



HAL
open science

Structural bias in aggregated species-level variables driven by repeated species co-occurrences: a pervasive problem in community and assemblage data

Bradford A Hawkins, Boris Leroy, Miguel Á. Rodríguez, Alexander Singer, Bruno Vilela, Fabricio Villalobos, Xiangping Wang, David Zelený

► To cite this version:

Bradford A Hawkins, Boris Leroy, Miguel Á. Rodríguez, Alexander Singer, Bruno Vilela, et al.. Structural bias in aggregated species-level variables driven by repeated species co-occurrences: a pervasive problem in community and assemblage data. *Journal of Biogeography*, Wiley, 2017, 44 (6), pp.1199-1211. 10.1111/jbi.12953 . hal-03334333

HAL Id: hal-03334333

<https://hal.archives-ouvertes.fr/hal-03334333>

Submitted on 3 Sep 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

1 Strapline: Special Paper

2

3 **Structural bias in aggregated species-level variables driven by**
4 **repeated species co-occurrences: a pervasive problem in community**
5 **and assemblage data**

6 Bradford A. Hawkins^{1*}, Boris Leroy², Miguel Á. Rodríguez³, Alexander Singer^{4,5,6}, Bruno
7 Vilela^{3,7}, Fabricio Villalobos^{7,8}, Xiangping Wang⁹ and David Zelený¹⁰

8 *¹Department of Ecology and Evolutionary Biology, University of California, Irvine,*
9 *CA 92697, USA*

10 *²UMR 7208 BOREA, Muséum National d'Histoire Naturelle, Sorbonne Universités, Université*
11 *Pierre et Marie Curie, Université de Caen, CNRS, IRD, Université des Antilles, Paris, France,*

12 *³Forest Ecology & Restoration Group, Department of Life Sciences, Universidad de Alcalá,*
13 *Alcalá de Henares, Madrid, Spain*

14 *⁴German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig,*
15 *Deutscher Platz 5e, 04103 Leipzig, Germany*

16 *⁵Department of Ecological Modelling, Helmholtz Centre for Environmental Research –*
17 *UFZ, Permoserstraße 15, 04318 Leipzig, Germany*

18 *⁶Swedish Species Information Centre, Swedish University of Agricultural Sciences, Box*
19 *7007, 750 07 Uppsala, Sweden*

20 *⁷Departamento de Ecologia, ICB, Universidade Federal de Goiás, Goiânia, Goiás, Brasil*

21 *⁸Red de Biología Evolutiva, Instituto de Ecología, A.C., Carretera antigua a Coatepec 351, El*
22 *Haya, 91070 Xalapa, Veracruz, Mexico*

23 ⁹*College of Forestry, Beijing Forestry University, 35 East Qinghua Road, Haidian, Beijing*
24 *100083, China*

25 ¹⁰*Institute of Ecology and Evolutionary Biology, National Taiwan University, Roosevelt Rd. 1,*
26 *10617 Taipei, Taiwan*

27

28

29

30

31

32

33 *Correspondence:
34 Bradford A. Hawkins
35 Department of Ecology & Evolutionary Biology
36 University of California, Irvine
37 Irvine, CA 92697 USA
38 E-mail: bhawkins@uci.edu

39

40

41

42

43

44 Word counts: abstract: 296; main text + references: 7536; 2 tables (1 page); 5 figs (1 colour; 2-3
45 journal pages); 1 online appendix

46

47 **ABSTRACT**

48 **Aim** Species attributes are often used to explain diversity patterns across
49 assemblages/communities. However, repeated species co-occurrences can generate spatial pattern
50 and strong statistical relationships between aggregated attributes and richness in the absence of
51 biological information. Our aim is to increase awareness of this problem.

52 **Location** North America

53 **Methods** We generated empirical species richness patterns using two data structures: (i) birds
54 gridded from range maps and (ii) tree communities from the US Forest Service Forest Inventory
55 and Analysis. We analysed richness using linear regression, regression trees, generalized additive
56 models, geographically weighted regression and simultaneous autoregression, with ‘random
57 intrinsic variables’ as predictors generated by assigning random numbers to species and
58 calculating averages in assemblages. We then generated simulations in which species with
59 cohesive or patchy distributions are placed with respect to the North American temperature
60 gradient with or without a broad-scale richness gradient. Random intrinsic variables are again
61 used as predictors of richness. Finally, we analysed one simulated scenario with random intrinsic
62 variables as both response and predictor variables.

63 **Results** The models of bird and tree richness often explained moderate to large proportions of the
64 variance. Regression trees, geographically weighted regression and simultaneous autoregression
65 were very sensitive to the problem; generalized additive models were moderately affected, as was
66 multiple regression to a lesser extent. In the virtual data, the variance explained increased with
67 increasing species co-occurrences, but neither range cohesion, a richness gradient nor spatial

68 autocorrelation in predictors had major impacts. The problem persisted when the response
69 variable was also a random intrinsic variable.

70 **Main Conclusions** Repeated species co-occurrences can generate strong spurious relationships
71 between richness and aggregated species attributes. It is important to realize that models utilizing
72 assemblage variables aggregated from species-level values, as well as maps illustrating their
73 spatial patterns, cannot be taken at face value.

74
75 Key words: **community structure, community weighted means, geographical ecology,**
76 **intrinsic variables, spatial analysis, species composition, species co-occurrence, species**
77 **richness gradients, trait analysis**

78

79 Running title: Structural bias in assemblage/community analysis

80

81 INTRODUCTION

82 Community ecology, geographical ecology, ecological biogeography and some aspects of
83 macroecology and macroevolution frequently utilize metrics generated across communities or
84 assemblages. One fundamental pattern shared across all of these fields is spatial variation in
85 species richness, which can be quantified in grains ranging from small plots, for many ecological
86 questions, to entire continents, for biogeographical and macroevolutionary questions. Beginning
87 in the 1960's (Pianka, 1967), analyses of non-insular, broad-scale diversity gradients primarily
88 focused on quantifying relationships with components of the environment, which depending on
89 the grain/extent of the analysis and the taxon, normally included one or more measures of climate,
90 often supplemented with non-climatic variables such as, inter alia, area, topography, productivity,
91 soil or water properties, distance from source pools, or geological history (see Field *et al.* [2009]
92 for a compilation of case studies and the variables that have been considered). A major concern
93 of these analyses has been ranking the 'importance' of potential drivers of diversity, generally by
94 comparing regression coefficients or the relative statistical explanatory power of predictors.
95 Irrespective of the specific metrics, most analyses utilized extrinsic predictor variables, defined as
96 variables generated independent of the species in the plots, transects or grid cells. The majority of
97 the environmental predictors, particularly climatic variables, also contain strong spatial structure,
98 which were presumed to directly or indirectly generate the species richness patterns. There is a
99 very extensive literature associated with the analysis of such spatially structured data (e.g. Ripley,
100 1981; Haining, 2003; Dale & Fortin, 2014).

101 Recently, there has been increased interest in the analysis of intrinsic variables, defined as
102 variables calculated from attributes of the species known or assumed to be present in each
103 assemblage or community. Two that have been used for some time as response variables in

104 assemblage-based analyses include body size (with particular reference to Bergmann's Rule;
105 Blackburn & Hawkins, 2004; Diniz-Filho *et al.*, 2007; Olalla-Tárraga *et al.*, 2010; Slavenko &
106 Meiri, 2015) and range size (with reference to Rapoport's Rule; Stevens, 1989; Hawkins & Diniz-
107 Filho, 2006; Morin & Lechovicz, 2011). Other intrinsic variables, such as metrics generated
108 using the position of each species in a phylogeny, have also been correlated with species richness
109 patterns, often in combination with extrinsic predictors (Kerr & Currie, 1999; Hawkins *et al.*,
110 2005; Svenning *et al.*, 2008; Belmaker & Jetz, 2015). However, with the development of
111 community phylogenetics (Webb *et al.*, 2002) and trait-based approaches to studying community
112 size and structure (Shipley, 2010), the use of intrinsic variables as both response and predictor
113 variables in assemblage/community analyses is rapidly expanding (e.g., Swenson & Enquist,
114 2007; Jansson & Davies, 2008; Mayfield *et al.*, 2010; Swenson *et al.*, 2012, 2016; Dubuis *et al.*,
115 2013; Stuart-Smith *et al.*, 2013; Hawkins *et al.*, 2014; Leingärtner *et al.*, 2014; Albouy *et al.*,
116 2015; Belmaker & Jetz, 2015; Blonder *et al.*, 2015; Enquist *et al.*, 2015; Finegan *et al.*, 2015;
117 Godoy *et al.*, 2015; Honorio Coronado *et al.*, 2015; Lima-Mendez *et al.*, 2015; Seymour *et al.*,
118 2015; Šimová *et al.*, 2015; Stevens & Gavilanez, 2015; Zhang *et al.*, 2015; Biswas *et al.*, 2016;
119 Boucher-Lalonde *et al.*, 2016; González-Maya *et al.*, 2016; Kimberly *et al.*, 2016; Marin &
120 Hedges, 2016; Pfautsch *et al.*, 2016; de la Riva *et al.*, 2016). The assumption or hypothesis
121 underlying all such analyses is that species attributes sort geographically according to their
122 responses to the abiotic and biotic environment. Here we show that these biologically meaningful
123 assumptions cannot be evaluated from standard statistical associations of intrinsic variables
124 measured at the community or assemblage level.

125 Patterns of species richness are by their nature spatial, which raises a number of statistical
126 and inferential issues. The issue of spatial autocorrelation has been known to ecologists at least

127 since Legendre (1993), as has the problem that collinearity among predictors can be driven by a
128 joint environmental driver. However, a third ubiquitous and potentially serious analytical issue
129 related to the use of intrinsic variables in spatial analysis appears to have largely escaped notice.
130 We illustrate through the use of bird range maps, plot data for trees, and biologically plausible
131 simulated data sets an analytical problem associated with the use of intrinsic variables in
132 assemblage- and community-focused analyses conducted in a spatial context. The problem arises
133 whether the intrinsic variables are predictor or also as response variables, although our primary
134 focus is on analyses of species richness as the response variable.

