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

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

9 [1] School of Environmental Sciences, University of Liverpool. Liverpool, UK.
10 [2] Programa de Pós-Graduação em Ecologia e Evolução, Universidade do Estado do Rio de
11 Janeiro. Rio de Janeiro, Brazil.
12 [3] Departamento de Biologia, Pontifícia Universidade Católica do Rio de Janeiro. Rio de Janeiro,
13 Brazil.
14 [4] Departamento de Ecologia, Universidade do Estado do Rio de Janeiro. Rio de Janeiro, Brazil.
15 [5] Department of Bioscience, Norwegian Institute of Bioeconomy Research. Bergen,
16 Norway.
17 [6] Division of Biological Sciences, University of Montana. Missoula, US.

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

Abstract

26   **Background.** Ecological communities tend to be spatially structured due to
27 environmental gradients and/or spatially contagious processes such as growth, dispersion and
28 species interactions. Data transformation followed by usage of algorithms such as Redundancy
29 Analysis (RDA) is a fairly common approach in studies searching for spatial structure in
30 ecological communities, despite recent suggestions advocating the use of Generalized Linear
31 Models (GLMs). Here, we compared the performance of GLMs and RDA in describing spatial
32 structure in ecological community composition data. We simulated realistic presence/absence
33 data typical of many β-diversity studies. For model selection we used standard methods
34 commonly used in most studies involving RDA and GLMs.
35   **Methods.** We simulated communities with known spatial structure, based on three real
36 spatial community presence/absence datasets (one terrestrial, one marine and one freshwater).
37 We used spatial eigenvectors as explanatory variables. We varied the number of non-zero
38 coefficients of the spatial variables, and the spatial scales with which these coefficients were
39 associated and then compared the performance of GLMs and RDA frameworks to correctly
40 retrieve the spatial patterns contained in the simulated communities. We used two different
41 methods for model selection, Forward Selection (FW) for RDA and the Akaike Information

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

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

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

56
Introduction
58
59   Ecological communities tend to be spatially structured in response to environmental
60 gradients that are themselves organized in space, or to spatially contagious processes such as
61 growth, dispersion, and species interactions (Legendre & Legendre, 2012; Peres-Neto &
62 Legendre, 2010). Thus, disentangling the causes of spatial structure and identifying spatial
63 variability and different scales of organization in natural communities is a central question in
64 ecology (Legendre, 1993). Answering this question requires the construction of explanatory
65 variables based on spatial relationships among sites (Dray et al., 2006). One approach
66 extensively used to create spatial variables and/or control for spatial autocorrelation in residuals
67 is an eigenvector-based method, called Moran's eigenvector maps (MEMs, Dray et al., 2006).
68 This method creates spatial explanatory variables representing structure on a range of spatial
69 scales from the spatial relationships among sampling sites. These variables can be used for a
70 broad range of goals, from controlling for phylogenetic autocorrelation in ecological data (Diniz-
71 Filho et al., 2012) to searching for spatial structure in natural communities, even when
72 irregularly sampled (*e.g.* Bauman et al., 2016; Neves et al., 2015).

73   In many studies the response variables for which ecologists seek to find spatial structure
74 are community composition datasets containing either abundances or presence/absence
75 information (here, we focus on the latter). For community ecology studies, Redundancy Analysis
76 (RDA) is one of the most popular strategies due to its versatile framework, well-established
77 literature and abundant toolkits available for implementation (see Blanchet, Legendre, Bergeron,
78 & He, 2014; Borcard, Legendre, & Drapeau, 1992; Eisenlohr & Oliveira-Filho, 2015; Saiter,
79 Eisenlohr, Barbosa, Thomas, & Oliveira-Filho, 2015). The RDA algorithm searches for optimal
80 linear combinations (in the least-squares sense, see Legendre & Legendre, 2012) of the
81 explanatory variables that best explain the variation in the transformed community composition
82 data (Legendre & Gallagher, 2001; Borcard et al., 2011; Blanchet et al., 2014). The usual
83 approach then consists of establishing the global significance of the relationship between the
84 response matrix and all the explanatory variables, after which a subset of explanatory variables is
85 usually selected by stepwise procedure such as Forward Selection (FW, *sensu* Blanchet et al.,

86 2008). The most common approach uses two thresholds for variable selection: a significance
87 level α and the adjusted R² (see below and Blanchet et al., 2008 for details). This whole
88 framework will hereafter be called RDA/FW for brevity. A statistic related to the Akaike
89 Information Criteria (AIC, Akaike, 1973) has also been suggested for RDA model selection
90 (Godínez-Domínguez & Freire, 2003), but it has been shown to perform poorly and will not be
91 further explored here (Bauman, Drouet, Dray, et al., 2018).
92   However, methods based on least-squares such as RDA are unlikely to perform well
93 when applied to data that violate the assumption of constancy in the mean-variance relationship.
94 This assumption is usually violated by datasets containing many zeros including abundance
95 (count or semi-quantitative) and presence/absence (binary) data. Data transformation does not
96 always solve such problems (O'Hara & Kotze, 2010; Warton, 2018), although least-squares can
97 give reasonably robust tests of the significance of regression coefficients (Ives, 2015). In general,
98 algorithmic methods such as RDA do not take into account the statistical properties of the
99 response variable, such as the distribution of variances and how the response changes along
100 spatial/environmental gradients (Ferrier et al., 2007; Warton et al., 2012, 2015, 2018). More
101 recently, Generalized Linear Models (GLMs) have been proposed as an alternative model-based
102 approach to the analysis of presence/absence or count data (Wang et al., 2012; Warton et al.,
103 2015; Yee, 2006). The use of GLMs has long been established for univariate analyses and related
104 approaches for multivariate count data are now available (O'Hara & Kotze, 2010; Warton,
105 2018). The usual approach to selection of explanatory variables in this approach is Akaike's
106 Information Criterion (AIC: Akaike, 1973; Wagenmakers & Farrell, 2004). This framework will
107 hereafter be named GLM/AIC.
108   Here, we compared the performance of the RDA/FW and GLM/AIC approaches to
109 selecting spatial explanatory variables for community presence/absence data by measuring the
110 proportion of spatial patterns contained in simulated communities they could correctly retrieve.
111 There have been some studies of simulated multivariate count data (Warton et al., 2012), but
112 presence/absence data are particularly important in spatial studies because they are often the only
113 data that can be collected consistently over large spatial extents. We therefore compare the
114 performance of RDA/FW and GLM/AIC methods for the selection of MEM spatial variables
115 (including one special case, the asymmetric eigenvector map or AEM) from realistic simulated
116 presence/absence data. We used spatial variables as our predictors since we were interested in
117 discovering whether varying the spatial scales in which communities were structured would
118 affect model performance. We generated simulated data sets with predefined spatial structure
119 based on three real data sets, under two different ecological interpretations of presence/absence
120 data. First, we assumed that species are truly present at some sites and absent at others, and are
121 detected if present (simulated presence method, SPM). Alternatively, absences may represent
122 failure to detect species that are truly present. In this case, we simulated species abundances,
123 followed by a simulated sampling step to obtain presence/absence data (simulated abundances
124 method, SAM).
125
126 **Materials & Methods**
127
128 *Baseline Datasets*
129
130   We compared the two approaches to spatial variable selection using simulated community
131 data based on three real community composition datasets with a range of properties:

