



**HAL**  
open science

## Generating spatially constrained null models for irregularly spaced data using Moran spectral randomization methods

H. H. Wagner, Stéphane Dray

► **To cite this version:**

H. H. Wagner, Stéphane Dray. Generating spatially constrained null models for irregularly spaced data using Moran spectral randomization methods. *Methods in Ecology and Evolution*, 2015, 6, pp.1169-1178. 10.1111/2041-210x.12407. hal-02018998

**HAL Id: hal-02018998**

<https://univ-lyon1.hal.science/hal-02018998v1>

Submitted on 29 Jan 2024

**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.

2015

## Generating spatially-constrained null models for irregularly spaced data using Moran spectral randomization methods

Post-print/Accepted manuscript

Helene H. Wagner

Stéphane Dray

Wagner, H. H. and Dray, S. (2015), Generating spatially constrained null models for irregularly spaced data using Moran spectral randomization methods. *Methods Ecol Evol*, 6: 1169–1178. doi:10.1111/2041-210X.12407

---

This is the peer reviewed version of the following article: Wagner, H. H. and Dray, S. (2015), Generating spatially constrained null models for irregularly spaced data using Moran spectral randomization methods. *Methods Ecol Evol*, 6: 1169–1178, which has been published in final form at [doi:10.1111/2041-210X.12407](https://doi.org/10.1111/2041-210X.12407). This article may be used for non-commercial purposes in accordance with [Wiley Terms and Conditions for Self-Archiving](#).

### HOW TO CITE TSPACE ITEMS

**Always cite the published version**, so the author(s) will receive recognition through services that track citation counts, e.g. Scopus. If you need to cite the page number of the TSpace version (original manuscript or accepted manuscript) because you cannot access the published version, then cite the TSpace version **in addition to** the published version using the permanent URI (handle) found on the record page.

1 **Generating spatially-constrained null models for irregularly spaced**  
2 **data using Moran spectral randomization methods**

3 Helene H. Wagner<sup>1</sup> and Stéphane Dray<sup>2</sup>

4

5

6 <sup>1</sup> Department of Ecology and Evolutionary Biology, University of Toronto, 3359

7 Mississauga Road, Mississauga, ON, Canada L5L 1C6. Phone: +1-905-569-4702, email address:

8 [helene.wagner@utoronto.ca](mailto:helene.wagner@utoronto.ca)

9 <sup>2</sup> Université de Lyon, F-69000, Lyon ; Université Lyon 1 ; CNRS, UMR5558, Laboratoire

10 de Biométrie et Biologie Evolutive, F-69622, Villeurbanne, France.

11

12 **Word count**

13 Abstract (limit: 350 words): **348**

14 Text (incl. references, tables, figure captions, without Supporting Materials;

15 max. 6000 – 7000 words): **6962**

16

17 **Abstract**

- 18 1. Spatial autocorrelation jeopardizes the validity of statistical inference, e.g., correlation and  
19 regression analysis. Restricted randomization methods can account for the effect of spatial  
20 autocorrelation in the observed data by building it into an empirical null model for hypothesis  
21 testing. This can be achieved e.g. based on conditional simulation, which fits a highly  
22 parameterized geostatistical model to the observed spatial structure, or, for data observed on a  
23 regular transect or grid, with Fourier spectral randomization methods that can flexibly model  
24 spatial structure at any scale. This paper uses Moran eigenvector maps to extend spectral  
25 randomization to irregularly spaced samples.
- 26 2. We present different algorithms to perform restricted randomization to suit different types of  
27 research questions: individual randomization of each variable, joint randomization of a group  
28 of variables while keeping within-group correlations fixed, and randomization with a fixed  
29 correlation between original data and randomized replicates (e.g., as input for simulation  
30 studies). The performance of the proposed Moran spectral randomization methods for  
31 regularly and irregularly spaced samples is assessed with correlation analysis of simulated  
32 data.
- 33 3. Moran spectral randomization closely matched the spatial structure of original simulated data  
34 sets, with identical or nearly identical Moran's  $I$  values and power spectra, depending on the  
35 algorithm. In correlation analysis of two stationary spatially autocorrelated variables, Moran  
36 spectral randomization produced correct type I error rates for stationary spatial data, even for  
37 very small and highly irregular samples, but was sensitive to linear trend. When one or both

38 variables lacked spatial structure, Moran spectral randomization tests were more conservative  
39 than correlation *t*-tests.

40 4. The proposed Moran spectral randomization method requires a minimum of parameterization  
41 and is able to address multivariate data with spatial structure at multiple scales, with the  
42 option of controlling levels of correlation with the original data. It can provide technically  
43 unlimited numbers of randomizations even for small samples while closely maintaining the  
44 spatial characteristics of uni- or multivariate data at all spatial scales. The method is  
45 applicable for correlation analysis of stationary, autocorrelated spatial or temporal series.  
46 Further research should assess whether the method can be extended to multiple regression  
47 analysis.

#### 48 **Keywords**

49 Spatial autocorrelation, Fourier analysis, Moran eigenvector maps, correlation, time series  
50 analysis

#### 51 **Tweetable Abstract (max. 120 characters)**

52 Moran spectral randomization methods for irregularly spaced, spatially autocorrelated, uni-  
53 and multivariate data.

54

## 55 **Introduction**

56 Ecological data often exhibit spatial autocorrelation, which poses challenges for statistical  
57 inference (Legendre 1993). Spatial autocorrelation may affect parameter estimates, type I error  
58 rates and statistical power of tests, including significance tests for correlation, partial correlation  
59 and regression coefficients that are widely used in ecology (Kühn & Dormann 2012). Methods  
60 for dealing with autocorrelation in correlation inference include modified parametric tests with  
61 adjustments for effective sample size (Dutilleul *et al.* 1993; Dutilleul, Pelletier & Alpargu 2008)  
62 and restricted randomization tests, which preserve the observed spatial autocorrelation and build  
63 it into the null model of the statistical hypothesis test (Fortin & Jacquez 2000; Fortin, Jacquez &  
64 Shipley 2012). For data sampled on a regular grid, this may be achieved by permuting observed  
65 values with toroidal shift methods (Upton & Fingleton 1985), conditional simulation methods  
66 that generate new values with similar spatial structure as the observed data (Fortin & Dale 2005),  
67 or with spectral randomization methods based on Fourier or wavelet analysis (Deblauwe, Kennel  
68 & Coueron 2012).

69       Conditional autoregressive simulation (Cressie 1991) can be applied to irregularly spaced  
70 data, but involves fitting geostatistical models to the observed data and may require substantial  
71 computing time (Fortin & Jacquez 2000). Complications may arise (*i*) if the observed spatial  
72 pattern (e.g., species abundance) results from multiple processes (e.g., dispersal, competition,  
73 predation) acting at different spatial scales, so that the observed data exhibit spatial structure at  
74 multiple scales; (*ii*) when analyzing multivariate data, where it may be necessary to randomize  
75 one set of variables simultaneously, keeping their correlations constant, while randomizing their  
76 correlations with a second set of variables (e.g., to test associations between functional groups of

77 species while keeping associations within groups constant); or *(iii)* when the spatial relationships  
78 do not depend on geographic distance per se but on some definition of adjacency formalized in a  
79 neighbor matrix (e.g., in a step-wise model of gene flow within a network of discrete populations,  
80 or if the data relate to spatial units such as polygons, rather than point locations). The first two  
81 situations may require large numbers of geostatistical parameters to be fitted, whereas in the third  
82 case, distance-based geostatistical modeling may not be appropriate. Furthermore, conditional  
83 simulation methods are not suitable for simulating replicates with a fixed correlation with the  
84 original data (e.g., as needed for simulation studies to test performance of estimation methods in  
85 regression analysis of spatial data; Beale *et al.* 2010) as they control the parameters of the  
86 generating process, not the properties of the observed pattern. This paper provides a new  
87 approach, Moran Spectral Randomization (MSR), to generate spatially-structured random  
88 variables. Compared to existing approaches, MSR has the main advantages that *(i)* it can deal  
89 with irregularly spaced data, *(ii)* it considers a spatial neighbor matrix, rather than geostatistical  
90 modeling, and thus requires a minimum of parameterization, *(iii)* it preserves the multiscale  
91 properties of spatial structures, *(iv)* and it is able to address multivariate data or *(v)* to control  
92 levels of correlation with the original data.