135 A specific flavour of the problem was reported by Zelený & Schaffers (2012), who found
136 that mean Ellenberg indicator values, an intrinsic community-based variable used in vegetation
137 analysis, ‘inherited’ information about compositional similarity across communities, which then
138 resulted in overestimates of explained variance in correspondence analyses as well as in
139 regressions with species richness and inflated Type I error rates. They referred to this as a
140 ‘similarity issue’ caused by the fact that the same species often occur in multiple communities.
141 More recently, Peres-Neto *et al.* (2016) reported biased estimates of regression coefficients and
142 inflated Type I error rates between intrinsic community-based mean trait values and
143 environmental variables in the context of trait-environment analysis used in community ecology.
144 The problem does not require that the community data have explicit spatial structure, only that
145 some species occur in more than one community to the extent that some co-occurrences are
146 repeated (hereafter referred to as the co-occurrence problem). However, we might expect a priori
147 that the problem will be especially widespread in spatially structured assemblage data if there is
148 any overlap of species distributions caused by species-level responses to environmental gradients,
149 which will be rampant in datasets covering broad areas. To illustrate the severity of the problem

150 in two widely used types of data we first present analyses of the species richness patterns of North
151 American birds in their breeding ranges derived from range maps and tree community richness in
152 plots sampled by the United States Forest Service's Forest Inventory and Analysis (FIA). The
153 statistical models we generate use common linear, nonlinear, machine-learning and spatial
154 regression methods to quantify the strengths of associations among cell/plot species richness as
155 the response variable and sets of 'random intrinsic variables' as predictors, generated by assigning
156 random numbers as species attributes and calculating their cell/plot means. These attributes could
157 represent any quantitative physiological, morphological, ecological, behavioural or phylogenetic
158 variable generated from any taxon-level assignment of values.

159 In a second set of analyses we explore four potential influences on the problem of
160 particular relevance to ecologists and biogeographers, focusing on (i) levels of repeated species
161 co-occurrences, (ii) the spatial coherence of those occurrences, (iii) the existence of a strong
162 broad-scale richness gradient and (iv) the presence of spatial autocorrelation in the predictors.
163 For this we develop a set of simulated North Americas occupied by virtual species, to which each
164 species is given sets of random attributes as with the two data sets comprising real species. With
165 these random intrinsic variables as predictors we model the case in which a strong species
166 richness gradient is generated by species with cohesive ranges responding to the temperature
167 gradient found on the continent, followed by the case in which species still respond to
168 temperature but ranges lack coherence. Although less likely in real data of moderate to large
169 geographical extent, we also generate data sets without broad-scale richness gradients using
170 species with either cohesive or patchy ranges. Finally, we use the first of the simulated scenarios
171 to analyze community-level metrics in which random intrinsic variables comprise both response
172 and predictor variables. The latter analyses illustrate the potential extent of the problem when all

173 variables are intrinsic and generated from data containing repeated species co-occurrences.

174 **MATERIALS AND METHODS**

175 **North American birds**

176 Distribution maps were downloaded from BirdLife International
177 (<http://www.birdlife.org/datazone/info/spcdownload>, accessed in June, 2014), and breeding
178 ranges of the 1913 non-marine bird species in the region were extracted for analysis. The maps
179 were binned at a $0.5^\circ \times 0.5^\circ$ grain in a grid extending from the northern tip of Greenland to
180 Panama, and the presence-absence matrix (PAM) of 14,662 grid cells each containing at least 15
181 bird species was created. As intrinsic predictors of species richness we generated random
182 intrinsic variables, created by first assigning a real number between 0 and 1 taken from a uniform
183 random distribution as a species attribute to each bird species. We then calculated means for each
184 cell in the grid by averaging these random species attribute values for the birds found in the cell.
185 This two-step process was repeated 100 times to generate a population of 100 random intrinsic
186 variables for potential inclusion in statistical models of richness. Range-map based patterns of
187 species richness and species co-occurrences invariably have strong spatial autocorrelation due to
188 the high cohesiveness of most range maps. Data of this type are common in ecological
189 biogeography and geographical ecology.

190 **Trees in the conterminous United States**

191 We also generated a PAM for the 304 gymnosperm and angiosperm species in 104,588 plots
192 (each 0.07 h) in the US Forest Service's Forest Inventory and Analysis database
193 (<http://www.fia.fs.fed.us/>, accessed in January, 2012) that contained at least three species and
194 were in the conterminous USA. As with the birds, we generated 100 random intrinsic variables

195 by repeatedly assigning random species attributes to all species in the dataset and averaging their
196 values for species present in each plot, and these random intrinsic variables were then used as
197 predictors in statistical models of tree species richness. Because the data are plot-based counts,
198 species ranges are non-cohesive and expected to generate a substantially noisier and less spatially
199 autocorrelated richness pattern, although distributions are by no means random due to trees'
200 responses to spatially structured environmental drivers operating across a range of scales. This is
201 the data type used in community ecology, community phylogenetics and frequently in analyses of
202 altitudinal diversity gradients.

203 **Virtual North America**

204 We simulated species distributions in North America by defining their tolerances to annual mean
205 temperature (BIO 1 in WorldClim [Hijmans *et al.*, 2005]) within the 'virtualspecies' package in R
206 (Leroy *et al.*, 2016). To generate a species distribution, we simulated a Gaussian response to
207 temperature, defined by an optimum value and a thermal tolerance delimiting 99% of the area
208 under the Gaussian curve. We used this response to temperature to project the probability of
209 occurrence of the species in North America. Next, we converted probabilities of occurrence into
210 presence-absence with a probabilistic conversion. Lastly, we applied dispersal limitation with
211 two approaches: (i) a non-cohesive approach where a species distribution was limited to a defined
212 number of single-pixel habitat patches across North America; and (ii) a cohesive approach where
213 species distributions were limited to a cohesive range of size identical to its non-cohesive
214 counterpart. We expected the statistical problem to be most severe in the presence of a richness
215 gradient comprising species with cohesive ranges due to a higher level of repeated species co-
216 occurrences.

217 To sample species' optimal temperatures, we defined two scenarios: (i) a scenario with a
218 richness gradient (optimal temperatures more likely to be sampled at higher temperatures), and
219 (ii) a scenario with no richness gradient (optimal temperatures were randomly sampled along the
220 temperature gradient). Thermal tolerances were randomly sampled between 5° and 45°C for both
221 scenarios. These two scenarios were designed to test the co-occurrence effect on models where
222 there is a link between richness and a spatially structured environmental driver (temperature), and
223 where there is no link between richness and the environment, although the latter case is highly
224 unlikely in any real data set. For each scenario, we generated 2000 species, and we repeated the
225 process five times with different numbers of suitable habitat patches each time (250, 500, 1000,
226 2500 and 5000). We expected increases in numbers of available habitat patches to increase the
227 degree of co-occurrence among species. We characterized co-occurrence patterns by estimating
228 the C-score (Stone & Roberts, 1990) for each dataset/scenario. The C-score describes the average
229 pairwise value of species associations in a PAM, ranging from a lower bound of 0 (maximum
230 aggregation) to an undefined upper bound (Gotelli, 2000). Lower C-score values thus indicate
231 higher average co-occurrence across all species pairs. Given that a particular C-score is specific
232 to the PAM being analyzed, we used a modified version that normalizes the C-score according to
233 a general maximum derived from the data and thus can be compared across datasets (Dormann *et*
234 *al.*, 2008). To summarize, to facilitate interpretation of the results for the bird and tree data,
235 neither of which is replicated, we simulated a total of 20 virtual North Americas (two richness
236 scenarios × two range cohesiveness scenarios × five sizes of habitat patches).

237 As with the bird data, we generated a PAM for each scenario across the North American
238 grid and generated 100 random intrinsic variables by assigning random numbers as species
239 attributes and calculating assemblage means. These were selected as predictors of species

240 richness and for one scenario as the response variable as well.

241 **Statistical analyses**

242 A range of linear and non-linear modeling methods exist for analyzing assemblage/community
243 data focused on patterns of diversity, from which we selected five that have been commonly used
244 or are coming into common usage: ordinary least squares linear regression (MR), regression trees
245 (RT), generalized additive models (GAM), geographically weighted regression (GWR) and
246 simultaneous autoregression (SAR). These methods vary considerably in their underlying
247 assumptions and their ability to capture non-linear/non-stationary relationships, both of which are
248 widespread in broad-scale ecological datasets (Bini *et al.*, 2009) including our real and virtual
249 data. Because of the non-stationarity in the data, we selected geographically weighted regression
250 as our primary choice of a spatially explicit method, as it is explicitly designed to describe
251 spatially varying relationships among variables. Even so, because SAR is used by many workers
252 we evaluated its sensitivity to the co-occurrence problem using the bird data, the most strongly
253 spatially autocorrelated of the data sets. When evaluating the results using this method it should
254 be remembered that the coefficients are also sensitive to non-stationarity of the relationships
255 independent of repeated species co-occurrences (see Fotheringham *et al.* [2002], Bini *et al.*
256 [2009], Beale *et al.* [2010] and Hawkins [2012] for discussion of the assumptions underlying this
257 class of spatially explicit methods), so interpretation of the results contains some ambiguity.

258 The utilization of geographically weighted regression is also compromised by the fact that
259 we focused on a single bandwidth in the bird and tree data sets, 250.6 and 100.4, respectively;
260 generated by a preliminary evaluation of the method in the Geographically Weighted Regression
261 module in the SPATIAL ANALYSIS IN MACROECOLOGY program
262 (<https://www.ecoevol.ufg.br/sam/>). Model outputs are sensitive to the bandwidth, and selection

263 of appropriate bandwidths is itself a complex statistical issue (Cho *et al.*, 2010). Thus, changing
264 model parameters will change the results independently of the underlying structure of the data,
265 and the results presented here represent one of many possible outcomes. Even so, it provides a
266 warning that the method may be sensitive to the problem we describe in this paper.

267 Our rationale for selecting multiple modeling approaches was to evaluate the extent to
268 which the existing literature is likely affected by the co-occurrence problem. If the analytical
269 methods we evaluate are affected, it is likely that many other regression methods are affected as
270 well. At the very least we cannot rule out that possibility without examining all known methods,
271 which is beyond the scope of this paper. Zelený & Schaffers (2012) have already demonstrated
272 that correspondence analysis and correlation are sensitive to the problem.

