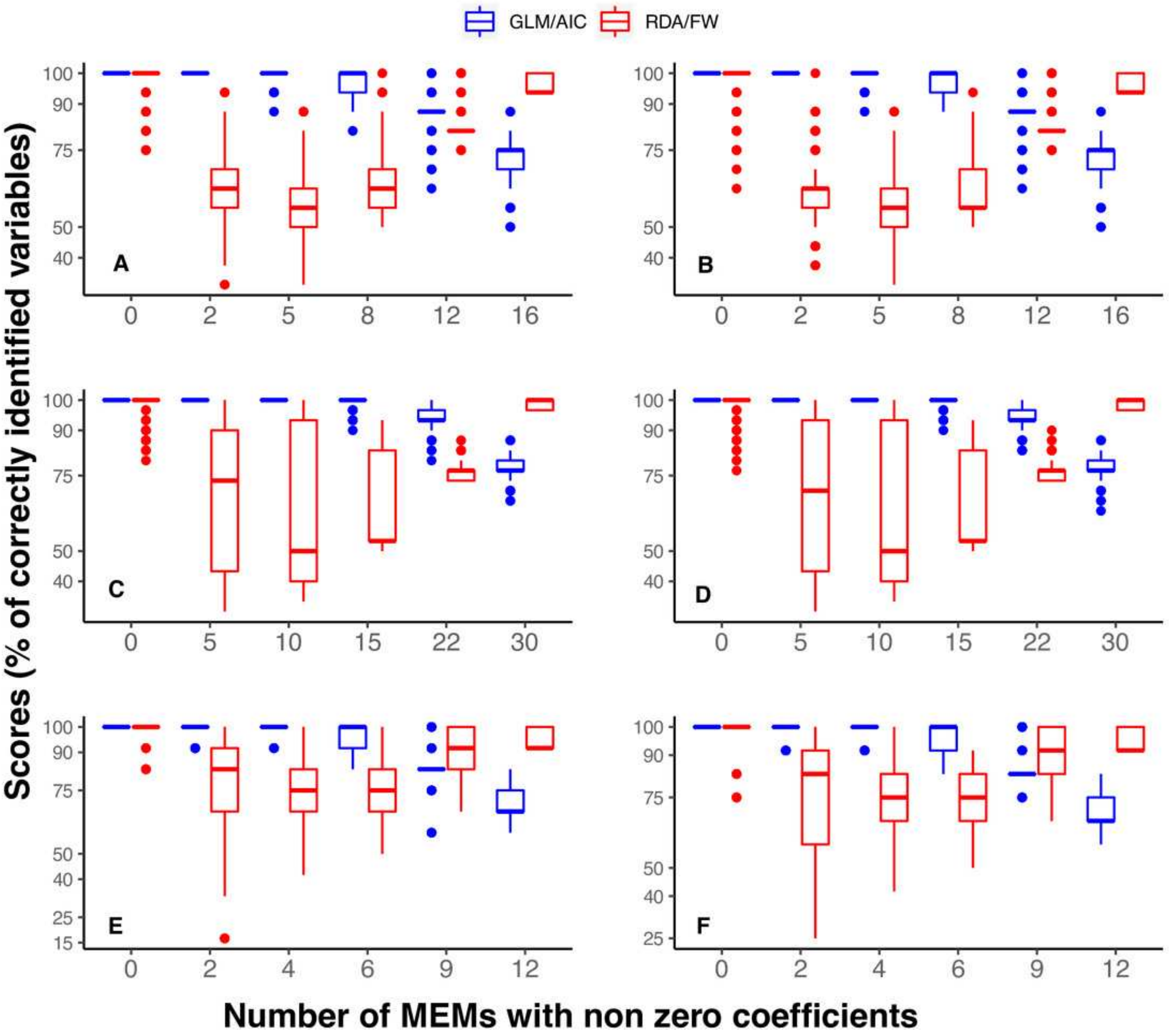
**Data acquisition (I)**: We used real data from marine, terrestrial and freshwater communities and their respective sampling site coordinates as our baseline datasets. **Obtaining response and predictor matrices (II)**: Those datasets were used to construct a response matrix of presence/absence data **Y** (1) and a matrix **X** of spatial explanatory variables called MEMs. The spatial variables were obtained from a pairwise site-by-site distance matrix **A** (2) and a connectivity matrix **B** (3) describing the spatial relationship among sites (see main text for specific decisions for each dataset). The Hadamard product of these two matrices generates the spatial weighting matrix **W** (4), which is then doubly centred and diagonalised, yielding eigenvectors to be used as spatial variables, represented below by matrix **X**. **Obtaining realistic coefficients for spatial variables (III).** From a Generalized Linear Model (GLMs) for the relationship between **Y** and **X** (5) we obtained a matrix **C** of realistic regression coefficients (6). **Using non-zero coefficients to model new presence/absence data with pre-defined spatial structure (IV)**: We sampled different numbers of non-zero coefficients from **C** under 14 distinct scenarios (see main text) to build a new matrix **C*** and then left-multiplied **C*** by **X (7)** to obtain matrix **Ŷ*.** This matrix represented the logit predicted probabilities of presence or a matrix of log abundances, depending on which of two models that differed, respectively, in assumptions regarding absences as real (simulated presence model, SPM) or artifacts derived from poor sampling (SAM). From **Ŷ*** we estimated (8) new presence/absence data **Y*** containing the spatial structure defined by **C***. **Using GLM/AIC and RDA/FW to select spatial models using the simulated presence/absence data (V):** Finally, we regressed **Y*** against **X** using the GLM/AIC and RDA/FW frameworks (9) to assess which MEMs would be correctly selected by those two methods. The performance of each method was mainly assessed by the proportion of MEM variables that were correctly included or excluded from final models by each method (10).