93       Conceptually, MSR is related to Fourier spectral randomization (FSR). FSR is based on  
94 Fourier analysis that decomposes an observed, regularly spaced spatial or temporal series into a  
95 set of orthogonal sinusoids with different frequencies, i.e., spatial or temporal scales (harmonic  
96 regression; Graybill 1976) using discrete Fourier transform (DFT; Cooley & Tukey 1965; Gauss  
97 1866). The power spectrum of the series thus describes how the variance of the data is distributed  
98 over the frequency components into which it may be decomposed, i.e., the squared correlation of

99 each sinusoid with the observed data quantifies the intensity of the observed pattern at the scale  
100 of the sinusoid.

101 In MSR, sinusoids are replaced by Moran eigenvector maps (MEM; Borcard & Legendre  
102 2002; Dray, Legendre & Peres-Neto 2006; Griffith & Peres-Neto 2006; Peres-Neto & Legendre  
103 2010) that can also be applied to irregular samples (Borcard & Legendre 2002). MEM are  
104 obtained through the eigen analysis of a spatial neighbor matrix which provides orthogonal  
105 eigenvectors maximizing the spatial autocorrelation measured by Moran's  $I$  (Moran 1950).  
106 Hence, MEM provides vectors allowing the decomposition of the variance of observed data,  
107 similar to the power spectrum in Fourier analysis.

108 We show that MSR provides technically unlimited numbers of randomizations even for  
109 small samples while maintaining the spatial characteristics of uni- or multivariate data at all  
110 spatial scales. Specifically, we show how to derive (i) replicate autocorrelated patterns under the  
111 null hypothesis of an absence of correlation with observed data, (ii) replicate autocorrelated  
112 multivariate patterns with constant correlation among variables, and (iii) replicate autocorrelated  
113 patterns that have a predefined correlation with observed data, as used in simulation studies (e.g.,  
114 Beale *et al.* 2010). The performance of the proposed Moran spectral randomization methods is  
115 assessed with correlation analysis of simulated data for different sampling designs.

## 116 **Methods**

### 117 *Spectral decomposition of variance and correlation using Moran Eigenvector Maps (MEM)*

118 Spatial component regression (SCR; Wagner 2013) uses MEM to decompose the correlation  $r_{xy}$   
119 between two variables  $\mathbf{x}$  and  $\mathbf{y}$ , observed at the same  $n$  sampling locations that may be regularly



120 or irregularly spaced, by spatial eigenvectors. It defines the correlation  $r_{xy}$  as the cross product of  
 121 the vectors  $\mathbf{r}_{xV}$  and  $\mathbf{r}_{yV}$  that contain the correlation coefficients between  $\mathbf{x}$  or  $\mathbf{y}$  and each column in  
 122 a matrix  $\mathbf{V}$  that is defined as a set of orthogonal and uncorrelated (Rodgers, Nicewander &  
 123 Toothaker 1984) spatial eigenvectors obtained by MEM. In matrix notation:

$$124 \quad r_{xy} = \mathbf{r}_{xV}^T \mathbf{r}_{yV}, \quad (1)$$

125 where  $^T$  indicates the transpose. The columns in  $\mathbf{V}$  are the eigenvectors of a symmetric and  
 126 doubly centered spatial weight matrix  $\mathbf{W}$  (Dray, Legendre & Peres-Neto 2006). Matrix  $\mathbf{W}$  is  
 127 constructed by first defining a neighbor matrix of size  $n \times n$  that contains values of one for pairs  
 128 of observations  $i$  and  $j$  if  $j$  is considered a neighbor of  $i$ , and zero otherwise. Each neighbor  $j$  of  
 129 observation  $i$  is then assigned a spatial weight, which may be binary or e.g. a function of the  
 130 inverse geographic distance between  $i$  and  $j$ , and weights may optionally be row-standardized.  
 131 Note that spatial eigenvectors in  $\mathbf{V}$  are defined as a function of  $\mathbf{W}$  only, without reference to the  
 132 values of variables  $\mathbf{x}$  or  $\mathbf{y}$  observed at the sampling locations, and  $\mathbf{V}$  is thus the same for  $\mathbf{x}$  and  $\mathbf{y}$ .  
 133 MEM results in a matrix  $\mathbf{V}$  where the spatial eigenvectors are already sorted from the largest  
 134 scale to the finest scale.

135 A set of  $n$  observations with a full-rank spatial weights matrix  $\mathbf{W}$  of size  $n \times n$  will result  
 136 in  $n - 1$  orthogonal and uncorrelated (Griffith 2000) eigenvectors  $\mathbf{V}_k$  associated with eigenvalues  
 137  $\lambda_k$ , while a single eigenvector with zero eigenvalue is dropped. Under these conditions, the vector  
 138  $\mathbf{r}_{xV}^2$  of squared correlations of  $\mathbf{x}$  with eigenvectors in  $\mathbf{V}$  forms a power spectrum (i.e., a spectral  
 139 decomposition of the variance of the observed data  $\mathbf{x}$  on the orthonormal basis  $\mathbf{V}$ ):

$$140 \quad r_{xV}^2 = \mathbf{r}_{xV}^T \mathbf{r}_{xV} = 1. \quad (2)$$

141 Note that if  $\mathbf{W}$  is not of full rank, there will be multiple zero eigenvalues and their  
 142 associated eigenvectors may not be uncorrelated. To ensure that eq. 2 holds in this situation, a  
 143 vector of ones is added to the subset of eigenvectors with null eigenvalues, the subspace spanned  
 144 by these eigenvectors is then re-orthogonalized, the unit eigenvector is removed, and the last  
 145 eigenvector in the subset is dropped. This procedure is implemented in the R function  
 146 ‘scores.listw’ provided in Supporting Material 3.

147 The variable  $\mathbf{x}$  can be decomposed on an orthonormal basis  $\mathbf{V}$ , so that  $\mathbf{x}$  can be fully  
 148 recreated from  $\mathbf{r}_{\mathbf{xV}}$ ,  $\mathbf{V}$ , and its mean  $\bar{\mathbf{x}}$  and standard deviation  $s(\mathbf{x})$ :

$$149 \quad \mathbf{x} = \bar{\mathbf{x}} + s(\mathbf{x})(n-1)^{0.5} \mathbf{V}\mathbf{r}_{\mathbf{xV}} \quad (3)$$

150 A key feature of MEM is the additive decomposition of Moran’s  $I$  (Dray, Legendre &  
 151 Peres-Neto 2006; Dray 2011). Eigenvalues  $\lambda_k$  are conveniently rescaled by multiplication with  $h$   
 152  $= n / \sum_{ij} w_{ij}$ , where  $w_{ij}$  are the elements of  $\mathbf{W}$ , so that the rescaled eigenvalue  $m_k = h\lambda_k$  corresponds  
 153 to Moran’s  $I_k$  of eigenvector  $\mathbf{v}_k$ , with  $k = \{1, \dots, n-1\}$ . According to Dray (2011), Moran’s  $I_{\mathbf{x}}$  of  
 154 variable  $\mathbf{x}$  can then be found by:

$$155 \quad I_{\mathbf{x}} = \mathbf{r}_{\mathbf{xV}}^2 \mathbf{T} \mathbf{m}, \quad (4)$$

156 which means that global spatial autocorrelation ( $I_{\mathbf{x}}$ ) of the variable  $\mathbf{x}$  is the sum of Moran’s  
 157  $I$  of all spatial eigenvectors, weighted by the variance each eigenvector explains in the observed  
 158 variable  $\mathbf{x}$  (i.e., weighted by the power spectrum).