273 For the real data sets (birds and trees) we generated sets of regression models of richness
274 using combinations of random intrinsic variables as predictors. Models using each method were
275 generated with one, three or five predictors, which is within the range of the number of predictors
276 evaluated by researchers. The sample size of the birds comprised the 14,662 cells containing at
277 least 15 species. For the trees computational limitations required randomly sampling 25,000 plots
278 supporting at least three species. The models of varying complexity were generated 100 times,
279 except in the case of the regression trees, for which 200 trees were generated in each case. Model
280 iterations used each random intrinsic variable in the one-predictor models or randomly selected
281 combinations of variables in the three- and five-predictor models. Evaluation of model fit
282 comprised coefficients of determination (R^2), or the model average R^2 in the case of
283 geographically weighted regression. We do not explicitly evaluate regression coefficients for four
284 of the five regression types, as they have no biological meaning with respect to sets of random
285 predictors and not all of the methods generate them. The exception is the SAR models, since they

286 are designed to account for the spatial autocorrelation in data and can generate high coefficients
287 of determination irrespective of the nature of the predictors; further, it is the more precise
288 coefficients generated by the method that justify its use (Beale *et al.*, 2010). Consequently, for
289 the SARs we determined how many of the coefficients across the models were significantly
290 different from 0. If the models often generate spurious coefficients it indicates that controlling for
291 spatial autocorrelation in the data does not remove the bias generated by the co-occurrence
292 problem. This also would represent one line of evidence that the problem we are evaluating in
293 this paper is not simply due to the spatial autocorrelation in the data, and we must look elsewhere
294 for an explanation.

295 Analysis of the 20 virtual scenarios comprised first fitting randomly selected sets of five
296 random intrinsic variables to species richness. Given the extremely strong fits found using
297 geographically weighted regression of the bird data and the large number of spurious regression
298 coefficients in the simultaneous autoregression models (see Results section), making the
299 sensitivity of both methods to the problem obvious, we excluded them from the analysis of the
300 simulated data. As before, each model was repeated 100 times using random combinations of
301 random intrinsic variables, and coefficients of determination were tallied.

302 The final analysis used one random intrinsic variable as a response variable and five
303 random intrinsic variables as predictors, derived from data in the 5000-patch, cohesive-ranges
304 scenario with a strong richness gradient. We repeated the analysis ten times with arbitrarily
305 chosen response variables, each replicated 100 times with random combinations of predictors.
306 Here we present the ‘best case’ and ‘worst case’ results, those with the lowest and highest mean
307 coefficients of determination among the sets of models of the 10 repetitions. Running models for
308 all 100 random intrinsic variables as response variables would expand the range of possible

309 results, but the results for 10 are sufficient to illustrate the potential severity of the problem when
310 using intrinsic variables derived from species presences as response variables in
311 assemblage/community analysis with strong spatial structure.

312

313 **RESULTS**

314 **Bird species richness**

315 The richness gradient generated by the breeding range maps is strongly spatially patterned (Fig.
316 1a), as is already well known (e.g., Cook, 1969; Orme *et al.*, 2005; Hawkins *et al.*, 2006).

317 Further, means generated from random attributes can contain obvious spatial structure across
318 multiple scales, as illustrated using three examples (Fig. 1b-d). Although the details of the spatial
319 patterns varied among the random intrinsic variables, they tended to share a common structure of
320 positive autocorrelation at small spatial scales and negative autocorrelation at very large scales, as
321 did the species richness gradient (Fig. 2). Further, statistical models of richness had moderate to
322 strong explanatory power across the model types (Table 1). Geographically weighted regression
323 was especially sensitive to covariation between random predictors and richness, and even a single
324 predictor variable generated very strong model fits. Simultaneous autoregression similarly
325 showed evidence of strong sensitivity; all 100 models generated at least one coefficient significant
326 at $P < 0.01$, and in 26 cases all five coefficients were significant (Appendix S1 in the Supporting
327 Information). Because of their ability to capture non-linear relationships, regression trees and
328 generalized additive models generated moderate to very strong models, despite the complete lack
329 of biological information in the predictors. Linear regression, due to the constraint of fitting
330 linear relationships, generated the weakest models on average, but even a single random predictor
331 could sometimes explain over half of the variance in richness (maximum $r^2 = 0.518$).

332 **Tree species richness**

333 The richness pattern for FIA plots is also spatially patterned, albeit noisy (Fig. 1e), as expected.
334 The range of richness values is low, also expected from the very small plot size (0.07 ha). At
335 least some of the random intrinsic variables also contain obvious spatial structure (Fig. 1f-h), and
336 all contain at least some small-scale positive autocorrelation with low to moderate levels of
337 broad-scale structure in many of them (Fig. 2b). Single predictor models of richness are in all
338 cases weaker than for the bird data, but regression trees and generalized additive models were
339 sensitive to the co-occurrence problem irrespective of the number of predictors (Table 2).
340 Geographically weighted regression was not as strongly impacted as for the bird data, but R^2 s
341 remained fairly high. In contrast, linear regression models were reasonably robust, perhaps only
342 because they are constrained to describe linear relationships. Our general finding is that although
343 both data sets are affected by the co-occurrence problem there are differences with respect to their
344 sensitivity, and these differences could at least potentially reflect that the plot data have (i) a
345 weaker broad scale species richness gradient, (ii) lower levels of spatial autocorrelation, and (iii)
346 lower levels of species co-occurrences (see next section). We explore these issues with the virtual
347 scenarios.

348 **Virtual North America**

349 The simulations provided evidence that all data likely to be analyzed by biogeographers are
350 sensitive to some extent to the co-occurrence problem, at least for the analytical methods we
351 examined (Fig. 3). It made rather little difference in the average model R^2 s whether the data were
352 derived from cohesive or patchy ranges (cf. Fig. 3a and c) or if they contained a broad-scale
353 species richness gradient (cf. Fig. 3a and b). The only data structure that did not generate
354 spurious models in at least some cases was when they are derived from patchy species

355 distributions in the absence of a richness gradient (Fig. 3d), a very unlikely structure in data
356 collected across any moderately strong environmental gradient.

357 Two consistent patterns in the virtual scenarios were that multiple regression models are
358 less strongly impacted than regression trees or generalized additive models, and the problem
359 becomes increasingly more severe with increasing levels of repeated species co-occurrences for
360 all analytical methods and three of four data structures (Fig. 3a,b,c.). We also note that the levels
361 of co-occurrence in some of the virtual scenarios were very similar to those found in both the bird
362 (Fig. 3a) and tree (Fig. 3c) data, and higher levels of co-occurrence are found in the bird than in
363 the trees, undoubtedly due in part to the cohesive ranges in the former.

364 Despite the results from the simultaneous regressions, it is possible that the spatial
365 autocorrelation found in all real data is at least part of the problem. We examined this by
366 quantifying the spatial patterns of the response and predictor variables in the virtual scenarios of
367 cohesive versus patchy ranges with C-scores near 0.79 (see Fig. 3a and c). If spatial
368 autocorrelation is the root of the problem, we expect both data sets to contain broadly similar
369 spatial patterning given that matched model fits (percent of variance explained) are similar in both
370 data sets despite the fact that the ranges that underlie the variables are structurally quite different.

371 Unsurprisingly, cohesive ranges generated similar patterns of spatial autocorrelation
372 between species richness and many of the random intrinsic variables (positive short-distance and
373 negative long-distance autocorrelation, Fig. 4a), so it is perhaps not surprising that model fits
374 were very high (Fig. 3a). However, using patchy ranges to generate a richness gradient
375 effectively removed the spatial pattern in the random intrinsic variables across all scales without
376 affecting the pattern in richness (Fig. 4b). Despite the almost complete spatial decoupling of
377 patterns in richness and the predictors, model fits remained high (Fig. 3c). Therefore, the

378 analytical problem can exist independent of any spatial autocorrelation in the predictors. On the
379 other hand, spatial patterning in the broad sense must have a role to play when groups of species
380 respond similarly to an environmental gradient, as the models are minimally impacted when
381 species do not respond to a spatially structured environmental gradient and are patchily
382 distributed (Fig. 3d).

383 **Traits as response variables**

384 The co-occurrence problem persists when the focus of an analysis is itself an intrinsic variable,
385 although not as severely (Fig. 5). In the subset of random intrinsic variables selected as response
386 variables both multiple regression and generalized additive models were moderately impacted,
387 whereas regression tree models remained very strong even though none of the variables in the
388 analysis, including the response variable, carry meaningful information.

389 **DISCUSSION**

390 Following Zelený & Schaffers (2012), we find that the community-focus widely used in ecology,
391 biogeography and macroecology suffers from a potentially severe structural problem with obvious
392 ramifications. First and foremost, any metric, whether physiological, morphological, behavioural,
393 functional, phylogenetic, or ecological, that is generated at the assemblage/community level by
394 assigning values to species and averaging them for the species present within a cell/plot can have
395 internal statistical relationships of no biological significance across communities. Thus, the
396 problem is likely to be widespread in community-based analyses in which species share multiple
397 sites. Most worrying in our context is that the statistical bias generated by repeated species co-
398 occurrences among sites is not slight in most biologically plausible scenarios, especially when
399 multiple intrinsic variables are involved. That sets of intrinsic variables derived from random

400 numbers can sometimes generate >90% explanatory power in statistical models of species
401 richness in spatially structured assemblages/communities suggests that no result using actual traits
402 or other attributes can be trusted, however strong the model may be. It also follows that it is not
403 possible to compare with confidence goodness-of-fits, regression coefficients or other measures
404 of variable importance or rank in analysis involving multiple intrinsic predictors. In some
405 situations, where levels of co-occurrence are low, multiple regression appears to be robust, but
406 without detailed analysis it is not possible to know why because of the multiple problems with
407 linear regression that have been identified when used to analyze spatially structured data
408 (Fotheringham *et al.*, 2002; Grace & Bollen, 2005; Bini *et al.*, 2009; Hawkins, 2012). We are
409 unable to address this complex set of statistical issues here.

410 Secondly, we expected range cohesion to exacerbate the analytical problem by generating
411 potentially spurious spatial autocorrelation among intrinsic predictor variables that would then
412 link to the underlying spatial autocorrelation in richness. If true this would identify results based
413 on range map data as being particularly unreliable, whereas the plot data normally generated by
414 community ecologists would be less impacted due to lower levels of autocorrelation. However,
415 our virtual data indicate that strong spurious relationships can occur in plot data without spatial
416 autocorrelation in the intrinsic predictors as long as richness itself is spatially structured (see Fig.
417 4b). Although it was possible to generate data with minimal apparent impact on the three
418 statistical methods (see Fig. 3d) few real data sets will have this structure, and so no data should
419 be considered a priori to be immune to the problem, and the presence or absence of spatial
420 autocorrelation is not definitive evidence that no problem exists, as long as some level of repeated
421 species co-occurrence exists across communities/assemblages.