(I)
Species
Sites
Coordinates
Sites
(II)
(1)
(2)
(3)
Sites
Sites
A
A1:Distance
A2:Distance
A3:Distance
(4)
B
B1:Minimum spanning tree
B2:Delaunay triangulation
B3:River flow
Sites
W
Sites
Species
Y
Sites
MEMs
X
Sites
(III)
(5)
Ŷ = XC
(6)
Species
C
MEMs
Sampling realistic coefficients (C) from real data X and Y
(IV)
(7)
Assigning new simulated coefficients C* using values from C for K
non-zero coefficients for each simulation scenario.
XC* = Ŷ*
Using C* to create logit probabilities of presence or log abundances (Ŷ*) either by SPM/SAM models.
From Ŷ* new presence/absence data Y* with pre-defined spatial structure is obtained
(8)
Y*
(V)
(9)
Y* ~ X
Linear regression or GLM + stepwise forward selection
(10)
Variable Selection: correct inclusion or exclusion of MEMs

# Figure 2

Overall performance comparison between GLM/AIC (blue) and RDA/FW (red) methods on simulated presence/absence data.

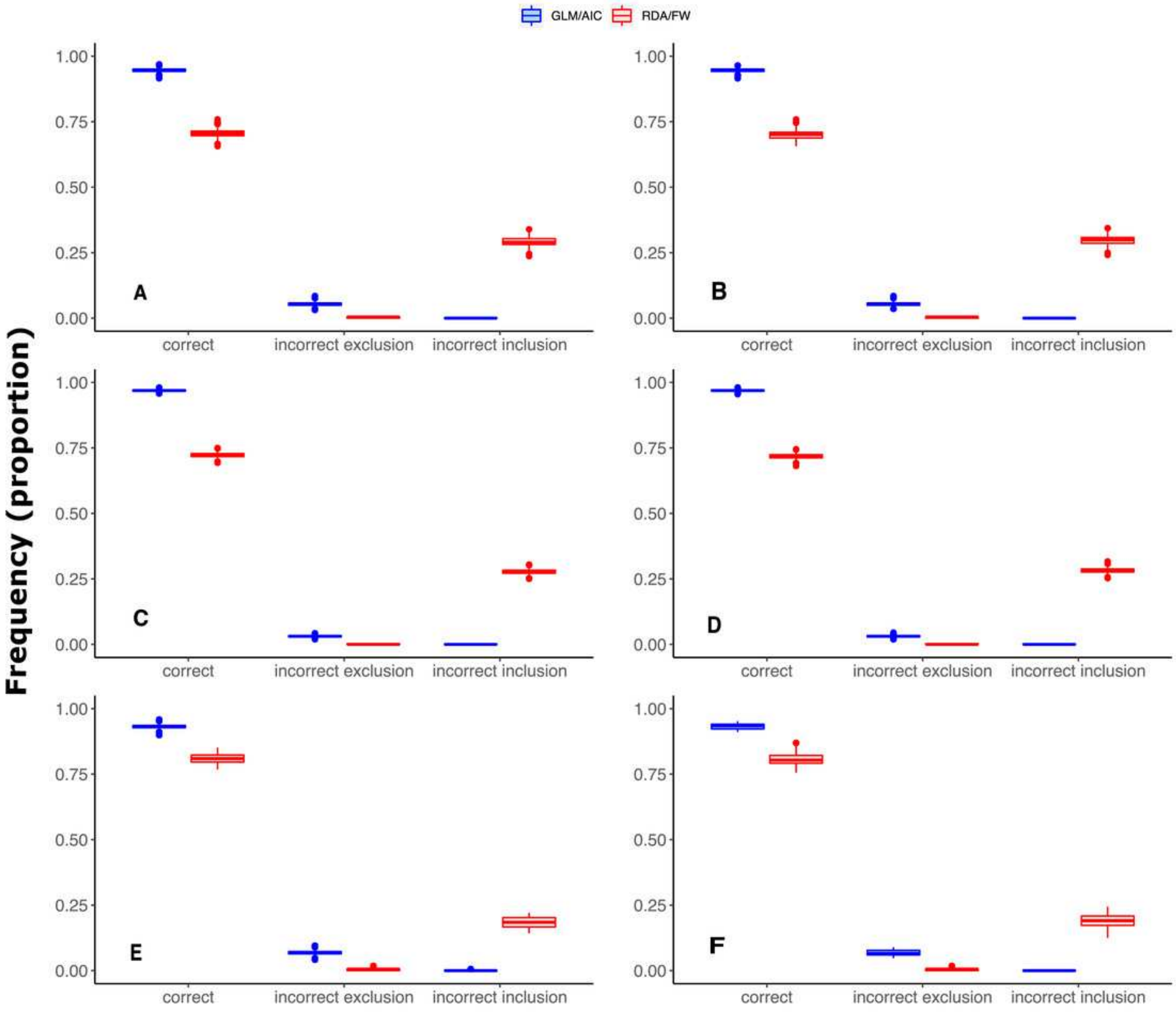
Scores were measured by counting the percentage of MEMs correctly included/excluded from the final model out of the total number of variables in each dataset (A = 16, B = 30, C= 2). This comparison was made across varying numbers of MEMs with non-zero coefficients (x axis). (A, D) simulated data based on subtidal macroalgae in Ilha Grande Bay ; (B, E) data based on plant species from Scottish grassland and (C, F) data based on aquatic macroinvertebrate insect species from a river in Brazil. Panels A, C and E depict results where community presence/absence data was simulated directly from real coefficients (SPM, see main text) whereas B, D and F show simulation results where presence/absence data was estimated from expected abundances (SAM).


GLM/AIC
RDA/FW
Scores (% of correctly identified variables)
A
B
C
D
E
F
100
90
75
50
40
25
15
0
2
5
8
12
16
10
15
22
30
4
6
9
Number of MEMs with non zero coefficients

# Figure 3

Differences in performance between GLM/AIC and RDA/FW frameworks regarding the proportion of incorrect inclusions/exclusions of explanatory variables across 1000 simulations for each method.