### 159 ***Basic algorithm for Moran spectral randomization (MSR)***

160 The spatial structure of  $\mathbf{x}$  can be defined by its global level of autocorrelation ( $I_{\mathbf{x}} = \mathbf{r}_{\mathbf{xV}}^2 \mathbf{T} \mathbf{m}$ , eq. 4)  
 161 and its multiscale decomposition defined by its power spectrum ( $\mathbf{r}_{\mathbf{xV}}^2$ , eq. 2). In this context, MSR

162 aims to find a vector of coefficients  $\mathbf{a} = [a_1, \dots, a_{n-1}]^T$  so that the random variable  $\mathbf{x}_{\text{rand}} = (n-1)^{0.5}$

163  $\mathbf{V}\mathbf{a}$  satisfies different conditions. Squared values in  $\mathbf{a}$  should be a power spectrum so that:

$$164 \quad \sum_{i=1}^{n-1} a_i^2 = 1 \quad (\text{C1})$$

165 To preserve the global level of autocorrelation  $I_x$ , the vector  $\mathbf{a}$  must satisfy equation 4 and

166 this leads to the second condition:

$$167 \quad I_x = \sum_{i=1}^{n-1} a_i^2 m_i = \sum_{i=1}^{n-1} r_{\mathbf{xv}_i}^2 m_i \quad (\text{C2})$$

168

169 Lastly, to preserve the multiscale structure, the vector  $\mathbf{a}$  must satisfy equation 2 and this

170 leads to the third condition:

$$171 \quad a_i^2 = r_{\mathbf{xv}_i}^2 \quad (\text{C3})$$

172

173 We provide different algorithms to find adequate values for  $\mathbf{a}$  satisfying strictly or softly the

174 different conditions. Condition C1 is strictly preserved in all algorithms.

175

176 *The Singleton procedure*

177 Choosing  $|a_i| = r_{\mathbf{xv}_i}$  satisfies both conditions C2 and C3. Hence, the solution consists in

178 randomizing the sign of correlations independently for each single eigenvector (i.e.,  $a_i = \pm r_{\mathbf{xv}_i}$ ).

179 This approach allows also generating multivariate replicates with constant correlation among

180 variables (see below).

181

182 *The Pair procedure*

183 This procedure aims to satisfy strictly condition C2 and softly condition C3. Instead of preserving  
 184 the contribution of each single eigenvector in the multiscale decomposition, we consider pairs of  
 185 eigenvectors  $\mathbf{v}_i$  and  $\mathbf{v}_j$ . This subset is stored by column in the matrix  $\mathbf{V}_k$ . In this case, the condition  
 186 C3 is partially satisfied by preserving the proportion of variance explained by a pair of  
 187 eigenvectors. We obtain the new condition C3':

$$188 \quad R_{\mathbf{xv}_k}^2 = a_i^2 + a_j^2 = r_{\mathbf{xv}_i}^2 + r_{\mathbf{xv}_j}^2 \quad (\text{C3}')$$

189 The condition C3' corresponds to the equation of a circle of radius  $R_{\mathbf{xv}_k} = \sqrt{R_{\mathbf{xv}_k}^2}$ . To preserve the  
 190 level of autocorrelation for the pair of eigenvectors, the condition C2 becomes:

$$191 \quad I_{\mathbf{xv}_k} = a_i^2 m_i + a_j^2 m_j = r_{\mathbf{xv}_i}^2 m_i + r_{\mathbf{xv}_j}^2 m_j \quad (\text{C2}')$$

192 The resolution of the system of equations C2' and C3' leads exactly to the solution provided by  
 193 the singleton procedure. An alternative is to satisfy only the condition C3' that corresponds to the  
 194 equation of a circle. The solutions are obtained by drawing a random angle  $\Phi_{\text{rand}}$  from a uniform  
 195 distribution between  $[0 - 2\pi]$  and then compute the values

$$196 \quad \left\{ a_i = R_{\mathbf{xv}_k} \cos(\Phi_{\text{rand}}); a_j = R_{\mathbf{xv}_k} \sin(\Phi_{\text{rand}}) \right\}. \text{ Note that this method does not completely preserve}$$

197 the global autocorrelation values (condition C2' and C2). However, it will be considered as a  
 198 valuable alternative as it can be easily extended to generate multivariate replicates with constant  
 199 correlation among variables or univariate replicates with fixed correlation with observed data (see  
 200 below).

201

202 *The Triplet procedure*

203 This procedure aims to satisfy strictly condition C2 and softly condition C3. In this case, the  
 204 multiscale decomposition is preserved for triplets of eigenvectors  $\mathbf{v}_i$ ,  $\mathbf{v}_j$  and  $\mathbf{v}_l$  stored by column in  
 205 matrix  $\mathbf{V}_k$ . We thus obtain the new condition C3'':

$$206 \quad R_{\mathbf{xv}_k}^2 = a_i^2 + a_j^2 + a_l^2 = r_{\mathbf{xv}_i}^2 + r_{\mathbf{xv}_j}^2 + r_{\mathbf{xv}_l}^2 \quad (\text{C3}'')$$

207 The condition C3'' corresponds to the equation of a sphere of radius  $R_{\mathbf{xv}_k}$ . To preserve the level of  
 208 autocorrelation for the triplet of eigenvectors, the condition C2 becomes:

$$209 \quad I_{\mathbf{xv}_k} = a_i^2 m_i + a_j^2 m_j + a_l^2 m_l = r_{\mathbf{xv}_i}^2 m_i + r_{\mathbf{xv}_j}^2 m_j + r_{\mathbf{xv}_l}^2 m_l \quad (\text{C2}'')$$

210 To satisfy both conditions C2'' and C3'', a first random angle  $\Phi_{\text{rand}}$  is drawn from a  
 211 uniform distribution between  $]0 - 2\pi[$  (bounds excluded to avoid division by zero). Then, a

$$212 \quad \text{second angle is computed by } \Theta_{\text{rand}} = \text{asin} \left( \sqrt{\frac{I_{\mathbf{xv}_k} / R_{\mathbf{xv}_k}^2 - m_l - (m_i - m_l) \sin^2(\Phi_{\text{rand}})}{(m_j - m_i) \sin^2(\Phi_{\text{rand}})}} \right) = \text{asin}(\sqrt{Z}).$$

213 The solutions are then given by

$$214 \quad \left\{ a_i = \pm R_{\mathbf{xv}_k} \cos(\Theta_{\text{rand}}) \sin(\Phi_{\text{rand}}); a_j = \pm R_{\mathbf{xv}_k} \sin(\Theta_{\text{rand}}) \sin(\Phi_{\text{rand}}); a_l = \pm R_{\mathbf{xv}_k} \cos(\Phi_{\text{rand}}) \right\}.$$

215

### 216 *Practical implementation for Pair and Triplet procedures*

217 MSR consists in randomly redistributing the variance of  $\mathbf{x}$  within blocks of spatial eigenvectors  
 218  $\mathbf{V}_k$ . The following steps are performed:

- 219 1. Divide the  $n-1$  eigenvectors in  $K$  sets  $\mathbf{V}_k$  ( $k = \{1, \dots, K\}$ ) (see below for details). If the  
 220 number of eigenvectors is not a multiple of 2 (Pair) or 3 (Triplet), there will be one  
 221 incomplete block  $\mathbf{V}_k$ , whose (randomly selected) eigenvectors are treated by the singleton  
 222 procedure.