422 Yet another ramification of the co-occurrence problem is that although spatial structure in
423 intrinsic variables is not required for co-occurrences to be an issue in statistical analysis and
424 ecological inference, spatially autocorrelated data are often used to generate maps showing
425 aggregated assemblage/community trait values at the sub-continental, continental or global extent
426 (Hawkins & Diniz-Filho, 2006; Morin & Lechowicz, 2011; Jetz *et al.*, 2012; Swenson *et al.*,
427 2012; Hawkins *et al.*, 2014; Šímová *et al.*, 2014; Belmaker & Jetz, 2015). The patterns in such
428 maps can be visually striking and yet at least potentially biologically uninformative. Thus, if
429 repeated species co-occurrences contain spatial structure, which they will if multiple species
430 respond similarly to the environment, it is not surprising that climate or other spatially structured
431 environmental variables could generate relatively strong regression models when trait values are
432 response variables. It does not follow that such patterns must be artefactual if the trait of interest
433 actually drives the species distributions; the problem is that any trait can contain spatial structure
434 due to the co-occurrence problem even if it is distributed independently of the environment (see
435 Fig. 1 for examples).

436 We are aware of two published solutions to the impact of repeated species co-occurrences
437 on community-level metrics. One is the permutation method proposed by Zelený & Schaffers
438 (2012) to correct the inflated Type I error. Their modified permutation test first calculates
439 observed test statistics (like Pearson's r coefficients for correlation or F -values for regression or
440 ANOVA) of relationships between cell/plot mean species attributes and sample attributes. Then,
441 these observed statistics are compared with the null distribution of expected test statistics,
442 calculated between cell/plot means of randomly permuted species attributes and sample attributes.
443 Note the difference of this approach and the use of null models in evaluating functional or
444 phylogenetic diversity indices (e.g. Mason *et al.*, 2013, with community weighted means being

445 one of them) based on calculating standardized effect sizes (SES, or z -scores). However, while
446 SES is devised to correct for the effect of species richness influencing the absolute values of these
447 indices, it does not solve the problem of repeated co-occurrences, which is not directly related to
448 species richness. The modified permutation test of Zelený & Schaffers (2012) does correct for
449 inflated Type I error but also does not correct regression coefficients or model fits. Whereas
450 accurate significance testing may be necessary and sufficient for many ecological applications, it
451 is of limited value for broad-scale analyses, particularly of diversity gradients, in which the focus
452 is typically on ranking the relative contributions of potential explanatory variables to compare
453 potential underlying processes. The challenge of distinguishing strong and weak predictors of
454 species richness gradients has generated much of the discussion in the ecological literature
455 evaluating methods for estimating regression coefficients for spatially structured data (e.g.,
456 Lennon, 2000; Diniz-Filho *et al.*, 2003; Dormann *et al.*, 2007; Hawkins, 2012; Kühn & Dormann,
457 2012). Uncertainty about ranking potential ‘effects’ of predictors makes disentangling the
458 contributions of the many hypothesized influences on diversity gradients difficult, and species co-
459 occurrences add yet another layer of difficulty for evaluating intrinsic variables.

460 The second approach to the problem of which we are aware is an adaptation of the fourth-
461 corner method by Peres-Neto *et al.* (2012, 2016), which claims to be immune to both the bias in
462 regression coefficients and inflated Type I error rate. This method is in fact a special case of
463 correlation between cell/plot means of species attributes (traits) and sample attributes, in which
464 both species and sample attributes are standardized, and the correlation itself is weighted by row
465 sums of the species composition matrix. These row sums represent the sum of species abundances
466 in the cells/plots, which in the case of presence/absence species composition data equal species
467 richness. This method may be suitable for community data relating species traits to

468 environmental variables, which is sometimes done by the original fourth-corner method.
469 However, in our opinion its current formulation cannot be used in the context of the analysis of
470 species richness, since using correlation weighted by species richness to analyze the relationship
471 between species richness and one or more intrinsic variables has no theoretical justification.
472 Further development of this approach may lead to a solution to the problem we address here, but
473 it is not obvious to us how to accomplish this.

474 Although not a solution per se, a relatively straight-forward approach to evaluate if
475 repeated co-occurrences might be a problem in a data set would be to conduct a separate set of
476 regressions using cell/plot means calculated from repeatedly re-randomized trait values. If 100 or
477 more iterations of such regressions always generate very low coefficients of determination it
478 suggests that patterns of repeated co-occurrences are not generating serious structural bias for the
479 statistical method being evaluated. On the other hand, if at least some models using repeatedly
480 randomized trait values are moderate to strong, confidence in the results will have to be limited
481 until a formal analytical solution is devised.

482 To conclude, there is clearly a potentially serious analytical problem with community-
483 based metrics as predictors of species richness gradients, but a methodological solution to the co-
484 occurrence problem with respect to understanding diversity patterns is not yet available. Until it
485 is, workers should be aware that inferences from maps of assemblage/community-level metrics
486 for any class of attribute, as well as analyses based on them using commonly used statistical
487 methods, can be much less certain than they appear.

488

489

490

491 **ACKNOWLEDGEMENTS**

492 We thank Pedro Peres-Neto for discussion of the problem addressed in this paper. We also thank
493 the anonymous reviewers and Oliver Schewieger for their valuable critiques of the ms. FV is
494 supported by a CNPq BJT (“Science without Borders”) fellowship. BV was supported by a
495 CAPES grant for doctoral studies. Work by MAR was supported by the Spanish Ministry of
496 Economy and Competitiveness (grant: CGL2013-48768-P). XW was supported by the
497 National Natural Science Foundation of China (31370620) and the State Scholarship Fund of
498 China (2011811457). DZ was supported by the Ministry of Science and Technology (MOST
499 105-2621-B-002-004).

500

501

502

503 **Table 1** Means (and SD) of coefficients of determination (R^2) of four types of statistical models
 504 of the species richness of North American birds (see Fig. 1) across 14,462 cells in a continental
 505 grid including one, three or five ‘random intrinsic variables’ as predictors. Each predictor
 506 variable represents mean cell values of random numbers taken from a uniform distribution
 507 between 0 and 1 and assigned to species. LR = linear regression, RT= regression trees, GAM =
 508 generalized additive models, GWR = geographically weighted regression.

509	No. of				
510	Predictors	LR	RT	GAM	GWR
511					
512	One	0.145	0.321	0.304	0.936
513		(0.143)	(0.087)	(0.159)	(0.004)
514	Three	0.310	0.702	0.584	0.952
515		(0.148)	(0.061)	(0.126)	(0.004)
516	Five	0.437	0.853	0.732	0.964
517		(0.114)	(0.032)	(0.074)	(0.003)

518

519 The requisite numbers of predictors were randomly selected from a population of 100 random
 520 intrinsic variables. Each model type was run with 100 combinations of predictors, or each
 521 predictor once in the one-predictor models. The regression tree values were calculated from 200
 522 component trees in random forest models generated in the ‘randomForest’ package in R. The
 523 simple and multiple regression models comprise linear terms of predictors with no interactions,
 524 and the degrees of freedom for the smooth terms in the GAMs were estimated using the

525 Generalized Cross Validation criterion (for details see the `gam` function in the ‘`mgcv`’ R package).

526 See the text for the details of the GWR models.

527

528 **Table 2** Means (and SD) of coefficients of determination (R^2) of four types of statistical models
 529 of the species richness of trees in US Forest Service Forestry Inventory and Analysis plots (see
 530 Fig. 1) including one, three or five ‘random intrinsic variables’ as predictors. Each predictor
 531 variable represents mean plot values of random numbers taken from a uniform distribution
 532 between 0 and 1 and assigned to species. LR = linear regression, RT = regression trees, GAM =
 533 generalized additive models, GWR = geographically weighted regression. Modelling details as in
 534 Table 1.

535 No. of				
536 Predictors	LR	RT	GAM	GWR
538 One	0.021	0.146	0.184	0.443
	(0.025)	(0.016)	(0.036)	(0.006)
540 Three	0.058	0.444	0.441	0.465
	(0.034)	(0.021)	(0.038)	(0.010)
542 Five	0.084	0.593	0.542	0.495
	(0.042)	(0.020)	(0.030)	(0.012)

544

545

546 **REFERENCES**

- 547 Albouy, C., Leprieur, F., Le Loc'h, F., Mouquet, N., Meynard, C. N., Douzery, E. J. P. &
548 Mouillot, D. (2015) Projected impacts of climate warming on the functional and
549 phylogenetic components of coastal Mediterranean fish biodiversity. *Ecography*, **38**, 681–
550 689.
- 551 Beale, C. M., Lennon, J. J., Yearsley, J. M., Brewer, M. J. & Elston, D. A. (2010) Regression
552 analysis of spatial data. *Ecology Letters*, **13**, 246-264,
- 553 Belmaker, J. & Jetz, W. (2015) Relative roles of ecological and energetic constraints,
554 diversification rates and region history on global species richness gradients. *Ecology Letters*,
555 **18**, 563–571.
- 556 Bini, L. M. *et al.* (2009) Coefficient shifts in geographical ecology: an empirical evaluation of
557 spatial and non-spatial regression. *Ecography*, **32**, 193–204.
- 558 Biswas, S. R., Mallik, A. U., Braithwaite, N. T. & Wagner, H. H. (2016) A conceptual framework
559 for the spatial analysis of functional trait diversity. *Oikos*, **125**, 192–200.
- 560 Blackburn, T. M. & Hawkins, B. A. (2004) Bergmann's rule and the mammal fauna of northern
561 North America. *Ecography*, **27**, 715–724.
- 562 Blonder, B., Nogués-Bravo, D., Borregaard, M. K., Donoghue II, J. C., Jørgensen, P. M., Kraft,
563 N. J. B., Lessard, J.-P., Morueta-Holme, N., Sandel, B., Svenning, J.-C., Violle, C., Rahbek,
564 C. & Enquist, B. J. (2015) Linking environmental filtering and disequilibrium to
565 biogeography with a community climate framework. *Ecology*, **96**, 972–985.
- 566 Boucher-Lalonde, V., Morin, A. & Currie, D. J. (2016) Can the richness-climate relationship be
567 explained by systematic variations in how individual species' ranges related to climate?
568 *Global Ecology and Biogeography*, **25**, 527-539.