132   A) Presence/Absence of 110 marine benthic macroalgae species from a Rapid Assessment
133   Program for biodiversity of 42 sample sites spanning roughly 2000 km^2 at Ilha Grande
134   Bay, Rio de Janeiro, Brazil (tropical southwest Atlantic) (Carlos-Júnior et al., 2019,
135   permit number IBAMA/RJ:031/04);
136   B) Presence/Absence of 588 plant species from grassland covering 500 km^2 of Scotland's
137   coast. Data were collected from 3639 5 × 5 m quadrats from 94 sites. We used sites as
138   our sample units, treating species as present when they occurred in at least one quadrat at
139   a site, and absent otherwise (see Lewis et al., 2014 for more information);
140   C) Presence/Absence of 47 freshwater aquatic insect species collected from 30 sample sites
141   in five tributaries of the Guapiaçú River basin, Brazil which covers about 40 km² (Feijó-
142   Lima in prep, permit number INEA-RJ: 019-2014).
143   For each of the datasets we used the geographical coordinates (maps and sampling sites in
144 Supplemental Figure S1) to calculate spatial explanatory variables for regression (Fig. 1). We
145 chose MEMs as our spatial variables since they are commonly used to describe spatial structure
146 in ecological studies. Moreover, in contrast to coarser methods such as trend-surface analysis,
147 MEMs are a flexible method, capable of describing all spatial scales provided by the sampling
148 design (Borcard et al., 2011). They are also more flexible and powerful than the method of
149 principal coordinates of neighbor matrices (PCNMs, a special case of distance-based MEMs)
150 (Bauman, Drouet, Dray, et al., 2018; Bauman, Drouet, Fortin, et al., 2018; Borcard & Legendre,
151 2002; Dray et al., 2006). One needs two matrices to build the MEM variables for a given set of
152 site coordinates: matrix **B** describing the connectivity among the geographical sampling sites and
153 matrix **A** describing the weights of such connections. The Hadamard product of these two
154 matrices generates the spatial weighting matrix (matrix **W**), which is then doubly centred and
155 diagonalized, yielding eigenvectors to be used as spatial variables. For ecological studies, the
156 processes of interest are usually those generating positive autocorrelation, and it is therefore
157 common to use only MEMs associated with positive eigenvalues (as in this study). For studies in
158 which negative spatial autocorrelation is also of interest (*e.g.* where negative interactions such as
159 competitive exclusion, predation, etc are suspected), the eigenvectors associated with negative
160 eigenvalues can also be separately used (Bauman, Drouet, Dray, et al., 2018). We made
161 decisions about **B** and **A** for each dataset based on our ecological knowledge of the spatial
162 structure of these regions, since our goal was to simulate communities with ecologically sensible
163 spatial structures. Therefore, for dataset A we chose the minimum spanning tree (**B**) with
164 Euclidian linear distances as weights (**A**). Our decision was based on the shape of the bay and the
165 fact that the main water movements make the sampling sites geographically compartmentalised
166 in subregions where sites are likely to be minimally connected (Carlos-Júnior et al., 2019).
167 Similarly, spatial organisation in dataset B could be sensibly described in terms of Delaunay
168 triangulation (**B**) with Euclidian weights (**A**). Despite some degree of connectivity among all
169 sites, pairs of sites could be mostly associated not to their immediate neighbours but rather as a
170 function of their distances. This is due to cultural differences in land management. For example,
171 northern and western islands share cultural histories, which is reflected in species composition
172 (Lewis et al., 2014). Directional spatial processes in ecological data, such as those observed in
173 rivers, are well described by a special case of MEMs called asymmetric eigenvector maps
174 (AEMs, Blanchet, Legendre, & Borcard, 2008), which were used for constructing variables for
175 dataset C. In MEMs, larger eigenvalues are associated with broader-scale spatial structures while
176 smaller eigenvalues represent fine-scale spatial structures. This allowed us to control the spatial
177 scale of variation in community structure. Dataset A had 16 positive MEMs from 42 sites,

178   dataset B had 30, and dataset C had 12 AEMs with positive autocorrelation. For computation of
179   the MEMs for the three datasets we used the packages *adespatial* (version 0.3-7, Dray et al.,
180   2019) and *spdep* (version 0.7-4, Bivand & Piras, 2015; Bivand, Hauke, & Kossowski, 2013).
181
182   *Simulating communities with chosen spatial drivers*
183
184   We simulated realistic communities with known spatial structure, based on the three
185   datasets. We used spatial eigenvectors as explanatory variables. We varied the number of MEMs
186   with non-zero coefficients and created new binary (presence/absence) communities (with the
187   same number of sites and same expected number of species as the real ones) using two different
188   modelling scenarios. These simulated communities reflected the effect of those MEMs with non-
189   zero coefficients. By varying the number and ordering of the non-zero coefficients, we could
190   therefore control the spatial structure and scale of the simulated community data (see scheme in
191   Fig. 1 and Table 1).
192   In order to simulate new binary communities under the simulated presence method (SPM,
193   in which species are always detected if present), we first estimated a coefficient matrix $\mathbf{C}$ of size
194   (m variables + 1 (first) row with intercepts) $\times$ p species from each real data set. This was
195   achieved using the `manyglm` function with binomial errors in R package *mvabund* (version
196   3.11.9, Wang et al., 2012), with explanatory matrix $\mathbf{X}$ (n sites $\times$ m positive MEMs + an initial
197   column of 1's). The matrix $\mathbf{C}$ gives the effect of each explanatory variable on the logit-
198   transformed probabilities of presence. The *mvabund* package provides a GLM framework for
199   multivariate response data.
200   We then created new hypothetical scenarios by generating a new coefficient matrix $\mathbf{C}^*$,
201   of the same size as $\mathbf{C}$, whose elements c_{kj}^* are given by
202