- 223 2. Repeat for each complete subset  $\mathbf{V}_k$ .
- 224 2.1. Determine the pooled variance in  $\mathbf{x}$  explained by the eigenvectors ( $R_{\mathbf{xv}_k}^2$ ), the
- 225 associated autocorrelation ( $I_{\mathbf{xv}_k}$ ) and (for Pair) the observed angle  $\Phi_{\mathbf{x}.k} = \text{atan2}(r_{\mathbf{xv}_j}, r_{\mathbf{xv}_i})$ .
- 226 2.2. Sample the angle  $\Phi_{\text{rand}}$ .
- 227 2.3. For the triplet procedure, compute the second angle  $\Theta_{\text{rand}} = \text{asin}(\sqrt{Z})$ . If the value
- 228 of  $\Phi_{\text{rand}}$  leads to a value of  $Z$  not comprised in  $[0-1]$ , then step 2.2. is repeated. If this
- 229 existing condition is not satisfied after a number of trials (e.g., 100), the Singleton
- 230 procedure is used for each eigenvector in the block  $\mathbf{V}_k$ .
- 231 2.4. Determine the coefficients  $a_i, a_j$  for the Pair procedure, as

232  $\left\{ a_i = R_{\mathbf{xv}_k} \cos(\Phi_{\text{rand}}); a_j = R_{\mathbf{xv}_k} \sin(\Phi_{\text{rand}}) \right\}$ , or  $a_i, a_j$  and  $a_l$  for the Triplet procedure, as

233  $\left\{ a_i = \pm R_{\mathbf{xv}_k} \cos(\Theta_{\text{rand}}) \sin(\Phi_{\text{rand}}); a_j = \pm R_{\mathbf{xv}_k} \sin(\Theta_{\text{rand}}) \sin(\Phi_{\text{rand}}); a_l = \pm R_{\mathbf{xv}_k} \cos(\Phi_{\text{rand}}) \right\}$

- 234 3. Optional: Compute the new random variable  $\mathbf{x}_{\text{rand}} = \bar{\mathbf{x}} + s(\mathbf{x})(n-1)^{0.5} \mathbf{V}_{\mathbf{r}_{\mathbf{xv}}}$ .

235

236 Two or more observed variables may be randomized separately by sampling  $\Phi_{\text{rand}}$

237 independently for each variable and for each subset  $\mathbf{V}_k$ . This basic procedure will randomize the

238 correlation between replicates of different variables.

239 Subsets  $\mathbf{V}_k$  (step 1) can be defined randomly or by selecting consecutive eigenvectors. As

240 Triplet preserves values of Moran's  $I_{\mathbf{x}}$  (condition C2''), both procedures can be used but

241 empirical experiments showed that the selection of consecutive eigenvectors often produces

242 undetermined results in step 2.3. (i.e.,  $Z \notin [0-1]$ ) and thus leads to the use of the singleton

243 procedure. Hence, the random assignment of eigenvectors in subsets should be preferred with the  
 244 Triplet procedure. Here, we restricted random assignment to triplets with all positive or all  
 245 negative eigenvalues.

246 Pair procedure does not preserve the global level of autocorrelation (i.e.,  $I_{\mathbf{x}} \neq I_{\mathbf{x}_{rand}}$ ). As the  
 247 spatial eigenvectors in  $\mathbf{V}$  are automatically sorted by decreasing Moran components, i.e., rescaled  
 248 eigenvalues  $m_k$ , pairs of consecutive eigenvectors will have similar autocorrelation. In this case,  
 249 the use of subsets of consecutive eigenvectors should be preferred as it ensures that the  
 250 randomized replicate  $\mathbf{x}_{rand}$  will have very similar spatial structure as the original variable  $\mathbf{x}$  (i.e.,  
 251  $I_{\mathbf{x}} \approx I_{\mathbf{x}_{rand}}$ ). Supporting Material 1, Fig. S1.1 illustrates the effect of shifting variance between  
 252 consecutive spatial eigenvectors.

### 253 ***Generating multivariate replicates with constant correlation among variables***

254 To jointly randomize  $m$  variables, i.e., a data matrix  $\mathbf{X}$  with  $n$  rows and  $m$  columns, while  
 255 maintaining their correlations across all spatial scales, Singleton or Pair procedures may be used.  
 256 For the Singleton procedure, the sign for correlations with each eigenvector is randomized once  
 257 for all  $m$  variables, so that their signs either all change or all remain unchanged. For Pair, a single  
 258  $\Phi_{rand}$  is sampled independently for each subset  $\mathbf{V}_k$  and applied to all  $m$  variables. Then, step 2.4 is  
 259 performed for each variable  $\mathbf{x}$  using  $\Phi_{rand} + \Phi_{\mathbf{x},k}$  instead of  $\Phi_{rand}$ :

260  $\left\{ a_i = R_{x_{v_k}} \cos(\Phi_{\mathbf{x},k} + \Phi_{rand}); a_j = R_{x_{v_k}} \sin(\Phi_{\mathbf{x},k} + \Phi_{rand}) \right\}$ . See Supporting Material 2 for a

261 mathematical proof.

**262 *Generating replicates with fixed correlation with observed data***

263 To obtain a replicate  $\mathbf{x}_{\text{rand}}$  with similar spatial structure as the observed variable  $\mathbf{x}$  and a  
264 fixed correlation  $r_{\text{fix}}$  with  $\mathbf{x}$ , the Pair method should be used with  $\Phi_{\text{rand}} = \text{acos}(r_{\text{fix}})$ , the same for all  
265  $k$ , and step 2.4 should be performed using  $\Phi_{x,k} \pm \Phi_{\text{rand}}$ . The sign needs to be randomized  
266 independently for each subset  $\mathbf{V}_k$ , i.e., each pair of consecutive spatial eigenvectors. The  
267 effectiveness of this procedure to obtain multiple non-identical replicates is investigated in the  
268 simulation study below.

**269 *Illustration with simulated data***

270 All simulations and analyses were performed in R 3.1.1 (R Core Team 2014). The development  
271 version of the ‘adespatial’ package available at [https://r-forge.r-project.org/R/?group\\_id=195](https://r-forge.r-project.org/R/?group_id=195)  
272 provides general-use functions to implement the proposed Moran spectral randomization  
273 procedures. Supporting Material 3 contains a customized version of these functions that can be  
274 used with the R code in Supporting Material 4 to reproduce the simulation study.

275 To illustrate the proposed randomization algorithms, we simulated five standardized  
276 variables with different spatial patterns on the same grid with  $40 \times 40$  cells (Supporting Material  
277 1, Fig. S1.1, top row). X0 served as baseline without spatial pattern and was sampled randomly  
278 from a standard normal distribution using R function ‘rnorm’. X1 and X2 each represent  
279 stationary spatial patterns at small (X1) and large scale (X2). The two variables were generated  
280 with the function ‘RFsimulate’ of the R package ‘RandomFields’ using an isotropic exponential  
281 variogram model with variance = 5 and nugget = 0 for both variables, and with scale = 1 for X1  
282 and scale = 3 for X2. Variable X3, which has a more complex spatial structure, was derived as



283  $X_3 = X_0 + X_1 + X_2$  and is thus correlated with these variables.  $X_4$  represents the same spatial  
284 pattern as  $X_1$  but with an added linear trend along the y-axis, violating the assumption of  
285 stationarity.