569 Cho, S.-H., Lambert, D. M. & Chen, Z. (2010) Geographically weighted regression bandwidth
570 selection and spatial autocorrelation: an empirical example using Chinese agriculture data.
571 *Applied Economics Letters*, **17**, 767-772.

572 Cook, R. E. (1969) Variation in species density of North American birds. *Systematic Zoology*, **18**,
573 63-84.

574 Dale, M. R. T. & Fortin, M.-J. (2014) *Spatial analysis: a guide for ecologists*. 2nd Ed., Cambridge
575 University Press, Cambridge, UK.

576 Diniz-Filho, J. A. F., Bini, L. M. & Hawkins, B. A. (2003) Spatial autocorrelation and red
577 herrings in geographical ecology. *Global Ecology and Biogeography*, **12**, 53–64.

578 Diniz-Filho, J. A. F., Bini, L. M., Rodríguez, M. Á., Rangel, T. F. L. V. B. & Hawkins, B. A.
579 (2007) Seeing the forest for the trees: partitioning ecological and phylogenetic components of
580 Bergmann’s rule in European Carnivora. *Ecography*, **30**, 598–608.

581 Dormann, C. R., McPherson, J. M., Araújo, M. B., Bivand, R., Bolliger, J., Carl, G., Davies, R. G.,
582 Hirzel, A., Jetz, W., Kissling, W. D., Kühn, I., Ohlemüller, R., Peres-Neto, P. R., Reineking,
583 B., Schröder, B., Schurr, F. M. & Wilson, R. (2007) Methods to account for spatial
584 autocorrelation in the analysis of species distributional data: a review. *Ecography*, **30**, 609–
585 628.

586 Dormann, C. F., Gruber B. & Freund, J. (2008). Introducing the bipartite package: analysing
587 ecological networks. *R News*, **8/2**, 8–11

588 Dubuis, A., Rossier, L., Pottier, J., Pellissier, L., Vittoz, P. & Guisan, A. (2013) Predicting current
589 and future spatial community patterns of plant functional traits. *Ecography*, **36**, 1158–1168.

590 Enquist, B. J., Norberg, J., Bonsor, S. P., Violle, C., Webb, C. T., Henderson, A., Sloat, L. L. &
591 Savage, V.M. (2015) Scaling from traits to ecosystems: developing a general trait driver

592 theory via integrating trait-based and metabolic scaling theories. *Trait-Based Ecology –*
593 *From Structure to Function. Advances in Ecological Research*, **52**, 249–318.

594 Field, R., Hawkins, B. A., Cornell, H. V., Currie, D. J., Diniz-Filho, J. A. F., Guégan, J.-F.,
595 Kaufman, D. M., Kerr, J. T., Mittelbach, G. G., Oberdorff, T., O'Brien, E. M. & Turner, J.
596 R. G. (2009) Spatial species-richness gradients across scales: a meta-analysis. *Journal of*
597 *Biogeography*, **36**, 132–147.

598 Finegan, B., Peña-Claros, M., de Oliveira, A., Ascarrunz, N., Bret-Harte, M. S., Carreño-
599 Rocabado, G., Casanoves, F., Díaz, S., Velepucha, P. E., Fernandez, F., Licona, J. C.,
600 Lorenzo, L. Negret, B. S., Vaz, M. & Poorter, L. (2015) Does functional trait diversity
601 predict above-ground biomass and productivity of tropical forests? Testing three alternative
602 hypotheses. *Journal of Ecology*, **103**, 191–201.

603 Fotheringham, A.S., Brunsdon, C. & Charlton, M. (2002) *Geographically weighted regression:*
604 *the analysis of spatially varying relationships*. Wiley, Chichester.

605 Godoy, O., Rueda, M. & Hawkins, B. A. (2015) Functional determinants of forest recruitment
606 over broad scales. *Global Ecology and Biogeography*, **24**, 192–202.

607 González-Maya, J. F., Viquez-R, L. R., Arias-Alzate, A., Belant, J. L. & Ceballos, G. (2016)
608 Spatial patterns of species richness and functional diversity in Costa Rican terrestrial
609 mammals: implications for conservation. *Diversity and Distributions*, **22**, 43–56.

610 Gotelli, N. J. (2000) Null models of species co-occurrence patterns. *Ecology*, **81**, 2606 – 2621.

611 Grace, J.B. & Bollen, K.A. (2005) Interpreting the results from multiple regression and structural
612 equation models. *Bulletin of the Ecological Society of America*, **86**, 283–295.

613 Haining, R. (2003) *Spatial data analysis*. Cambridge University Press, Cambridge, UK.

614 Hawkins, B. A. (2012) Eight (and a half) deadly sins of spatial analysis. *Journal of*

615 *Biogeography*, **39**, 1–9.

616 Hawkins, B. A. & Diniz-Filho, J. A. F. (2006) Beyond Rapoport's rule: evaluating range size
617 patterns of New World birds in a two-dimensional framework. *Global Ecology and*
618 *Biogeography*, **15**, 461–469.

619 Hawkins, B. A., Diniz-Filho, J. A. F. & Soeller, S. A. (2005) Water links the historical and
620 contemporary components of the Australian bird diversity gradient. *Journal of*
621 *Biogeography*, **32**, 035–1042.

622 Hawkins, B. A., Diniz-Filho, J. A. F., Jaramillo, C. A. & Soeller, S. A. (2006) Post-Eocene
623 climate change, niche conservatism, and the latitudinal diversity gradient of New World
624 birds. *Journal of Biogeography*, **33**, 770–780.

625 Hawkins, B. A., Rueda, M., Rangel, T. F., Field, R. & Diniz-Filho, J. A. F. (2014) Community
626 phylogenetics at the biogeographical scale: cold tolerance, niche conservatism and the
627 structure of North American forests. *Journal of Biogeography*, **41**, 23–38.

628 Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G. and Jarvis, A. (2005) Very high
629 resolution interpolated climate surfaces for global land areas. *International Journal of*
630 *Climatology*, **25**, 1965–1978.

631 Honorio Coronado, E. N., *et al.* (2015) Phylogenetic diversity of Amazonian tree communities.
632 *Diversity and Distributions*, **21**, 1295–1307.

633 Jansson, R. & Davies, T. J. (2008) Global variation in diversification rates of flowering plants:
634 energy vs. climate change. *Ecology Letters*, **11**, 173–183.

635 Jetz, W., Thomas, G. H., Joy, J. B., Hartmann, K. & Mooers, A. O. (2012) The global diversity of
636 birds in time and space. *Nature*, **491**, 444–448.

637 Kerr, J. T. & Currie, D. J. (1999) The relative importance of evolutionary and environmental

638 controls on broad-scale patterns of species richness in North America. *Ecoscience*, **6**, 329–
639 337.

640 Kimberly, A., Blackburn, G. A., Whyatt, J. D. & Smart, S. M. (2016) How well is current plant
641 trait composition predicted by modern and historical forest spatial configuration?
642 *Ecography*, **39**, 67–76.

643 Kühn, I. & Dormann, C. F. (2012) Less than eight (and a half) misconceptions of spatial analysis.
644 *Journal of Biogeography*, **39**, 995–998.

645 Legendre, P. (1993) Spatial autocorrelation: trouble or new paradigm? *Ecology*, **74**, 1659–1673.

646 Leingärtner, A., Krauss, J. & Steffan-Dewenter, I. (2014) Species richness and trait composition
647 of butterfly assemblages change along an altitudinal gradient. *Oecologia*, **175**, 613–623.

648 Lennon, J. J. (2000) Red-shifts and red herrings in geographical ecology. *Ecography*, **23**, 101–
649 113.

650 Leroy, B., Meynard, C. N., Bellard, C. & Courchamp, F. (2016) virtual species, an R package to
651 generate virtual species distributions. *Ecography*, **39**, 599–607.

652 Lima-Mendez, G., *et al.* (2015) Determinants of community structure in the global plankton
653 interactome. *Science*, **348**, doi: 10.1126/science.1262073.

654 Marin, J. & Hedges, S. B. (2016) Time best explains global variation in species richness of
655 amphibians, birds and mammals. *Journal of Biogeography*, **43**, 1069–1079.

656 Mason, N. W. H., de Bello, F., Mouillot, D., Pavoine S. & Dray S. (2013). A guide for using
657 functional diversity indices to reveal changes in assembly processes along ecological
658 gradients. *Journal of Vegetation Science*, **24**, 794–806.

659 Mayfield, M. M., Bonser, S. P., Morgan, J. W., Aubin, I., McNamara, S. & Vesk, P. A. (2010)
660 What does species richness tell us about functional trait diversity? Predictions and evidence

661 for responses of species and functional trait diversity to land-use change. *Global Ecology*
662 *and Biogeography*, **19**, 423–431.

663 Morin, X. & Lechowicz, M. J. (2011) Geographical and ecological patterns of range size in North
664 American trees. *Ecography*, **34**, 738–750.

665 Olalla-Tárraga, M. Á., Bini, L. M., Diniz-Filho, J. A. F. & Rodríguez, M. Á. (2010) Cross-species
666 and assemblage-based approaches to Bergmann's rule and the biogeography of body size in
667 *Plethodon* salamanders of eastern North America. *Ecography*, **33**, 362–368.

668 Orme, C. D. L. *et al.* (2005) Global hotspots of species richness are not congruent with endemism
669 or threat. *Nature*, **436**, 1016–1019.

670 Peres-Neto, P.R., Leibold, M.A. & Dray, S. (2012) Assessing the effects of spatial contingency
671 and environmental filtering on metacommunity phylogenetics. *Ecology*, **93**, S14–S30.

672 Peres-Neto, P. R., Dray, S. & ter Braak, C. J. F. (2016) Linking trait variation to the environment:
673 critical issues with community-weighted mean correlation resolved by the fourth-corner
674 approach. *Ecography*, doi: 10.1111/ecog.02302.