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

203
204
205   where $\grave{F}_b$ is the empirical distribution function of c_{kj} (k=2, 3, ..., m+1, j= 1, 2, ..., p)
206   (Evans et al., 2000), and the b_{kj}^* are sampled with replacement. The set K defines to which rows
207   of $\mathbf{C}^*$ the non-zero coefficients were allocated: we studied 14 such sets (see below and Table 1 a-
208   c). In other words, we used the originally-estimated intercepts in each simulation (first row of
209   eqn 1), and drew those coefficients assigned to non-zero values (second row of eqn 1) from the
210   empirical distribution of all the originally-estimated explanatory variable coefficients. We
211   sampled the values of the non-zero coefficients from the empirical distribution in order to
212   simulate plausible but not fixed spatial structures. Table 1 depicts for each dataset how the non-
213   zero coefficients were assigned for each dataset and simulation scenario (see below).
214   We then calculated predicted probabilities of presence $\grave{p}_{ij}$ for the jth species at the ith site.
215   Given the matrix $\grave{\mathbf{Y}} = \mathbf{X}\mathbf{C}^*$ (n sites $\times$ p species) of predicted logit probabilities of presence, the
216   predicted probability of presence is
217

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

218
219

220   The simulated presence/absence value for species *j* at site *i* was sampled from a Bernoulli
221   distribution with success probability $\grave{p}_{ij}$. The result is a community matrix with the same number
222   of sites and the same expected number of species as the real community, and with realistic
223   coefficients for spatial eigenvectors. As in the maximum likelihood estimation done by
224   `manyglm` (Wang et al., 2012), species and sites were assumed conditionally independent when
225   generating simulated presence/absence data, given the values of the explanatory variables. Our
226   simulated communities correspond to the simple case in which presence/absence patterns are
227   affected by environmental variables but not interspecific interactions. Nevertheless, interspecific
228   interactions could be well relevant to real world systems and other models (Godsoe & Harmon
229   2012; Anderson, 2017).
230   Since GLMs are specified correctly for presence/absence data generated this way, we
231   would expect them to perform well. We therefore devised a second ecologically meaningful
232   simulation method in which absences arise from the sampling protocol, called the simulated
233   abundance method (SAM). The two simulation methods differ in whether they assume we have
234   true absences or sampling-related absences. Note that it is not possible to simulate binary data
235   directly using RDA, because RDA does not generate predicted probabilities of presence. Instead,
236   we treated $\grave{Y}$ as log expected abundances and exponentiated each element to get expected
237   abundances λ. Then we calculated the probability of detecting the species under Poisson
238   sampling (*i.e.* the probability of drawing a value of at least 1 from a Poisson distribution with
239   parameter λ), which is

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

243   Finally, we generated a Bernoulli random variable with success probability $\grave{p}_{ij}$ to produce
244   a simulated presence-absence observation. Both GLM and RDA are mis-specified for data
245   generated in this way. Codes for both the SPM and SAM simulation frameworks and all the
246   datasets used in our simulations are available as supplemental information (Data S1, S2 and S3).
247   We compared GLM and RDA variable selection under up to 14 different scenarios,
248   differing in the number of non-zero coefficients (*nVar*) and whether these coefficients were
249   associated with fine or broad spatial scales. We simulated up to six different choices of the
250   number of MEM variables creating the spatial structure in the data (*i.e.* having non-zero
251   coefficients): none, approximately one sixth, approximately one third, approximately half,
252   approximately three-quarters, and all (Table 1 a-c, rows). We also simulated three different
253   spatial scales of the patterns. As mentioned above, MEMs associated with larger eigenvalues
254   represent broader spatial scales. We ordered the MEMs in descending order of eigenvalues and
255   arranged the non-zero coefficients within matrix $\mathbf{C}^*$ in three different ways (Table I a-c,
256   columns): only broad-scale MEMs with non-zero coefficients (scaling 1); only fine-scale MEMs
257   with non-zero coefficients (scaling 2); half broad-scale, half fine-scale (scaling 3). Because not
258   every combination of number of non-zero coefficients and spatial scaling is possible (*e.g.* it is
259   not possible to assign one non-zero coefficient in scaling 3), there were 14 possible combinations
260   overall for each dataset (Table 1). The main steps of the simulation scheme are summarized in
261   Fig. 1.

263   *RDA and GLM*
264   We used the default RDA function from the R package *vegan* (version 2.4-1, Oksanen et
265   al., 2016), with simulated community composition as the response variable, and MEMs

266   associated with positive eigenvalues generated from geographical coordinates of the sample sites
267   as explanatory variables. In order to perform a transformation-based RDA (Borcard et al., 2011;
268   Blanchet et al., 2014) we used the Ochiai coefficient, which is the Hellinger transformation
269   analogue for binary data, as recommended by Legendre & Gallagher (2001) and Borcard et al.
270   (2011).
271   Binomial GLMs were fitted to the same data using the `manyglm` function in R package
272   *mvabund* (Wang et al., 2012). We fitted our models using a logistic regression (logit link
273   function for binomial response), with species compositional data as the multivariate response
274   variable and MEMs as predictors. No interaction terms were included, following common
275   practice in spatial modelling of community data.
276
277   *Comparing model selection between RDA and GLM frameworks*
278
279   We compared the results of model selection between the approach usually taken in the
280   RDA and a somewhat-similar approach for GLMs. For RDA, we used the forward selection with
281   double stopping criterion following Blanchet et al. (2008), beginning with a global test of
282   significance (model with all spatial predictors) and carrying on with the variable selection if the
283   global model was significant. The forward selection itself consists of a stepwise procedure
284   including in the model the variable contributing the most to the adjusted R^2. The procedure stops
285   either when the next variable with the highest contribution is not significant (first stopping
286   criterion) or causes the adjusted R^2 to be bigger than that of the global model (*i.e.* containing all
287   variables; second criterion). This is implemented in the function `ordiR2step` in the *vegan*
288   package (Oksanen et al., 2016). For GLM, we used forward selection with a stopping rule based
289   on minimum Akaike Information Criterion (AIC) (Akaike, 1973; Wagenmakers & Farrell,
290   2004). The selection procedure started from a model with intercept only and added one
291   explanatory variable at a time, until no further improvement in the sum of AIC over each of the
292   response variables was possible. We used this approach because the usually large number of
293   MEMs makes it difficult to compare the AIC sum over all possible GLMs.
294   The performance of each method on simulated data was mainly assessed by two criteria.
295   First, we assessed how many MEMs with zero coefficients were incorrectly included in the final
296   model. Second, we assessed how many MEMs with non-zero coefficients were incorrectly
297   excluded from the final model. Also, we assessed overall accuracy (score) as the percentage of
298   MEMs whose inclusion/exclusion status was correct. The goals of ecological studies are usually
299   not directly related to the inclusion/exclusion of individual MEM variables, but instead to
300   identify spatial pattern, represented by a linear combination of MEMs. However, since the
301   MEMs form a basis for the space spanned by the transformed spatial weighting matrix, such a
302   linear combination is unique (Fraleigh & Beauregard, 1995, pages 197-198). Furthermore, the
303   MEMs are orthogonal, so that each represents a qualitatively distinct aspect of spatial pattern.
304   Therefore, if an individual MEM is incorrectly included or excluded, the estimated spatial
305   pattern is qualitatively wrong.
306   We further explored the ability of each method to capture spatial pattern using a graphical
307   approach (Article S1). For each real dataset and each method, we haphazardly picked one
308   simulated data set. We plotted the MEM decompositions of both the true and estimated spatial
309   patterns. We chose the scenarios in which each method had the worst performance in terms of
310   correctly including/excluding variables, in order to determine whether in such cases, overall
311   spatial pattern would still be captured.