286 The full sample ( $n = 1600$ ) was subsampled in different ways, resulting in ten different  
287 samples per simulated data set: a full sample ('Full 1600'); a regular sampling design ('Regular  
288 400'), obtained by discarding every second row and column of the full sample; and a random  
289 sample of the same size ('Random 400'). The remaining sampling designs were chosen to  
290 illustrate the effect of stronger irregularity in the spacing of observations, as well as a further  
291 reduction in sample size. The 70 sampling locations of the well-known oribatid mite data set  
292 (Borcard, Legendre & Drapeau 1992; Borcard & Legendre 1994), sampled within an original  
293 extent of  $2.5 \times 10$  m, were mapped onto a  $40 \times 13$  grid, resulting in sampling design 'Orib 70',  
294 covering one third of the extent of the full sample. The 13 levels of x-coordinates in 'Orib 70'  
295 were repeated three times to obtain sample 'Orib 210' with  $n = 210$ , covering the extent of the  
296 full sample. To further explore the behavior of Moran spectral randomization for small and  
297 highly irregularly-spaced samples, we subsampled 'Orib 70' in five different ways to a sample  
298 size of 32 – 35: a random subsample ('Random 35'); an artificially spaced out sample, deleting  
299 every second level of y coordinates ('Spaced 32'); a clumped sample, retaining only y-  
300 coordinates  $< 23$  ('Clumped 35'); a long-stretched sample, retaining only x-coordinates  $< 8$   
301 ('Long 34'); and a bi-modally clustered sample, retaining only y-coordinates that were either  $<$   
302  $13$  or  $> 30$  ('Bimodal 34').

303 For the 'Full 1600' and 'Regular 400' sampling designs, we applied queen's case neighbor  
304 definition, resulting in an average of 5.06 and 4.44 neighbors per sampling location, respectively.

305 For all eight irregular sampling designs, we defined neighbors with Gabriel graphs (Supporting  
306 Material, Fig. S1.2), resulting in averages of 3.37 – 4.06 neighbors per sampling location. For all  
307 sampling designs,  $\mathbf{W}$  was defined using inverse distance weights followed by row-  
308 standardization. This means that nearby neighbors receive more weight than more distant ones,  
309 and the weights of all neighbors sum to one for each sampling location.

### 310 *Evaluation of method performance*

#### 311 *Statistical properties of replicates*

312 Replicates of spatially autocorrelated variables should mimic the spatial pattern of the original  
313 data at all spatial scales but be statistically independent of these. When plotted in space, high and  
314 low values should thus show similar spacing but their physical locations should be randomized.  
315 In statistical terms, MSR replicates should have the same overall spatial autocorrelation (Moran's  
316  $I$ ) and the same multi-scale spatial structure (power spectrum) as the original variables but, on  
317 average, be uncorrelated with the original data, with the same distribution of empirical correlation  
318 coefficients as expected from independently simulated data.

319 To compare these statistical properties between MSR procedures, we simulated a data set  
320 with variables X0 – X4, as defined above, and subsampled it for each of the ten sampling  
321 designs. For each combination of variable and sampling design, we generated 30 replicates each  
322 with the following methods: Singleton, Pair, Triplet, and, for benchmarking, fully independent  
323 replicates generated by simulating 30 additional data sets ('Null'). We determined for each  
324 combination of simulated variable  $\mathbf{x}$ , sampling design and procedure: (i) the correlation of  
325 replicates with the original variable (type I error rate of a  $t$ -test for the correlation of  $\mathbf{x}$  with its

326 replicates); (ii) the preservation of global autocorrelation  $I_x$  and of the power spectrum (mean  
327 correlation between the power spectrum  $\mathbf{r}_{xv}^2$  of  $\mathbf{x}$  and the power spectra of its replicates); and (iii)  
328 for the pairs method only, the bias and precision of Moran's  $I$  of replicates compared to Moran's  
329  $I$  of  $\mathbf{x}$ . Bias in Moran's  $I$  was independent of the magnitude of Moran's  $I$  (controlling for  
330 sampling design), hence absolute, not relative deviations were assessed. The entire simulation  
331 was repeated 100 times and values averaged across trials.

### 332 *Spectral randomization tests for correlation*

333 A randomization test for the correlation between two spatial variables sampled from  
334 uncorrelated populations should have a correct type I error rate  $\alpha$ , i.e., if  $\alpha = 0.05$ , the null  
335 hypothesis of no correlation should be rejected in 5% of independent cases (where the expected  
336 correlation is zero), and high power to detect linear dependence (where the expected correlation  
337 different from zero). To assess the performance of MSR in significance tests of correlation, we  
338 ran 5000 simulations and estimated type I error rates from independent variable pairs, and power  
339 from correlated variable pairs. For each simulation run, we generated two data sets,  $X_{0_1} - X_{4_1}$   
340 and  $X_{0_2} - X_{4_2}$ , where each pair of corresponding variables (e.g.,  $X_{2_1}$  and  $X_{2_2}$ ) was simulated  
341 independently with the geostatistical parameters defined above and thus had the same expected  
342 spatial autocorrelation structure, though empirical values of Moran's  $I$  may differ. The expected  
343 correlation between any two variables from different data sets was zero (linear independence),  
344 whereas the expected correlation between  $X_3$  and  $X_0$ ,  $X_1$ , or  $X_2$ , or between  $X_4$  and  $X_1$ , from  
345 the same data set was non-zero (linear dependence). For each sampling design and variable pair,  
346 we tested the correlation with four different methods: correlation  $t$ -test (using R function

347 ‘cor.test’) and three MSR randomization tests with 199 replicates each, using Pair, Triplet  
348 (random triplets), and Singleton procedures.

### 349 *Spectral randomization with additional constraints*

350 To illustrate the implementation of further constraints, we simulated 100 data sets and created 30  
351 replicates per data set with the following algorithms: joint MSR of all variables X0 – X4 to  
352 maintain their cross-correlations, using Singleton and Pair methods, and MSR with  $r_{\text{fix}} = \{0, 0.3, -$   
353  $0.5, 0.7\}$  to obtain replicates that have a fixed correlation with the original variables. We  
354 evaluated the distribution of correlation coefficients between the original variables and their  
355 replicates, and the distribution of correlation coefficients among replicates.

## 356 **Results**

### 357 *Statistical properties of replicates*

358 For stationary patterns X0 – X3, MSR methods created patterns that were visually comparable to  
359 the original data but differed in the locations of low and high values (Supporting Material 1, Fig.  
360 S1.3). For X4, a non-stationary pattern with linear trend along the y-axis, replicates included a  
361 randomization of the trend angle (Supporting Material 1, Fig. S1.3, bottom). Pairs randomized  
362 trend direction more flexibly than Triplet and Singleton methods (Supporting Material 1, Fig.  
363 S1.4).

364 MSR replicates of a random normal variable (X0) were uncorrelated on average (mean = 0)  
365 with the original variable, but the distribution of correlations had higher variance than expected  
366 from true random variables (not shown). This is reflected in higher rates of statistically

367 significant correlations between each original variable  $\mathbf{x}$  and its replicates, based on a parametric  
368  $t$ -test for regression coefficients (Fig. 1). On average, correlations with  $\mathbf{x}$  were highest for the  
369 Singleton method, lower for Triplet, and lowest for Pair. For variable X2 with stationary, large-  
370 scale spatial autocorrelation, replicates of all methods showed increased rates of significant  
371 correlations with  $\mathbf{x}$ , with differences between sampling designs that appear to be related to  
372 differences in average Moran's  $I$  (Supporting Material, Fig. S1.5).