675 Pianka, E. R. (1967) On lizard species diversity: North American flatland deserts. *Ecology*, **48**,
676 333–351.

677 Pfautsch, S., Harbusch, M., Wesolowski, A., Smith, R., Macfarlane, C., Tjoelker, M. G., Reich, P.
678 B. & Adams, M. A. (2016) Climate determines vascular traits in the ecologically diverse
679 genus *Eucalyptus*. *Ecology Letters*, **19**, 240–248.

680 Ripley, B. D. (1981) *Spatial statistics*. Wiley Press, New York, NY, USA.

681 de la Riva, E. G., Pérez-Ramos, I. M., Tosto, A., Navarro-Fernández, C. M., Olmo, M., Marañón,
682 T. & Villar, R. (2016) Disentangling the relative importance of species occurrence,

683 abundance and intraspecific variability in community assembly: a trait-based approach at
684 the whole-plant level in Mediterranean forests. *Oikos*, **125**, 354–363.

685 Seymour, C. L., Simmons, R. E., Joseph, G. S. & Sliingsby, J. A. (2015) On bird functional
686 diversity: species richness and functional differentiation show contrasting responses to
687 rainfall and vegetation structure in an arid landscape. *Ecosystems*, **18**, 971–984.

688 Shipley, B. (2010) *From plant traits to vegetation structure. Chance and selection in the assembly*
689 *of ecological communities*. Cambridge University Press, Cambridge, UK.

690 Šimová, I., Violle, C., Kraft, N. J. B., Storch, D., Svenning, J.-C., Boyle, B., Donoghue J. C. II,
691 Jørgensen, P., McGill, B. J., Morueta-Holme, N., Piel, W. H., Peet, R. K., Regetz, J.,
692 Schildhauer, M., Spencer, N., Thiers, B., Wisser, S. & Enquist, B. J. (2015) Shifts in trait
693 means and variances in North American tree assemblages: species richness patterns are
694 loosely related to the functional space. *Ecography*, **38**, 649–658.

695 Slavenko, A. & Meiri, S. (2015) Mean body sizes of amphibian species are poorly predicted by
696 climate. *Journal of Biogeography*, **42**, 1246–1254.

697 Stevens, G. C. (1989) The latitudinal gradient in geographical range: how so many species coexist
698 in the tropics. *The American Naturalist*, **113**, 240–256.

699 Stevens, R. D. & Gavilanez, M. M. (2015) Dimensionality of community structure: phylogenetic,
700 morphological and functional perspectives along biodiversity and environmental gradients.
701 *Ecography*, **38**, 861–875.

702 Stone, L. & Roberts, A. (1990) The checkerboard score and species distributions. *Oecologia*,
703 **85**, 74–79.

704 Stuart-Smith, R. D., Bates, A. E., Lefcheck, J. S., Duffy, J. E., Baker, S. C., Thomson, R. J.,
705 Stuart-Smith, J. F., Hill, N. A., Kininmonth, S. J., Airoidi, L., Becerro, M. A., Campbell, S.

706 J., Dawson, T. P., Navarrete. S. A., Soler, G. A., Strain, E. M. A., Willis, T. J. & Edgar, G.
707 J. (2013) Integrating abundance and functional traits reveals new global hotspots of fish
708 diversity. *Nature*, **501**, 539–542.

709 Svenning, J.-C., Borchsenius, F., Bjorholm, S. & Balslev H. (2008) High tropical net
710 diversification drives the New World latitudinal gradient in palm (Arecaceae) species
711 richness. *Journal of Biogeography*, **35**, 394–406.

712 Swenson, N. G. & Enquist, B. J. (2007) Ecological and evolutionary determinants of a key plant
713 functional trait: wood density and its community-wide variation across latitude and
714 elevation. *Journal of Botany*, **94**, 451–459.

715 Swenson, N. G. *et al.* (2012) The biogeography and filtering of woody plant functional diversity
716 in North and South America. *Global Ecology and Biogeography*, **21**, 798–808.

717 Swenson, N. G., Weiser, M. D., Mao, L., Normand, S., Rodriguez, M. Á., Lin, L., Cao, M. &
718 Svenning, J.-C. (2016) Constancy in functional space across a species richness anomaly.
719 *The American Naturalist*, **187**, E83–E92.

720 Webb, C.O., Ackerly, D. D., McPeck, M. A. & Donoghue, M. J. (2002) Phylogenies and
721 community ecology. *Annual Review of Ecology, Evolution and Systematics*, **33**, 475–505.

722 Zelený, D. & Schaffers, A. P. (2012) Too good to be true; pitfalls of using mean Ellenberg
723 indicator values in vegetation analyses. *Journal of Vegetation Science*, **23**, 419–431.

724 Zhang, Y., Wang, R., Kaplan, D. & Liu, J. (2015) Which components of plant diversity are most
725 correlated with ecosystem properties? A case study in a restored wetland in northern China.
726 *Ecological Indicators*, **49**, 228–236.

727

728 **SUPPORTING INFORMATION**

729 Additional Supporting Information may be found in the online version of this article:

730

731 **Appendix S1** Results for SAR models.

732

733

734 **DATA ACCESSIBILITY**

735 The bird range maps are available at (<http://www.birdlife.org/datazone/info/spcdownload>, and the

736 FIA data are available at (<http://www.fia.fs.fed.us/>. All scripts used for the simulations are

737 available at <https://github.com/Farewe/CooccurrenceIssue>.

738

739

740

741 **BIOSKETCH**

742 **Bradford A. Hawkins** is interested in ecological and phylogenetic patterns across a range of

743 spatial scales, with a focus on linking local and biogeographical processes.

744

745 Editor: Jens-Christian Svenning

746 FIGURE LEGENDS

747 **Figure 1** Species richness pattern of (a) North American bird species derived from maps of
748 breeding ranges gridded at a $0.5^\circ \times 0.5^\circ$ grain, and (e) trees in 0.07 h plots recorded by the US
749 Forest Service Forest Inventory and Analysis. (b-d, f-h) Three examples of cell/plot means
750 (random intrinsic variables) in which random values between 0 and 1 were assigned to each
751 species of bird or tree. All colour schemes are in the natural-jenks scale from ArcGIS 10.3.

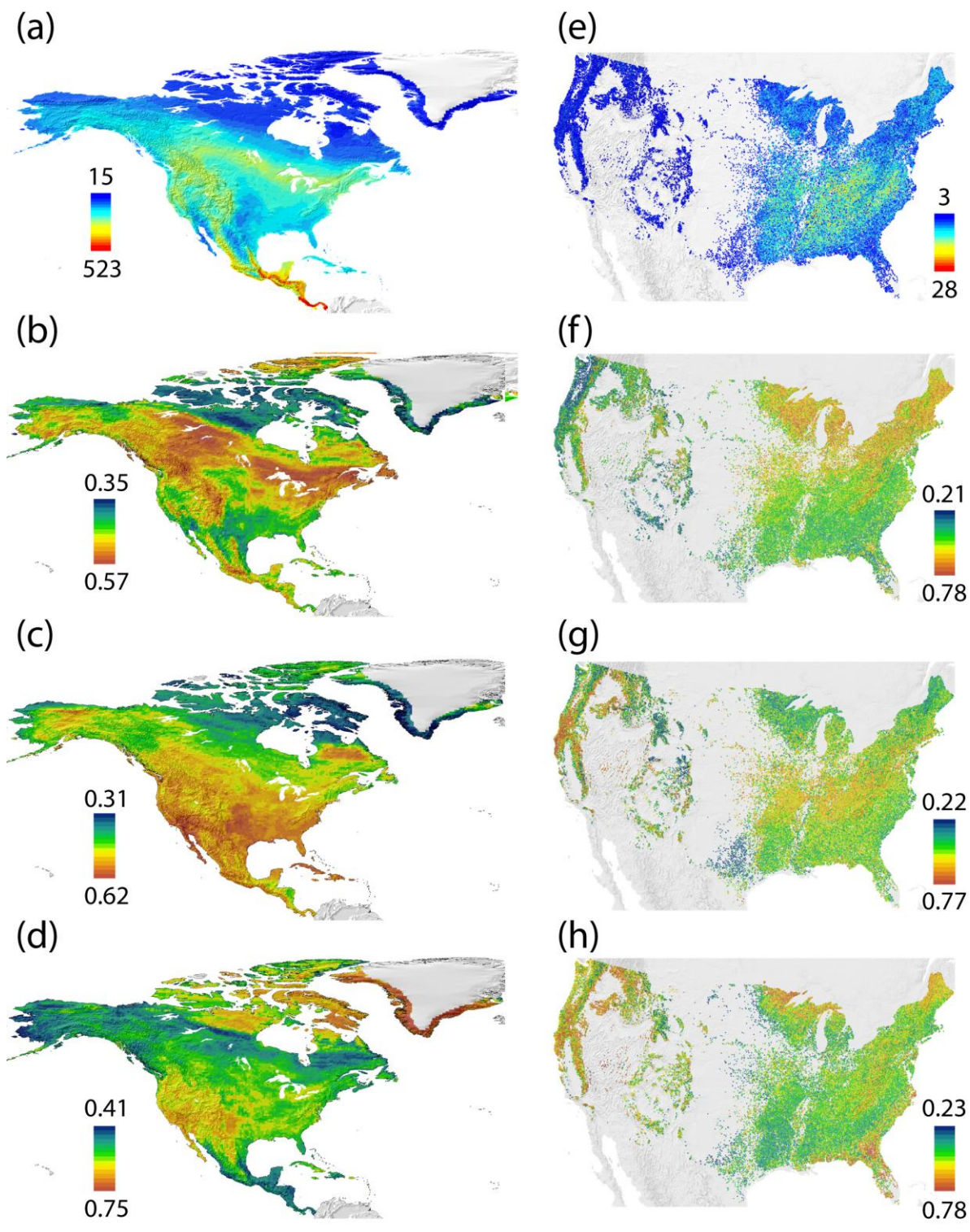
752
753 **Figure 2** Spatial autocorrelation structure (Moran's I) of (a) the North American bird species
754 richness pattern (black line) and 100 random intrinsic variables (gray lines), and (b) US tree
755 richness pattern and 100 random intrinsic variables. Note difference in scale of axes.