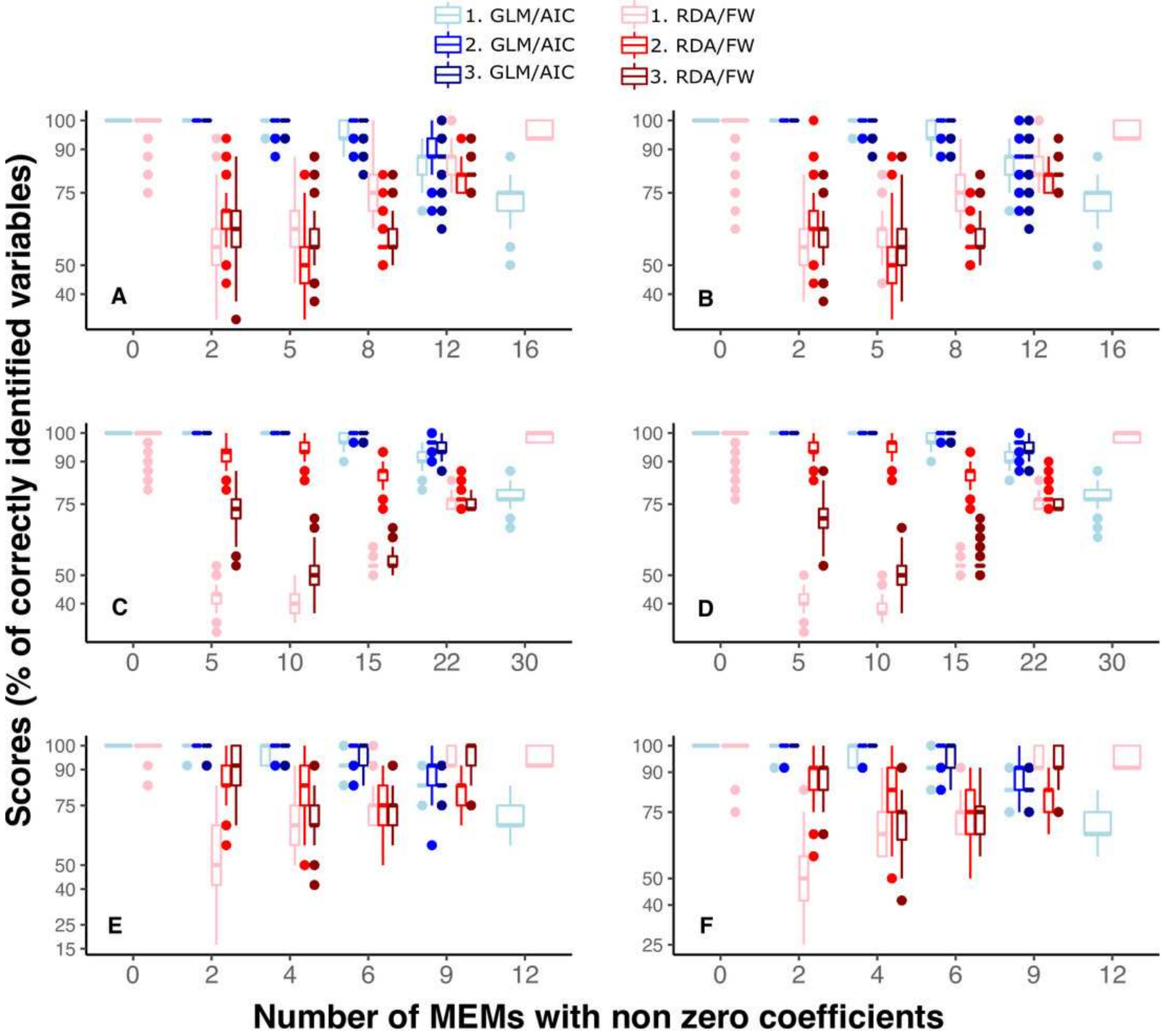
Panels A, C and E depict results where community presence/absence data was simulated direclty from real coefficients (SPM, see main text) whereas B, D and F show simulation results where presence/absence data was estimated from expected abundances (SAM). Panels A and D depict results for simulated data based on subtidal macroalgae in Ilha Grande Bay; panels B and E represent data based on plant species from Scottish grassland; and panels C and F represent data based on aquatic macroinvertebrate insect species from a river in Brazil. Darker lines represent mean values.


GLM/AIC
RDA/FW
Frequency (proportion)
1.00
0.75
0.50
0.25
0.00
A
B
C
D
E
F
correct
incorrect exclusion
incorrect inclusion

# Figure 4

Performance of GLM/AIC (blue) and RDA/FW (red) modelling approaches under variation in spatial scales of MEMs with non-zero coefficients.

Spatial scale was defined as broad (1), fine (2) or mixed (3) (where applicable). (A, B) simulated data based on macroalgae in Ilha Grande Bay ; (C, D) data based on plant species from Scottish grassland and (E, F) data based on aquatic macroinvertebrate insect species from a river in Brazil. Panels A, C and E depict results where community presence/absence data was simulated direclty from real coefficients (SPM) whereas B, D and F show simulation results where presence/absence data was estimated from expected abundances (SAM, see main text).


1. GLM/AIC
2. GLM/AIC
3. GLM/AIC
1. RDA/FW
2. RDA/FW
3. RDA/FW
Scores (% of correctly identified variables)
100
90
75
50
40
25
15
A
B
C
D
E
F
0
2
5
8
12
16
10
15
22
30
4
6
9
Number of MEMs with non zero coefficients