312   Finally, we calculated how much of the variation in response variables was explained by
313 each method using the adjusted R^2 for the linear model in RDA and its analogue for GLMs, the
314 D-value (Tjur, 2009). These two values cannot be directly compared since they are not exactly
315 equivalent, but their results could yield interesting insights and are made available as
316 supplemental information (see table results in Data S4).
317   For each of the combinations of conditions in Table I, 1000 simulated data sets were
318 generated under each of SPM and SAM. For each simulated data set, spatial explanatory
319 variables were selected using both GLM/AIC and RDA/FW.
320

Results

322

323   Overall, GLM/AIC outperformed RDA/FW in selecting spatial explanatory variables
324 when data were simulated under either SPM or SAM in all three scaling patterns (Fig.2).
325   In general, GLM/AIC had fairly predictable performance: it performed nearly perfectly
326 when few or none of the available variables had non-zero true coefficients (*i.e.* $nVar = 0$, $m/6$,
327 $m/3$ or $m/2$), but was less accurate when many or all the variables had non-zero true coefficients (
328 $nVar = 3m/4$ or $nVar = m$) (blue lines in Fig.2 A-E). There was also some discernible pattern in
329 RDA/FW's scores: it performed best at $nVar = 0$ and $nVar = m$, with intermediate values showing
330 a considerable decrease in selection success. The loss of accuracy for intermediate values of
331 $nVar$ (drop in red lines across different $nVar$ values in Fig.2 A-E) varied substantially among
332 datasets, making general inferences about results more difficult. There was little difference
333 between the results from the SPM and SAM simulations (Fig. 2B, D, F).
334   It is also noteworthy that when the model had a smaller number of variables to select
335 from (River dataset C with 12 MEMs), scores in GLM/AIC were higher, with virtually no
336 incorrect inclusion of variables, and incorrect exclusion of variables occurring on average in only
337 approximately 6% of all 14000 simulations over the whole set of replicates (Figure 3E). Under
338 the same conditions, RDA/FW's rate of success was approximately 81%, incorrectly including
339 variables at a rate of 18% (incorrect exclusions represented less than 1%) as depicted in Figure
340 3E.
341   Under both the SPM and SAM simulation methods, GLM/AIC differed substantially
342 from the RDA/FW framework in regard to the type of errors it most often produced. GLM/AIC
343 had virtually no incorrect inclusion of variables (Fig. 3, blue). However, when $nVar = 3m/4$ or
344 $nVar = m$ some variables that should be included in the final model were left out. Nevertheless,
345 GLM/AIC never had less than around 90% accuracy over all three datasets (overall mean = 96 ±
346 1.3% against 71 ± 1.7% from RDA/FW). On the other hand, RDA/FW often included more
347 variables than it should in the model (Fig. 3, red). Such errors especially occurred when
348 $0 < nVar \leq 3m/4$. Under some conditions, up to one third of the variables selected by RDA/FW
349 had zero coefficients.
350   MEM decompositions of true and estimated spatial structure provided a visual
351 assessment of the extent of the misspecification yielded by each method (Article S1). In all three
352 datasets, the worst performance of GLM/AIC corresponded to those models in which it should

353 have included all MEM variables (Fig. 2). Those scenarios represented communities structured at
354 all spatial scales (broad, intermediate and fine). Despite incorrectly excluding several individual
355 variables, GLM/AIC was capable of selecting subsets of variables that corresponded to all those
356 scaling categories (Article S1.2-S1.7). In contrast, RDA/FW performed worse when there were
357 few spatial variables (nVar = 5, nVar = 10 and nVar=2 for datasets A, B and C, respectively).
358 Under those conditions, incorrect inclusion of variables also resulted in the inclusion of incorrect
359 spatial scales. For example, in one simulation from dataset A (Article S1.8) the true spatial
360 structure contained only five MEMs describing finer spatial scale patterns (scaling 2 = MEMs
361 12-16). However, the final model selected by RDA/FW included 13 variables describing both
362 broad (MEMs 1-6) and intermediate spatial scales (MEMs 9, 11), along with the correct ones
363 (Article S1.9). Similar results were found in all three datasets (Article S1.10-S1.13). Moreover,
364 these incorrect inclusions of individual variables by RDA/FW resulted in the inclusion of MEM
365 variables associated to eigenvalues substantially different from the correct ones, representing
366 spatial scales much larger than those actually present in the data (Article S1.14). For matters of
367 space, we only plotted one failure example from each dataset for both GLM/AIC and RDA/FW.
368 However, the correct spatial structures within simulated communities and those structures
369 retrieved by both methods in all our simulations scenarios are available as supplemental data
370 (Data S5).
371   Under SPM simulations, the scale of spatial pattern (fine, broad or mixed: scaling 1, 2
372 and 3, respectively) had negligible effect on GLM/AIC performance (Fig.4A, C, E). A slight
373 difference in variable selection scores between scaling 1 to 2 and 3 was only found in one
374 modelling condition (Fig. 4, $nVar = 3^m/4$). On the other hand, scaling often affected the
375 performance of RDA/ FW, although there was no obvious general pattern across different
376 conditions and datasets (Fig.4A, C, E). Under SAM simulations, both frameworks performed
377 similarly to what was observed under SPM (Fig.4B, D, F).
378