373 Replicate simulations of the data generating process ('Null') showed high variability in  
374 their global Moran's  $I$  (Fig. 2, left: standard deviation rescaled by multiplication with  $(n - 1)^{0.5}$  to  
375 account for sample size) and power spectra (multi-scale spatial structure), as indicated by a low  
376 correlation between the power spectrum of  $\mathbf{x}$  and those of its replicates (Fig. 2, right). In contrast,  
377 the Singleton method completely preserved both Moran's  $I$  and power spectra. The Triplet  
378 method preserved Moran's  $I$  but resulted in some variation in the power spectrum, and the Pair  
379 method showed variation both in Moran's  $I$  and the power spectrum.

380 The Pair method does not strictly preserve Moran's  $I$  but its MSR replicates were generally  
381 unbiased for the random normal variable X0 (Supporting Material, Fig. S1.6). For the spatially  
382 autocorrelated variable X2, bias was negligible for large data sets but Moran's  $I$  of replicates  
383 showed a slight negative bias for the small sampling designs.

#### 384 *Spectral randomization tests for correlation*

385 When correlating two random normal variables X0<sub>1</sub> and X0<sub>2</sub>, or one random normal variable (X0)  
386 and one stationary, spatially autocorrelated variable (X1, X2 or X3), the parametric  $t$ -test  
387 produced correct type I error rates. Specifically, the empirical type I error rates fell within the  
388 95% confidence interval for a binomial distribution  $\text{bin}(n, p)$  with  $n = 5000$  simulated data sets

389 and  $p = \alpha = 0.05$  (Fig. 3 A, B). In these situations, all three MSR procedures (especially the  
390 Triplet method) showed slightly deflated type I error rates, making tests more conservative.

391 When correlating any two spatially autocorrelated variables ( $X_1$ ,  $X_2$ ,  $X_3$  or  $X_4$ ), the  $t$ -test  
392 showed high inflated type I error rates. MSR methods maintained correct or slightly deflated type  
393 I error rates for correlations among stationary variables ( $X_1$ ,  $X_2$  or  $X_3$ ), but showed considerably  
394 deflated error rates for the correlation between  $X_4$  and  $X_1$ ,  $X_2$  or  $X_3$  (Fig. 3 C, D). Error rates  
395 were independent of the observed level of Moran's  $I$  of the randomized variable. When  
396 correlating two non-stationary variables  $X_{4_1}$  and  $X_{4_2}$ , where both variables included trend along  
397 the  $y$ -axis, MSR tests showed generally lower type I error rates than  $t$ -tests (Fig. 3 E). Especially  
398 the Singleton and Triplet procedures were less affected by non-stationarity.

399 When testing the correlation between  $X_{0_1}$  and  $X_{3_1}$ , where the  $t$ -test was applicable ( $X_{0_1}$   
400 being a random normal), all three MSR methods had slightly lower statistical power to detect  
401 linear dependence than the  $t$ -test (Fig. 4). Averaged across the five smallest sampling designs,  
402 Pair reached approximately 98% of the power of the  $t$ -test, Triplet 97%, and Singleton 96%. This  
403 order remained constant for all other types of variable pairs with linear dependence, where the  $t$ -  
404 test was not applicable (results not shown).

#### 405 ***Spectral randomization with additional constraints***

406 Correlations between variable pairs with linear dependence ( $X_{3_1}$  vs.  $X_{0_1}$ ,  $X_{1_1}$ , or  $X_{2_1}$ ;  $X_{4_1}$  vs.  
407  $X_{1_1}$ ) calculated using Equation 1 were identical to Pearson correlation coefficients. When  
408 correlated variables were jointly randomized using the Singleton or Pair procedures, their cross-  
409 correlations were completely preserved (results not shown).

410           When each variable was randomized with the additional constraint of a fixed correlation  
411 with the original variable, using the Pair method, sampling designs with uneven sample size  $n$   
412 ('Random 35', 'Clumped 35') completely preserved the predefined correlation  $r_{\text{fix}}$  (Supporting  
413 Material, Fig. S1.2 A). For even  $n$ , the presence of an incomplete block (where the Singleton  
414 method was used for one randomly selected spatial eigenvector) introduced some variation in the  
415 correlation with the original variable. For large samples ( $n \geq 400$ ), this variation was negligible,  
416 but for small samples ( $n \leq 100$ ), there was considerable variation as well as a negative bias, so  
417 that on average, correlations with the original variable were slightly weaker than specified by  $r_{\text{fix}}$ .

418           For sampling designs with uneven  $n$ , the correlation among replicates generated with the  
419 constraint of  $r_{\text{fix}}$  differed systematically from the correlation among replicates generated without  
420 such a constraint (Supporting Material, Fig. S1.2 B). The mean correlation among replicates was  
421 predictable by  $r_{\text{fix}}^2$ , whereas the standard deviation varied with  $r_{\text{fix}}$  and with the level of spatial  
422 autocorrelation. The factor by which the standard deviation of the correlation among replicates  
423 was inflated, compared to replicates without the constraint of  $r_{\text{fix}}$ , followed  $2^{0.5}(1 - r_{\text{fix}}^2)$   
424 (Supporting Material, Fig. S1.2 C).

## 425 **Discussion**

### 426 *Comparisons of Moran spectral randomization procedures*

427 This paper presents algorithms for restricted randomization of irregularly spaced data with Moran  
428 spectral randomization that can be applied to a wide range of sampling designs for which  
429 restricted randomization methods were limited so far. Moran spectral randomization provides a  
430 framework to build spatially constrained null models allowing statistical inference in the presence

431 of spatial autocorrelation. While Moran spectral randomization has been developed here for two-  
432 dimensional spatial data, it is also applicable to one-dimensional transect data or time series  
433 analysis. The method produces a technically unlimited number of non-identical replicates even  
434 for small samples (except for the Singleton method with  $2^{(n-1)}$  unique replicates), without  
435 requiring additional parameters beyond the explicit definition of neighbors and spatial weights in  
436 matrix  $\mathbf{W}$ .

437         The MSR replicates mimic the spatial characteristics of the original variables in terms of  
438 global autocorrelation and multiscale patterns. These properties allow controlling the type I error  
439 rate of statistical tests (e.g., bivariate correlation) in the presence of autocorrelation. Ideally,  
440 replicates should also be uncorrelated to the original variable to ensure power. These two  
441 objectives (similar spatial structure and independence with the original variables) are antagonist  
442 and thus defined a gradient from considering the original variable (spatial properties fully  
443 preserved but no independence) to its full randomization (complete independence with the  
444 original variable but no spatial constraint). Hence, we provide different procedures that  
445 correspond to different trade-offs between these two extremes.

446         The Singleton procedure imposes the strongest spatial constraints and thus produces  
447 replicates with the highest degree of correlation with the original variable. It performs  
448 surprisingly well in randomization tests, even for small spatial samples, despite the limitation of  
449  $2^{(n-1)}$  unique replicates. The method is simple to implement and can also be used to jointly  
450 randomize a group of variables. The Singleton method should be preferred when it is important to  
451 strictly preserve the spatial characteristics of the original variables even if this means that its  
452 replicates are more correlated with the original variable.



453           The Pair procedure is the least strict concerning the spatial constraints and thus provides  
454 MSR replicates with the lowest degree of correlation with the original variables. Spatial  
455 characteristics are quite well preserved and thus it appears as good compromise and allows for  
456 joint randomization as well as randomization with fixed correlation. The Pair procedure should  
457 be used when it is considered important to randomize correlation at each spatial eigenvector and  
458 reduce overall correlation with the original variable, or when the slightly lower statistical power  
459 or the limited number of  $2^{n-1}$  unique replicates of the Singleton method are of concern.