756
757 **Figure 3.** Mean (± 1 SD) coefficients of determination of three types of five-predictor statistical
758 models of species richness for four simulated North American scenarios plotted against a measure
759 of species co-occurrences (C-score) calculated for five range size distributions: (a) data containing
760 a strong species richness generated by species with cohesive ranges, (b) no broad-scale richness
761 gradient generated by species with cohesive ranges, (c) a broad-scale richness gradient generated
762 by species with patchy (non-cohesive) ranges, and (d) data with no broad-scale richness gradients
763 generated by species with patchy ranges. Within each scenario C-scores vary depending on the
764 average realized range size of the species, which is influenced by the number of available patches
765 species can occupy. For comparison, C-scores and model fits (data from Tables 1 and 2) for birds
766 and trees are shown in the scenarios to which their data correspond. Model types are regression
767 tree (RT), generalized additive models (GAM) and multiple linear regression (MR).

768

769 **Figure 4** Spatial autocorrelation structure (Moran's I) of simulated data under the scenarios in
770 which the data contain a broad-scale species richness gradient generated by species with (a)
771 cohesive or (b) non-cohesive, 1000-patch distributions. Black lines describe the spatial structure
772 of richness and gray lines describe structure of 100 random intrinsic variables in each scenario.

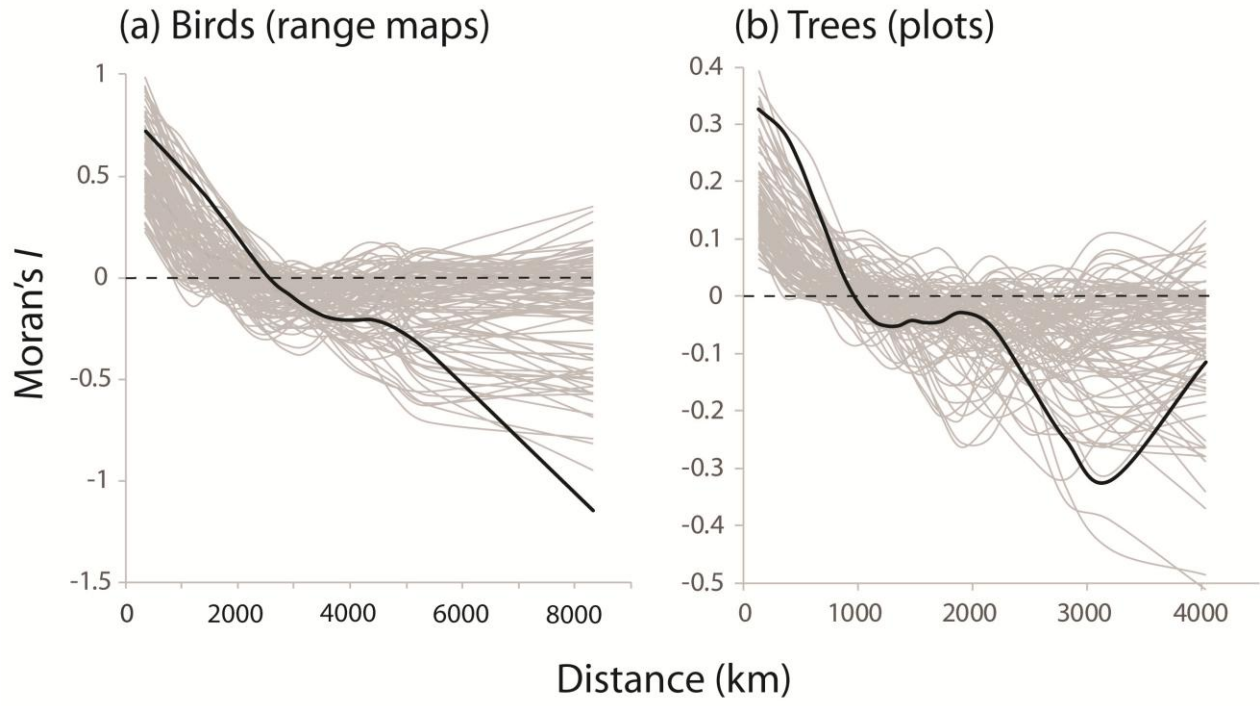
773
774 **Figure 5** Frequency distributions of coefficients of determination from three types of statistical
775 models generated using five random intrinsic variables as predictors against one random intrinsic
776 variable as the response variable, in a simulated North American scenario with a broad-scale
777 richness gradient generated by species with cohesive ranges. Each model type was iterated 100
778 times using random selections of predictors from a population of 99 random variables excluding
779 the variable used as the response. (a, c, e) The weakest models (best case) and (b, d, f) strongest
780 models (worst case) selected from analyses of 10 arbitrarily selected response variables. Vertical
781 dashed lines identify mean values.