Discussion

380

381   Here, we showed that a GLM/AIC-based method for finding spatial structure in
382 communities outperformed an RDA/FW-based method, for presence-absence data simulated
383 under two different ecologically plausible scenarios about how absences arise. We based our
384 simulated datasets on real datasets from marine, terrestrial and freshwater data. Notably,
385 differences in assumptions about how absences arise made little difference to performance. This
386 might be due to the structure of our community presence/absence datasets, which (like most
387 ecological datasets) had many rare species and, therefore, many expected abundances close to
388 zero. In such cases, the relationship between the community data and explanatory variables could
389 be approximated by a binomial GLM with a logit link function, even if this was not the correct
390 model (as in the SAM simulations). We therefore focus below on general patterns that apply
391 equally to both assumptions about absences, rather than on the details of these assumptions.
392   In selecting spatial explanatory variables, GLM followed by AIC-based model selection
393 (GLM/AIC) performed better than the widely-used approach of RDA followed by forward
394 selection (RDA/FW). Not only did GLM/AIC have better performance overall, but its

395   performance varied little between simulation conditions (Fig. 2). In contrast, RDA/FW
396   performed unpredictably, but often retained too many explanatory variables (Fig. 3).
397   The problems arising from data with non-Gaussian error distributions, such as classic
398   community presence and absence data, in a linear modelling framework are not new to science (
399   Legendre & Gallagher, 2001; McCullagh & Nelder, 1989; Wolda, 1981). Classical linear models
400   such as RDA (Legendre & Anderson, 1999; Legendre & Legendre, 2012) make assumptions
401   regarding constancy of variance in the data (ter Braak & Prentice, 1988) that cannot be true for
402   presence-absence data, even after data transformation (O'Hara & Kotze, 2010; Warton, 2018;
403   Warton et al., 2012). The problem may be negligible in some hypothesis testing situations (Ives,
404   2015). Regardless, incorrectly assuming linearity (and constant variance) may lead to serious
405   problems. Unfortunately, RDA is an algorithmic method that makes implicit decisions about the
406   distribution of variances (ter Braak & Prentice, 1988; Warton et al., 2012) and does not provide
407   the flexibility to separate systematic variation from random variation in the way that statistical
408   models such as GLMs do (Warton et al., 2015; and see O'Neil & Schutt, 2013 for differences
409   between algorithms and statistical models). New frameworks, such as using GLMs with
410   spatially-structured random effects (followed by variation partitioning to find environmental and
411   spatial components) have also been specifically proposed as a model-based alternative to MEMs
412   (Ovaskainen et al., 2017). Despite recent advances showing that better estimates could be
413   obtained by using sensible selection procedures, manipulating the data appropriately and/or by
414   splitting the analysis of the response data over shorter spatial/environmental gradients (Bauman,
415   Drouet, Dray, et al., 2018; Ives, 2015; Vieira et al., 2019), employing statistical models that
416   match the distribution of the response data is better practice in most cases (Ferrier et al., 2007;
417   Warton, 2018; Warton et al., 2015).
418   Another relevant aspect of the general performances of the two methods concerns the
419   peaks of performance in detecting spatial structure. The scores in the GLM/AIC framework were
420   close to ideal across datasets when the number of variables that should be selected was none or
421   was small relative to the number of variables available. The performance only decayed when
422   many or all of the available variables should have been retained in the final model. Thus, if a
423   few variables are responsible for most of the spatial structure in community composition,
424   GLM/AIC will usually outperform RDA/FW (Fig. 2). Considering that the majority of effects
425   could be derived from a small number of causes (Sullivan, 2019) in many biological systems,
426   GLM/AIC could presumably perform well on many real systems. On the other hand, RDA/FW
427   worked best precisely in situations thought unlikely in real systems, when no spatial structure is
428   present among communities (where GLM/AIC also performed equally well), or when
429   composition is structured at all possible spatial scales (i.e., $nVar = 0$ and $nVar = m$, respectively).
430   Moreover, when the model had a small number of variables to select from (River dataset, Fig.3E-
431   F), performance of RDA/FW was very variable (Fig. 3E-F).
432   The two approaches also differed in the ways they failed. GLM/more often included too
433   few variables, while RDA/FW more often included too many. This was consistent among all
434   three datasets under SPM and SAM simulations (Fig. 3) and is in contrast with results from
435   previous studies where GLMs produced higher Type I error rates compared to a linear model
436   (Ives, 2015). For beta diversity studies, where the aim is to identify the most important variables
437   associated with differences in community composition, leaving out a few variables that affect
438   composition is better, in our opinion, than including many variables whose effects are not
439   important. On the contrary, in other scenarios such as when one tries to select pivotal attributes
440   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 $\lfloor 3^m/4 \rfloor$). 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*

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

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

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

492 Bauman, D., Drouet, T., Dray, S., & Vleminckx, J. (2018). Disentangling good from bad
493 practices in the selection of spatial or phylogenetic eigenvectors. *Ecography*, *41*(10), 1638–
494 1649. https://doi.org/10.1111/ecog.03380

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

498 Bauman, D., Raspé, O., Meerts, P., Degreef, J., Ilunga Muledi, J., & Drouet, T. (2016).
499 Multiscale assemblage of an ectomycorrhizal fungal community: the influence of host
500 functional traits and soil properties in a 10-ha miombo forest. *FEMS Microbiology Ecology*,
501 *92*(10). https://doi.org/10.1093/femsec/fiw151