460           Lastly, the Triplet procedure that is supposed to combine the advantages of singleton and  
461 pairs did not perform as well as the other methods and could also be time-consuming (due to the  
462 existing conditions is step 2.3). Hence, it is not generally recommended in future work.

#### 463 ***Performance of Moran spectral randomization in correlation analysis***

464 Moran spectral randomization replicates preserved Moran's  $I$  of the original variable either  
465 perfectly (Singleton, Triplet methods) or approximately (Pair). In contrast, replicate simulations  
466 of the generating process (i.e., conditional simulation with known geostatistical parameters)  
467 resulted in considerable variation of Moran's  $I$  among replicates (Fig. 2, left). The marked  
468 quantitative difference illustrates an important conceptual distinction: spectral randomization  
469 methods randomize the observed pattern, whereas conditional simulation simulates independent  
470 outcomes of the underlying process, assuming that it has been correctly parameterized, e.g.,  
471 through geostatistical analysis of the empirical data. Preserving Moran's  $I$  may be advantageous  
472 when using replicates to control for the effect of spatial autocorrelation in the empirical data on  
473 statistical inference, such as correlation or regression analysis. Information on the variability of  
474 Moran's  $I$  given a specified model of the generating process will be most important when the

475 inference concerns the difference between two or more empirically observed patterns (Fortin *et*  
476 *al.* 2003; Remmel & Fortin 2013).

477 Moran randomization tests produced correct, or slightly deflated, type I errors for  
478 correlation analysis, independent of sample size or sample configuration, unless both variables  
479 exhibited non-stationarity, here in the form of linear trend along the y-axis (variable X4).  
480 Likewise, Fourier spectral randomization is known to be sensitive to trend. Recently, Deblauwe,  
481 Kennel & Coueron (2012) presented a wavelet-based spectral randomization method (dual-tree  
482 complex wavelet transform; DT-CWT) that proved more robust than DFT in correlation  
483 inference. Wavelet analysis has also been used to control spatial autocorrelation in regression  
484 analysis of gridded data (Carl & Kühn 2008), but such wavelet methods are not available for  
485 irregularly spaced data. Non-stationarity, which can occur in many different flavors, remains an  
486 unsolved problem especially in spatial regression analysis (Beale *et al.* 2010). Further research is  
487 needed to assess the performance of Moran spectral randomization in spatial regression analysis  
488 when stationarity assumptions are met and its sensitivity to various types of non-stationarity.

#### 489 ***Flexible algorithms for restricted randomization of irregularly spaced spatial data***

490 The proposed algorithms are applicable to a wide range of research questions and hypotheses. We  
491 presented MSR in the case of spatial autocorrelation but the method can also be used in the case  
492 of temporal or phylogenetic dependence (Peres-Neto 2006). Bivariate correlation analysis can be  
493 achieved by randomizing one variable, or randomizing both variables independently (Lennon  
494 2000). The possibility to jointly randomize several variables while maintaining their cross-  
495 correlations will be especially interesting in ecological community analysis that aims at assessing  
496 community interactions. Different groups of species (e.g., functional groups) may thus be

497 randomized separately, maintaining empirical species associations within groups but breaking  
498 them between groups. Molecular genetic analysis provides another example where alleles of the  
499 same locus may need to be randomized jointly but independently of other loci. While Moran  
500 spectral randomization can control correlations among jointly randomized variables even for non-  
501 normal data including binary variables, the randomized variables may show a different statistical  
502 distribution than the observed variables. Depending on the purpose of randomization, an iterative  
503 approach following the steps in iterative amplitude adjusted Fourier transform may be needed  
504 (Deblauwe, Kennel & Couteron 2012; Schreiber & Schmitz 1996), where the simulated values  
505 are replaced by the observed values according to their rank (though this may alter correlations  
506 between variables). Further research should focus on application of MSR to multivariate  
507 abundance or presence-absence data.

508       The Pair procedure can be used to generate replicates with similar multi-scale spatial  
509 structure (approximate preservation of Moran's  $I$  and power spectrum) and a predefined  
510 correlation  $r_{\text{fix}}$  with the original variable. This will be useful in simulation studies aimed e.g. at  
511 testing the performance of spatial regression methods in the presence of correlation among  
512 predictor variables (multi-collinearity; Beale *et al.* 2010; Dormann *et al.* 2013). In order to  
513 completely maintain  $r_{\text{fix}}$ , an uneven sample size  $n$  should be used, though the variation for even  $n$   
514 will be negligible for larger samples.

## 515 **Conclusion**

516 Moran spectral randomization allows to efficiently use regularly or irregularly spaced spatial data  
517 for assessing the correlation between two observed variables. It provides null distributions that  
518 explicitly incorporate the observed autocorrelation at all spatial scales without the need for

519 geostatistical parameter estimation as required for conditional simulation. Instead, Moran spectral  
520 randomization relies on the definition of spatial relationships in a spatial weights matrix  $\mathbf{W}$ , and  
521 the implications of the choice of  $\mathbf{W}$  on correlation and regression analysis with various methods,  
522 including those based on MEM, are not well understood (Stakhovych & Bijmolt 2009).

523 In the limited scope of simulations presented here, focusing on correlation analysis under  
524 stationary conditions, Moran spectral randomization produced correct or slightly conservative  
525 type I error rates and high statistical power (96% - 98% of the power of a parametric  $t$ -test). MSR  
526 thus performed well, even for very small and highly irregularly spaced samples. If the method  
527 can be extended to multiple regression analysis, Moran spectral randomization may provide a  
528 new avenue for the problem of correctly accounting for spatial autocorrelation in species  
529 distribution modeling when the parameters of the underlying spatial ecological process  
530 generating the spatial patterns are unknown (e.g.: Beale *et al.* 2010; Kühn & Dormann 2012).

### 531 **Acknowledgments**

532 This research was supported by a NSERC Discovery Grant to HHW. We thank Sucharita Ghosh  
533 and Erin Landguth for important discussions, and Marie-Josée Fortin, Kamran Safi, Shekhar  
534 Biwas, Guillaume Blanchet and two anonymous reviewers for helpful comments on earlier  
535 versions of the manuscript.

### 536 **Data Accessibility**

537 All necessary R code to reproduce the data is provided in Supporting Materials 3 & 4.

538 **References**

- 539 Beale, C.M., Lennon, J.J., Yearsley, J.M., Brewer, M.J. & Elston, D.A. (2010). Regression  
540 analysis of spatial data. *Ecology Letters*, **13**, 246-264.
- 541 Borcard, D. & Legendre, P. (1994). Environmental control and spatial structure in ecological  
542 communities: an example using oribatid mites (Acari, Oribatei). *Environmental and*  
543 *Ecological Statistics*, **1**, 37-61.
- 544 Borcard, D. & Legendre, P. (2002). All-scale spatial analysis of ecological data by means of  
545 principal coordinates of neighbour matrices. *Ecological Modelling*, **153**, 51-68.
- 546 Borcard, D., Legendre, P. & Drapeau, P. (1992). Partialling out the Spatial Component of  
547 Ecological Variation. *Ecology*, **73**, 1045-1055.
- 548 Carl, G. & Kühn, I. (2008). Analyzing spatial ecological data using linear regression and wavelet  
549 analysis. *Stochastic Environmental Research and Risk Assessment*, **22**, 315-324.
- 550 Cooley, J.W. & Tukey, J.W. (1965). An Algorithm for the Machine Calculation of Complex  
551 Fourier Series. *Mathematics of Computation*, **19**, 297-301.
- 552 Cressie, N. (1991). *Statistics for Spatial Data*. John Wiley & Sons, New York.
- 553 Deblauwe, V., Kennel, P. & Couteron, P. (2012). Testing Pairwise Association between Spatially  
554 Autocorrelated Variables: A New Approach Using Surrogate Lattice Data. *PLoS ONE*, **7**,  
555 e48766.
- 556 Dormann, C.F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G. *et al.* (2013). Collinearity:  
557 a review of methods to deal with it and a simulation study evaluating their performance.  
558 *Ecography*, **36**, 027-046.