782

783

Figure 1

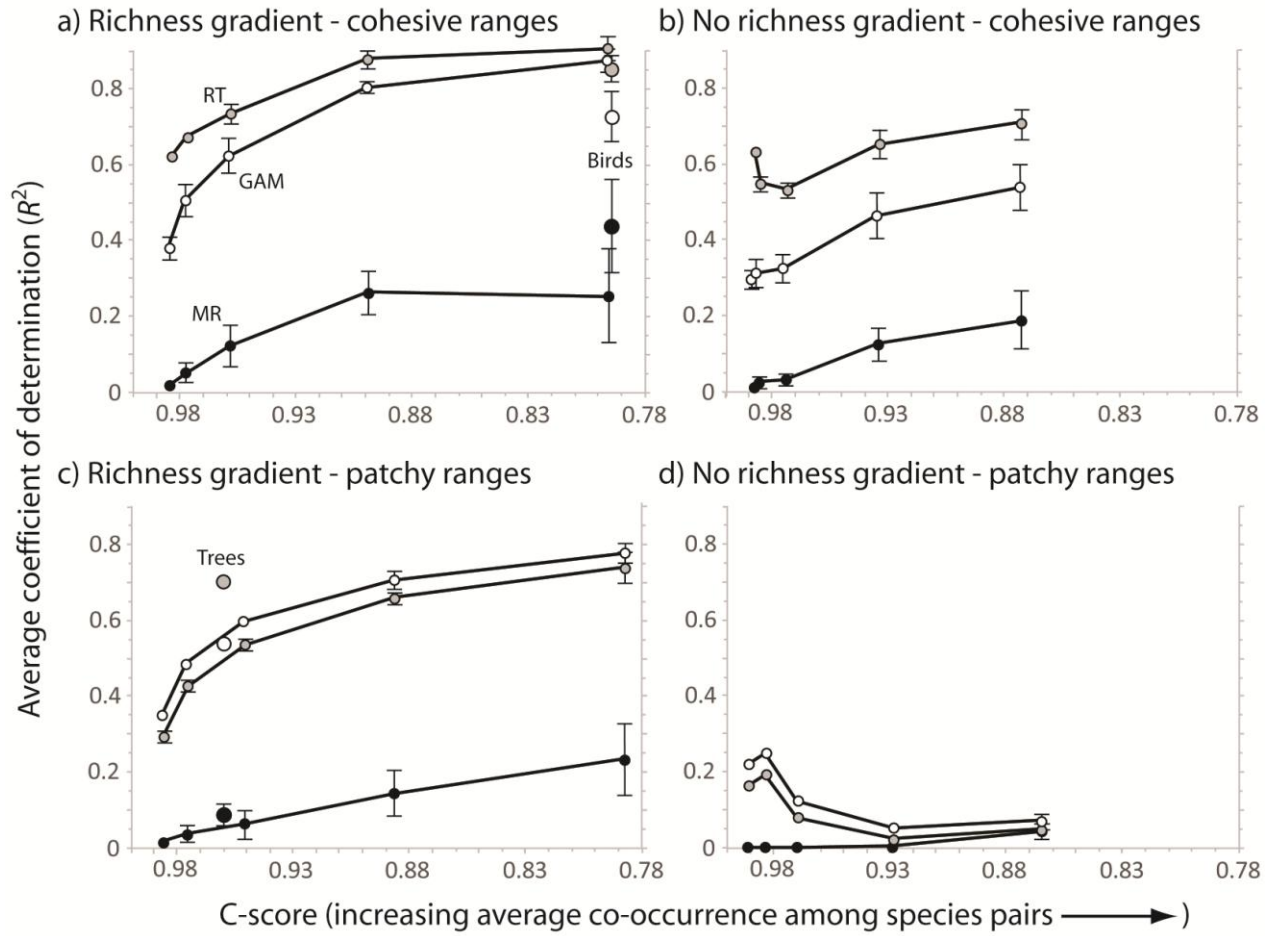


784

785

786

Figure 2

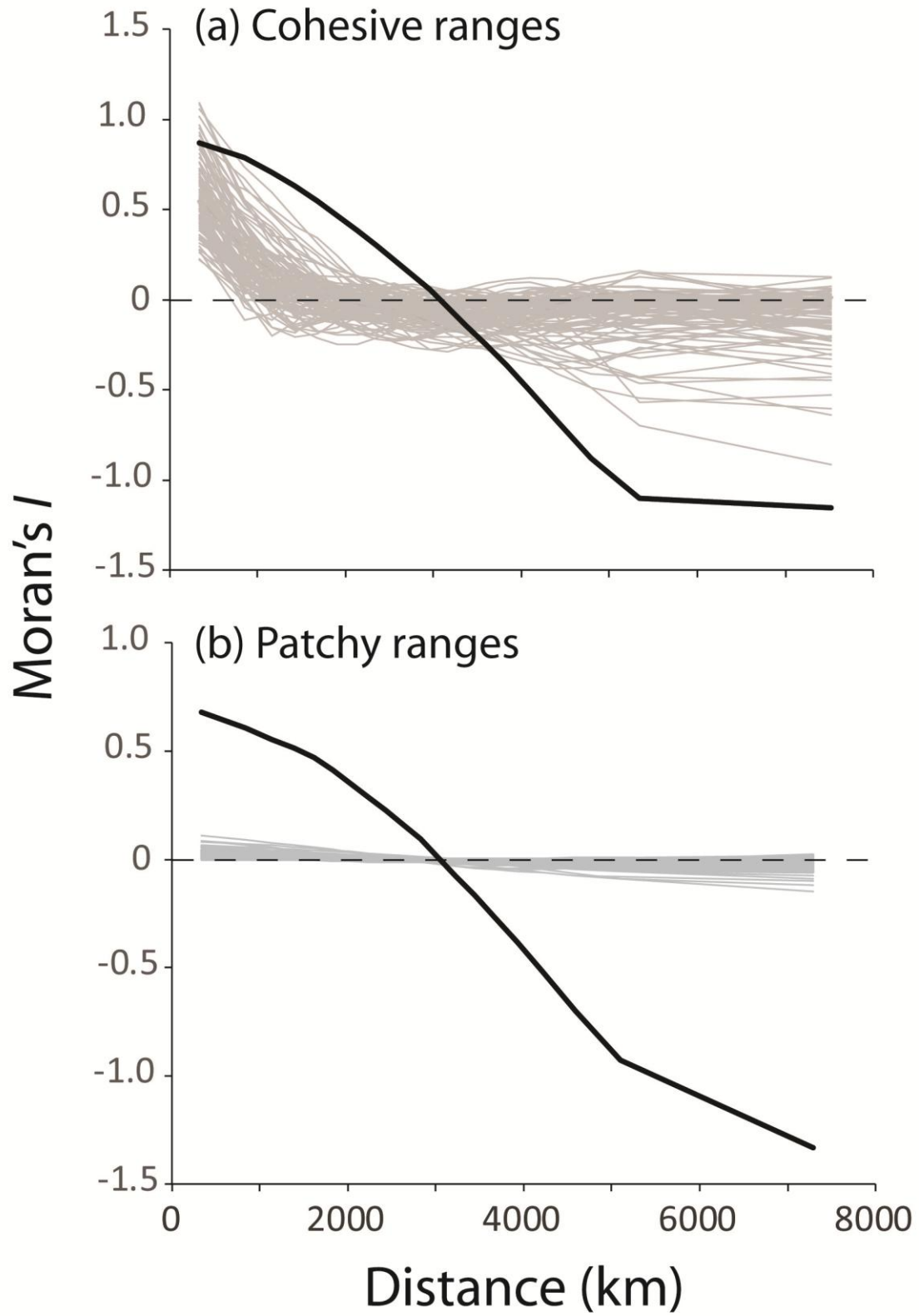


787

788

789

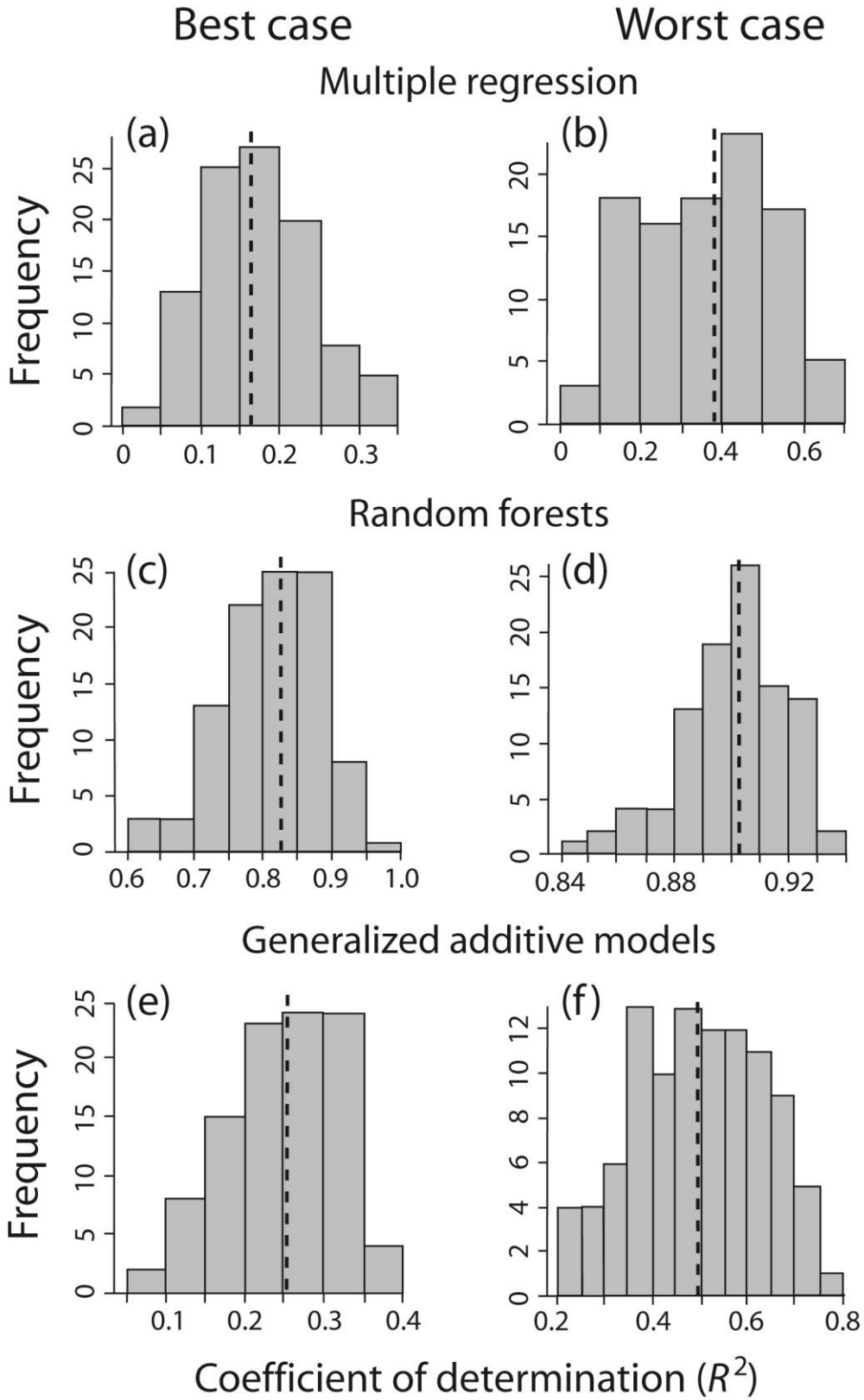
Figure 3



790

791

Figure 4



792

793

Figure 5

794

795

SUPPORTING INFORMATION

796 **Structural bias in aggregated trait variables driven by repeated**

797 **species co-occurrences: a pervasive problem in community and**

798 **assemblage data**

799

800 Bradford A. Hawkins, Boris Leroy, Miguel Á. Rodríguez, Alexander Singer, Bruno Vilela,

801 Fabricio Villalobos, Xiangping Wang and David Zelený

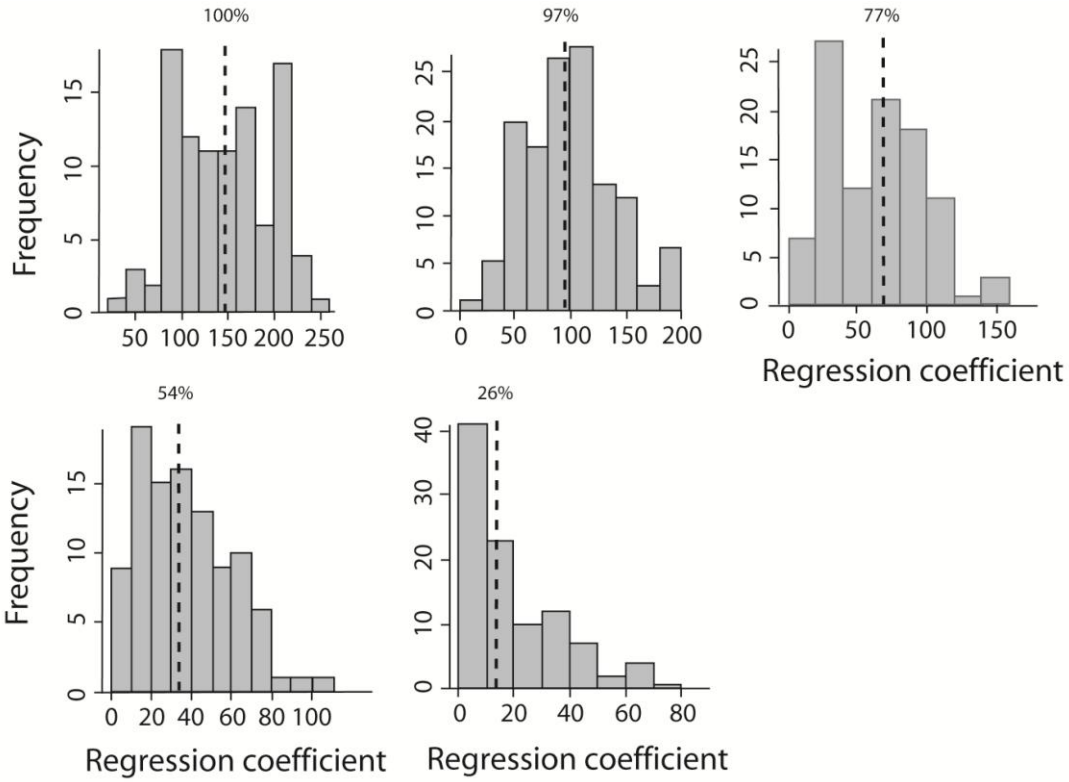
802

803 **Appendix S1** Frequency distributions of ranked regression coefficients (absolute values) of
804 simultaneous autoregressive (SAR) models of the species richness pattern of North American
805 birds in 14,662 grid cells, with five random intrinsic variables as predictors. The models were
806 iterated 100 times using randomly selected combinations of predictors from a population of 100
807 random intrinsic variables. The coefficients of determination of all iterations were extremely high
808 (mean = 0.980, SD = 0.001). The sensitivity to the co-occurrence problem was evaluated based
809 on the number of significant ($P < 0.01$) non-zero regression coefficients. For illustrative purposes
810 the five coefficients from each iteration have been ranked and the frequency distributions across
811 models of the strongest to weakest are plotted (dashed line = mean; note decreasing values from
812 top left to bottom right). The percentage of the 100 coefficients in each distribution that were
813 significant is provided above each panel.

814 The analyses were conducted using the `spautol()` function of the R package ‘Spdep’. We
815 used the sparse matrix decomposition method (Monte Carlo), which can handle large data sets
816 with thousands of observations. We used the `knearneigh()` function to create the matrix with the
817 indices of points belonging to the set of the k ($k = 4$) nearest neighbours. The matrix was

818 converted to a neighbours list with spatial weights using the knn2nb () and nb2listw () functions,
819 which were then used in the SAR models.

820



821

822

823