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

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

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

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

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

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

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

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

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

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

532   Carlos-Júnior, L. A., Spencer, M., Neves, D. M., Moulton, T. P., Pires, D. de O., e Castro, C. B.,
533   Ventura, C. R. R., Ferreira, C. E. L., Serejo, C. S., Oigman-Pszczol, S., Casares, F. A.,
534   Mantelatto, M. C., & Creed, J. C. (2019). Rarity and beta diversity assessment as tools for
535   guiding conservation strategies in marine tropical subtidal communities. *Diversity and*
536   *Distributions*. https://doi.org/10.1111/ddi.12896
537   Diniz-Filho, J. A. F., Bini, L. M., Rangel, T. F., Morales-Castilla, I., Olalla-Tárraga, M. Á.,
538   Rodríguez, M. Á., & Hawkins, B. A. (2012). On the selection of phylogenetic eigenvectors
539   for ecological analyses. *Ecography*, *35*(3), 239–249. https://doi.org/10.1111/j.1600-
540   0587.2011.06949.x
541   Dray, S., Bauman, D., Blanchet, F. G., Borcard, D., Clappe, S., Guenard, G., Jombart, T.,
542   Larocque, G., Legendre, P., Madi, N., & Wagner, H. H. (2019). *adespatial: Multivariate*
543   *Multiscale Spatial Analysis* (0.3-7). https://cran.r-project.org/package=adespatial
544   Dray, S., Legendre, P., & Peres-Neto, P. R. (2006). Spatial modelling: a comprehensive
545   framework for principal coordinate analysis of neighbour matrices (PCNM). *Ecological*
546   *Modelling*, *196*(3–4), 483–493. https://doi.org/10.1016/j.ecolmodel.2006.02.015
547   Eisenlohr, P. V., & Oliveira-Filho, A. T. de. (2015). Revisiting patterns of tree species
548   composition and their driving forces in the Atlantic Forests of Southeastern Brazil.
549   *Biotropica*, *47*(6), 689–701. https://doi.org/10.1111/btp.12254
550   Evans, M., Hastings, N., & Peacock, B. (2000). Statistical Distributions. In *New York* (Vol. 2,
551   Issue 4). Willey. https://doi.org/10.1002/9780470627242
552   Ferrier, S., Manion, G., Elith, J., & Richardson, K. (2007). Using generalized dissimilarity
553   modelling to analyse and predict patterns of beta diversity in regional biodiversity
554   assessment. *Diversity and Distributions*, *13*(3), 252–264. https://doi.org/10.1111/j.1472-
555   4642.2007.00341.x
556   Fitzpatrick, M. C., Sanders, N. J., Normand, S., Svenning, J. C., Ferrier, S., Gove, A. D., &
557   Dunn, R. R. (2013). Environmental and historical imprints on beta diversity: Insights from
558   variation in rates of species turnover along gradients. *Proceedings of the Royal Society B:*
559   *Biological Sciences*, *280*(1768). https://doi.org/10.1098/rspb.2013.1201
560   Fraleigh, J., & Beauregard, R. (1995). *Linear algebra* (3rd ed.). Addison Wesley.
561   Godínez-Domínguez, E., & Freire, J. (2003). Information-theoretic approach for selection of
562   spatial and temporal models of community organization. *Marine Ecology Progress Series*,
563   *253*, 17–24.
564   Godsoe, W., & Harmon, L. J. (2012). How do species interactions affect species distribution
565   models? *Ecography*, *35*(9), 811–820. https://doi.org/10.1111/j.1600-0587.2011.07103.x
566   Ives, A. R. (2015). For testing the significance of regression coefficients, go ahead and log-
567   transform count data. *Methods in Ecology and Evolution*, *6*(7), 828–835.
568   https://doi.org/10.1111/2041-210X.12386
569   Legendre, P., & Anderson, M. J. (1999). Distance-based redundancy analysis: Testing
570   multispecies responses in multifactorial ecological experiments. *Ecological Monographs*,
571   *69*(1), 1–24.
572   Legendre, P. (1993). Spatial Autocorrelation : Trouble or New Paradigm ? *Ecology*, *74*(6), 1659–
573   1673.
574   Legendre, P., & Gallagher, E. D. (2001). Ecologically meaningful transformations for ordination
575   of species data. *Oecologia*, *129*(2), 271–280. https://doi.org/10.1007/s004420100716
576   Legendre, P., & Legendre, L. (2012). *Numerical Ecology* (Third Engl). Elsevier Ltd.
577   Lewis, R. J., Pakeman, R. J., & Marrs, R. H. (2014). Identifying the multi-scale spatial structure

578   of plant community determinants of an important national resource. *Journal of Vegetation*
579   *Science*, 25(1), 184–197.
580 McCullagh, P., & Nelder, J. A. (1989). *Generalized Linear Models* (2° Edition). Chapman and
581   Hall/CRC.
582 Miller, A. (1990). *Subset Selection in Regression*. Chapman and Hall.
583 Neves, D. M., Dexter, K. G., Pennington, R. T., Bueno, M. L., & Oliveira Filho, A. T. (2015).
584   Environmental and historical controls of floristic composition across the South American
585   Dry Diagonal. *Journal of Biogeography*, 42(8), 1566–1576.
586   https://doi.org/10.1111/jbi.12529
587 O'Hara, R. B., & Kotze, D. J. (2010). Do not log-transform count data. *Methods in Ecology and*
588   *Evolution*, 1(2), 118–122. https://doi.org/10.1111/j.2041-210X.2010.00021.x
589 O'Neil, C., & Schutt, R. (2013). *Doing Data Science* (First). O'Reilly.
590 Oksanen, A. J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., Mcglinn, D., Minchin, P.
591   R., Hara, R. B. O., Simpson, G. L., Solymos, P., Stevens, M. H. H., & Szoecs, E. (2016).
592   *Package ' vegan .'*
593 Ovaskainen, O., Tikhonov, G., Norberg, A., Guillaume Blanchet, F., Duan, L., Dunson, D.,
594   Roslin, T., & Abrego, N. (2017). How to make more out of community data? A conceptual
595   framework and its implementation as models and software. *Ecology Letters*, 20(5), 561–
596   576. https://doi.org/10.1111/ele.12757
597 Peres-Neto, P. R., & Legendre, P. (2010). Estimating and controlling for spatial structure in the
598   study of ecological communities. *Global Ecology and Biogeography*, 19(2), 174–184.
599   https://doi.org/10.1111/j.1466-8238.2009.00506.x
600 Saiter, F. Z., Eisenlohr, P. V., Barbosa, M. R. V., Thomas, W. W., & Oliveira-Filho, A. T. de.
601   (2015). From evergreen to deciduous tropical forests: how energy–water balance,
602   temperature, and space influence the tree species composition in a high diversity region.
603   *Plant Ecology & Diversity*, 9(October), 1–10.
604   https://doi.org/10.1080/17550874.2015.1075623
605 Sullivan, W. (2019). Rockets, gauges, and pendulums: applying engineering principles to cell
606   biology. *Molecular Biology of the Cell*, 30(14), 1635–1640.
607   https://doi.org/10.1091/mbc.E19-02-0100
608 Ter Braak, C. J. F., & Prentice, I. C. (1988). A Theory of Gradient Analysis. *Advances in*
609   *Ecological Research*, 18(C), 271–317. https://doi.org/10.1016/S0065-2504(08)60183-X
610 Tjur, T. (2009). Coefficients of Determination in Logistic Regression Models—A New Proposal:
611   The Coefficient of Discrimination. *American Statistician*, 63(4), 366–372.
612   https://doi.org/10.1198/tast.2009.08210
613 Vieira, D. C., Brustolin, M. C., Ferreira, F. C., & Fonseca, G. (2019). segRDA: An R package
614   for performing piecewise redundancy analysis. *Methods in Ecology and Evolution*, 1(1),
615   2041-210X.13300. https://doi.org/10.1111/2041-210X.13300
616 Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using Akaike weights.
617   *Psychonomic Bulletin & Review*, 11(1), 192–196. https://doi.org/10.3758/BF03206482
618 Wang, Y., Naumann, U., Wright, S. T., & Warton, D. I. (2012). Mvabund- an R package for
619   model-based analysis of multivariate abundance data. *Methods in Ecology and Evolution*,
620   3(3), 471–474. https://doi.org/10.1111/j.2041-210X.2012.00190.x
621 Warton, D. I. (2018). Why you cannot transform your way out of trouble for small counts.
622   *Biometrics*, 74(1), 362–368. https://doi.org/10.1111/biom.12728
623 Warton, D. I., Foster, S. D., De'ath, G., Stoklosa, J., & Dunstan, P. K. (2015). Model-based