- 559 Dray, S. (2011). A New Perspective about Moran's Coefficient: Spatial Autocorrelation as a  
560 Linear Regression Problem. *Geographical Analysis*, **43**, 127-141.
- 561 Dray, S., Legendre, P. & Peres-Neto, P.R. (2006). Spatial modelling: a comprehensive  
562 framework for principal coordinate analysis of neighbour matrices (PCNM). *Ecological*  
563 *Modelling*, **196**, 483-493.
- 564 Dutilleul, P., Clifford, P., Richardson, S. & Hemon, D. (1993). Modifying the t test for assessing  
565 the correlation between two spatial processes. *Biometrics*, **49**, 305-314.
- 566 Dutilleul, P., Pelletier, B. & Alpargu, G. (2008). Modified *F* tests for assessing the multiple  
567 correlation between one spatial process and several others. *Journal of Statistical Planning*  
568 *and Inference*, **138**, 1402.
- 569 Fortin, M.-., Boots, B., Csillag, F. & Rempel, T.K. (2003). On the role of spatial stochastic  
570 models in understanding landscape indices in ecology. *Oikos*, **102**, 203-212.
- 571 Fortin, M. & Dale, M.R.T. (2005). *Spatial Analysis: A Guide for Ecologists*. Cambridge  
572 University Press, New York.
- 573 Fortin, M. & Jacquez, G.M. (2000). Randomization tests and spatially auto-correlated data.  
574 *Bulletin of the Ecological Society of America*, **81**, 201-205.
- 575 Fortin, M., Jacquez, G.M. & Shipley, B. (2012). Computer-intense methods. *Encyclopedia of*  
576 *Environmetrics*, 2<sup>nd</sup> edition (eds A.H. El-Shaarawi & W.W. Piegorsch), pp. 489-493. Wiley,  
577 Chichester.
- 578 Gauss, C.F. (1866). Theoria interpolationis methodo nova tractata. *Carl Friedrich Gauss Werke.*,  
579 pp. 265-327. Königliche Gesellschaft der Wissenschaften, Göttingen.

- 580 Graybill, F.A. (1976). Theory and Application of the Linear Model. Duxbury Press, Belmont,  
581 CA.
- 582 Griffith, D.A. (2000). Eigenfunction properties and approximations of selected incidence  
583 matrices employed in spatial analyses. *Linear Algebra and its Applications*, **321**, 95-112.
- 584 Griffith, D.A. & Peres-Neto, P.R. (2006). Spatial Modeling in Ecology: The Flexibility of  
585 Eigenfunction Spatial Analyses. *Ecology*, **87**, 2603-2613.
- 586 Kühn, I. & Dormann, C.F. (2012). Less than eight (and a half) misconceptions of spatial analysis.  
587 *Journal of Biogeography*, **39**, 995-998.
- 588 Legendre, P. (1993). Spatial autocorrelation: trouble or new paradigm? *Ecology*, **74**, 1659-1673.
- 589 Lennon, J.J. (2000). Red - shifts and red herrings in geographical ecology. *Ecography*, **23**, 101-  
590 113.
- 591 Moran, P.A.P. (1950). Notes on continuous stochastic phenomena. *Biometrika*, **37**, 17-23.
- 592 Peres-Neto, P.R. (2006). A unified strategy for estimating and controlling spatial, temporal and  
593 phylogenetic autocorrelation in ecological models. *Oecologia Brasiliensis*, **10**, 105–119.
- 594 Peres-Neto, P.R. & Legendre, P. (2010). Estimating and controlling for spatial structure in the  
595 study of ecological communities. *Global Ecology and Biogeography*, **19**, 174-184.
- 596 R Core Team (2014). R: A Language and Environment for Statistical Computing. R Foundation  
597 for Statistical Computing, Vienna, Austria.
- 598 Remmel, T. & Fortin, M.-J. (2013). Categorical, class-focused map patterns: characterization and  
599 comparison. *Landscape Ecology*, **28**, 1587-1599.
- 600 Rodgers, J.L., Nicewander, W.A. & Toothaker, L. (1984). Linearly Independent, Orthogonal, and  
601 Uncorrelated Variables. *The American Statistician*, **38**, 133-134.

- 602 Schreiber, T. & Schmitz, A. (1996). Improved Surrogate Data for Nonlinearity Tests. *Physical*  
603 *Review Letters*, **77**, 635-638.
- 604 Stakhovych, S. & Bijmolt, T.H.A. (2009). Specification of spatial models: A simulation study on  
605 weights matrices. *Papers in Regional Science*, **88**, 389-408.
- 606 Upton, G.J.G. & Fingleton, B. (1985). Spatial data analysis by example. Vol. 1: Point pattern and  
607 quantitative data. Wiley, New York.
- 608 Wagner, H.H. (2013). Rethinking the linear regression model for spatial ecological data. *Ecology*,  
609 **94**, 2381-2391.
- 610



611 **Figure captions**

612 Figure 1: Type I error rates for a  $t$ -test of the correlation between each observed variable and its  
613 replicates, for the random normal variable X0 (top row) and variable X2 with stationary, large-  
614 scale spatial autocorrelation. 'Null' refers to fully independent replicates of the generating  
615 process.

616 Figure 2: Spatial properties of replicates for fully independent replicates of the generating  
617 process ('Null') and for MSR replicates generated with Pair, Triplet or Singleton procedures. Left:  
618 bar chart of the mean standard deviation of Moran's  $I$  of replicates, rescaled by  $(n - 1)^{0.5}$  (to  
619 account for sample size) and pooled across sampling designs, separately for X0 (dark grey bars)  
620 and X2 (light gray bars). Note that Triplet and Singleton methods showed zero variation. Right:  
621 bar chart of the mean correlation between the power spectra of the original variable and its  
622 replicates, pooled across sampling designs, for X0 (dark gray bars) and X2 (light gray bars).

623 Figure 3: Boxplots of empirical type I error rates across 5000 simulation runs, for different  
624 types of variable pairs and randomization methods, pooled across all 10 sampling designs. Fig. A  
625 shows the correlation between two random normal variables, Fig. B the correlation between a  
626 random variable X0 and a variable with stationary spatial autocorrelation (X1, X2, or X3), Fig. C  
627 the correlation between any two variables with stationary spatial autocorrelation, Fig. D a  
628 correlation between a variable with non-stationary spatial autocorrelation (X4: linear trend along  
629 y-axis) and any variable with stationary spatial autocorrelation, and Fig. E the correlation  
630 between two variables with non-stationary spatial autocorrelation. The number below each  
631 boxplot reports the mean, and gray polygons indicate a 95% confidence interval for the expected  
632 type I error rate.

633 Fig. 4: Relative power of MSR randomization tests. Each boxplot shows the distribution of  
634 ten empirical power estimates to detect the correlation between variables X0 and X3, using Pair,  
635 Triplet or Singleton procedures, rescaled by the corresponding empirical power estimate of a  
636 parametric correlation  $t$ -test. The ten values for each method were obtained from 5000 simulated  
637 data sets, subsampled with each of the five sampling designs with sample sizes between 32 and  
638 35, randomizing either X0 or X3.

639

640