624   thinking for community ecology. *Plant Ecology*, *216*(5), 669–682.
625   https://doi.org/10.1007/s11258-014-0366-3
626   Warton, D. I., Lyonsy, M., Stoklosa, J., & Ivesz, A. R. (2016). Three points to consider when
627   choosing a LM or GLM test for count data. *Methods in Ecology and Evolution*, *7*, 882–890.
628   https://doi.org/10.1111/2041-210X.12552
629   Warton, D. I., Wright, S. T., & Wang, Y. (2012). Distance-based multivariate analyses confound
630   location and dispersion effects. *Methods in Ecology and Evolution*, *3*(1), 89–101.
631   https://doi.org/10.1111/j.2041-210X.2011.00127.x
632   Wolda, H. (1981). Similarity indices, sample size and diversity. *Oecologia*, *50*(3), 296–302.
633   https://doi.org/10.1007/BF00344966
634   Yee, T. W. (2006). Constrained additive ordination. *Ecology*, *87*(1), 203–213.
635   http://www.ncbi.nlm.nih.gov/pubmed/16634311
636

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).

Variable Selection: correct inclusion or exclusion of MEMs
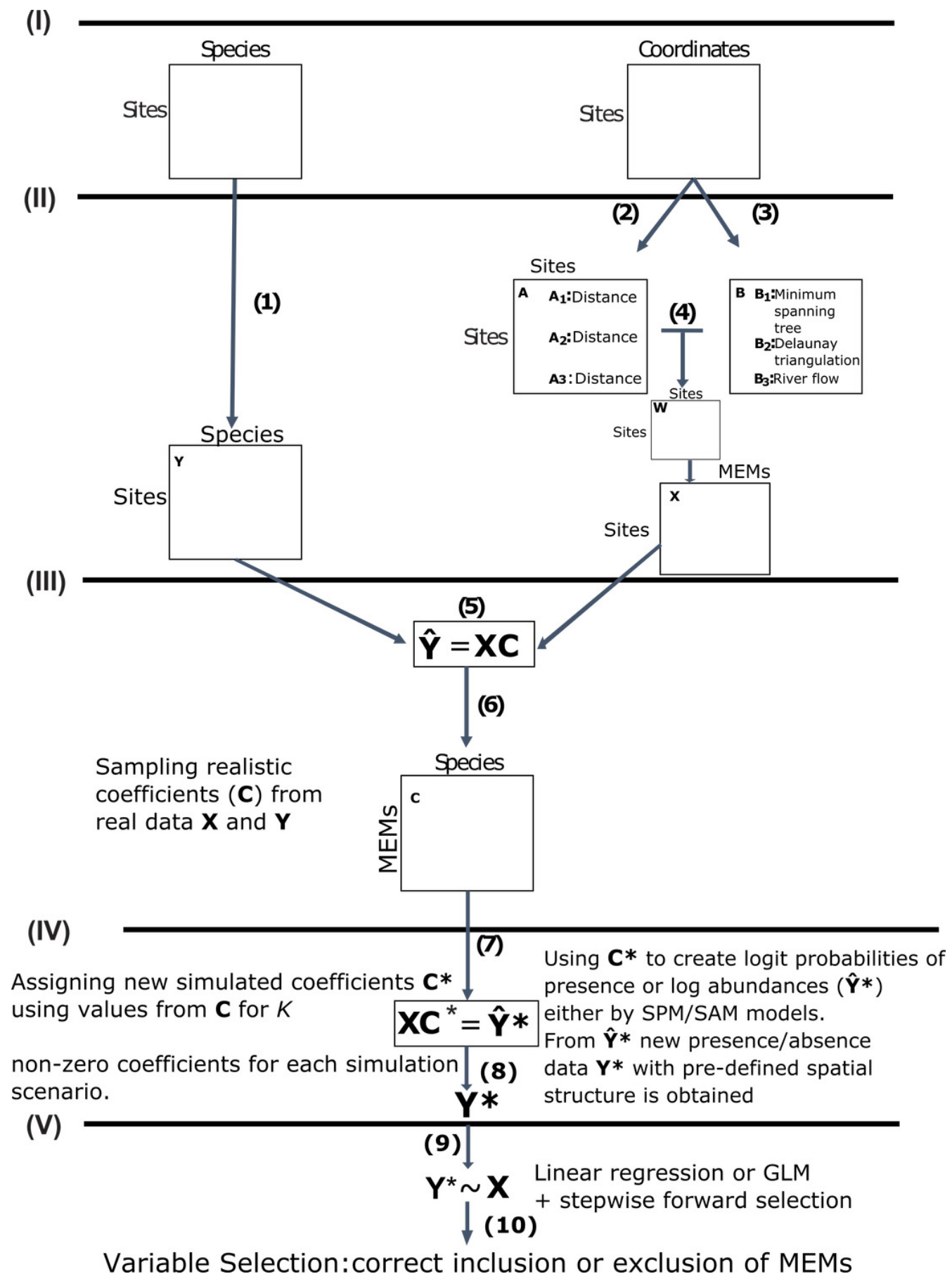
Figure 2

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).

anual to be reviewed
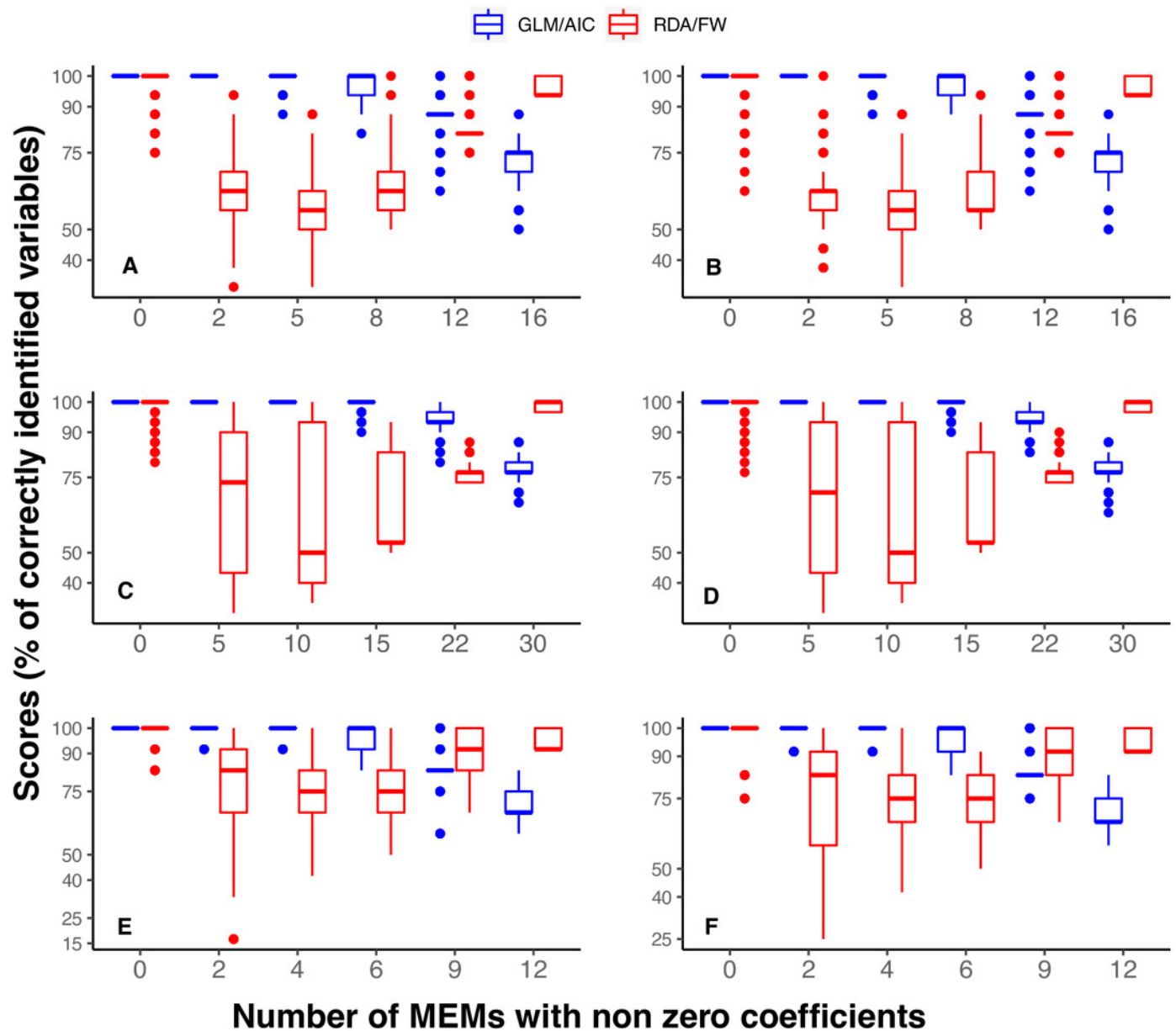
Figure 3

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.

Frequency (proportion)
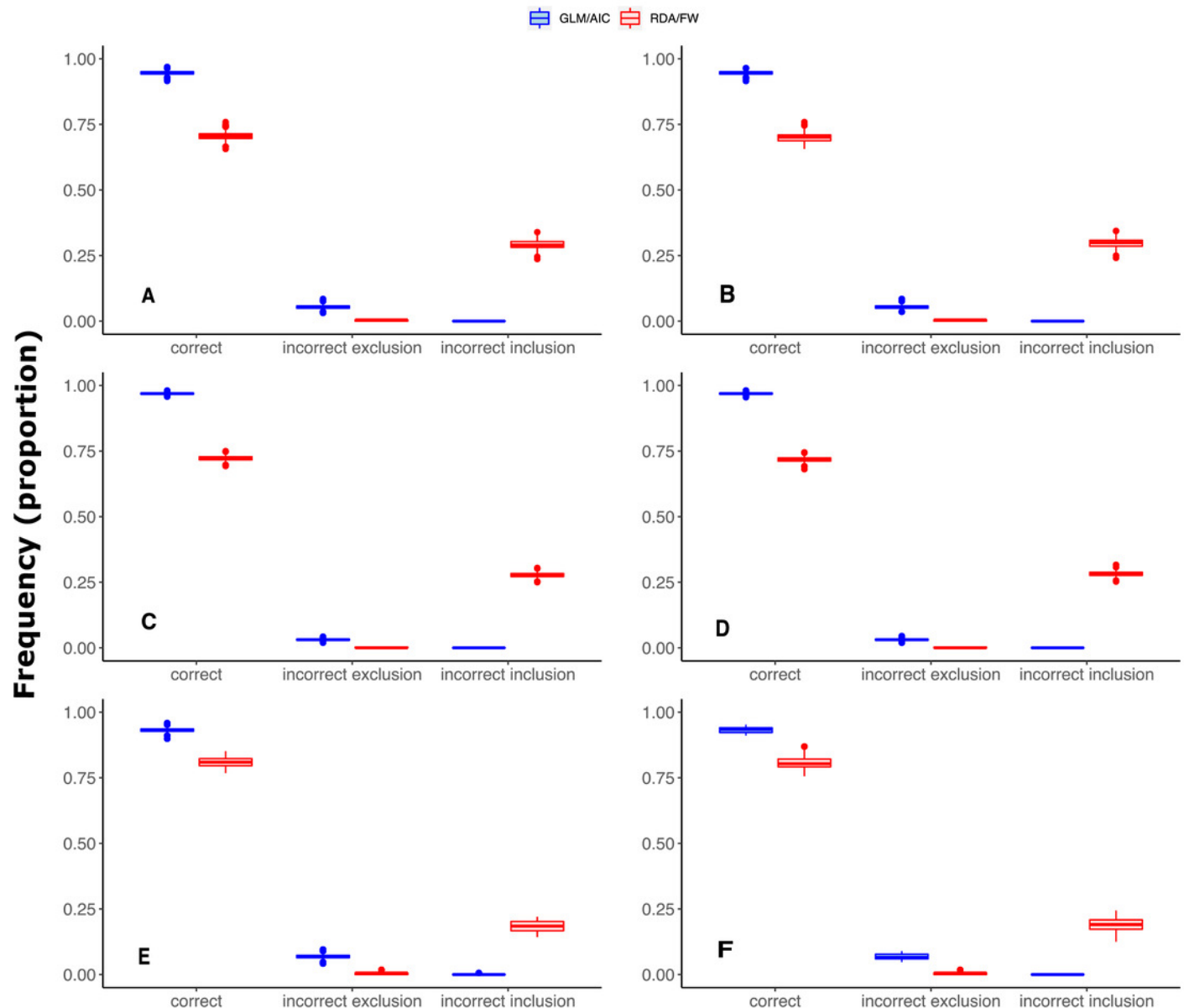
Figure 4

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).