Spatially-explicit predictions using spatial eigenvector maps

Author 1 Author 2

2024-07-21

Running headline: Spatially-explicit predictions

Abstract

1. In this paper, we explain how to obtain sets of descriptors of the spatial variation, which we call « predictive Moran's Eigenvector Maps » (pMEM), that can be used to make spatially-explicit predictions for any environmental variables, biotic or abiotic. It unites features of a method called « Moran's Eigenvector Maps » (MEM) and spatial interpolation, and produces sets of descriptors that can be used with any other modelling method, such as regressions, support vector machines, regression trees, artificial neural networks, and so on. The pMEM are the predictive eigenvectors produced by using a DWF in the construction of MEMs. Seven types of pMEM, each associated with one of seven different distance weighting functions (DWF), were defined and studied.

2. We performed a simulation study to determine the power of different types of pMEM eigenfunctions at making accurate predictions for spatially-structured variables.

3. We exemplified the application of the method to the prediction of the spatial distribution of 35 Oribatid mite species living in a peat moss (*Sphagnum*) mat on the shore of a Laurentian lake. We also provide an `R` language package called **pMEM** to make calculations easily available to end users.

4. The results indicate that anyone of the pMEMs obtained from the different distance weighting functions could be the best suited one to predict spatial variability in a given data set. Their application to the prediction of mite species distributions highlights the capability of pMEMs for predicting species distributions, and for providing spatially-explicit estimates of environmental variables that are useful for predicting species distributions.

Key-words: space, prediction, interpolation, mapping, Moran's I

Introduction

Spatial analysis hinges on the principle that natural features and conditions are not distributed haphazardly in space, but are organized as a consequence of the processes from which they originate (Forman and Godron 1986; Forman 1995; Legendre 1993). For instance, spatially-structured geological processes affected the sorting of minerals in the earth crust, the latter are eroded at various rates by the action of water, ice, or wind, thereby affecting the distribution of surface and ground waters which, in turn, are driving the distribution of microbes, fungi, plants, and animals at various scales in the landscape. Determining all the relevant natural processes influencing the distribution of ecosystem components in the landscape is often undermined by our lack of the necessary data (e.g., Pascoe et al. 2019; Antunes et al. 2020). Nevertheless, the combined effects of natural processes are readily visible as spatial structures in the form of mosaics of gradients, patches of various sizes, shapes and orientations, and so on. In such circumstances, it is helpful to model feature distribution directly from their spatial structures instead of relying on sparsely available environmental descriptors.

Spatial structuring entails that the probability of making a particular observation at a given location in space is conditional on the values observed at other points around that location. Consequently, it is possible to estimate values of a spatially-structured variables at locations in an area using a set of values of that variable sampled in the same area. Kriging (Matheron 1962) is an interpolation method that can be used for that purpose (Legendre and Fortin 1989; Pebesma 2004). Kriging relies on an estimator of the spatial variation, which is a function of the pairwise distances between locations, in order to weight the surrounding observations before averaging. Alternatively, a method called co-kriging enables one to use data from other observed variables to help predict the value of a variable of interest (Myers 1984). Kriging and co-kriging have long been shown to be useful for making spatially-explicit predictions.

Moran's eigenvector maps (MEM), which were proposed by Dray, Legendre, and Peres-Neto (2006), are sets of latent descriptors used to represent spatial variation in models. MEM provide sets of orthogonal (i.e., linearly independent) variables generated from the pairwise distances among the sampling sites, which are calculated from the spatial distances among the sites, or other types of spatial relationship matrices, describing, for example, the connectivity among the sites. Each of these latent variables, which is called a spatial eigenvector (hereafter referred to as an SEV), has a corresponding eigenvalue, which is related to, and indicates the spatial scale of, the spatial variation it describes. SEVs are used as descriptors of spatial variability in any sort of statistical model suitable to represent single or multiple random (dependent) variable(s), such as (generalized) linear regression, regression trees, gradient boosted trees (Mason et al. 1999; Chen and Guestrin 2016), Bayesian additive regression trees (Chipman, George, and McCulloch 2010), support vector machines

(Cortes and Vapnik 1995), artificial neural networks (Goodfellow, Bengio, and Courville 2016), and so on. MEMs get their name from the Moran's index of spatial autocorrelation (Moran 1950), as there is a simple relationship between the eigenvalue associated with an SEV and Moran's I index calculated for the largest distance class of that SEV. Spatial orthogonal eigenvectors whose eigenvalues were not strict linear functions of Moran's I have also been described, for instances PCNM by Borcard and Legendre (2002), ISOMAP with anisotropic SEV by Mahecha and Schmidtlein (2008), and AEM by Blanchet, Legendre, and Borcard (2008), among other papers (Griffith and Peres-Neto 2006).

From discrete to continuous domain

Each SEV from an MEM can be regarded as the set of values of an underlying spatial eigenfunction (hereafter referred to as an SEF) for the set of sampling sites for which the MEM has been calculated. An SEV originates from a discrete domain, which is a sample of locations meant to represent a population of locations, whereas its corresponding SEF has a continuous domain, and thus bears values for all locations in that population. To our knowledge, no study has explicitly addressed MEM from the perspective of continuous SEFs, rather than discrete, point-defined latent variables (but see, Guénard et al. 2016, 2017; and Guénard and Legendre 2018, for early applications of this idea). However, this aspect of MEM is instrumental in using the suite of spatial patterns described by MEM for spatially-explicit predictive modelling. As such, predictive spatial modelling using MEM opens the way to applying machine learning approaches in situations where spatial variation is important and should be represented in a way that meets the objective of producing spatially-explicit predictions.

Distance weighting

Crucially, all MEM-based SEF share the same calculation basis involving two matrices. The first is a binary connectivity matrix ($\mathrm{B} = [b_{i,j}]$), whose elements take the value 1 when sites i and j are linked together, and the value 0 when they are not linked. The second is a spatial weight matrix ($\mathrm{A} = [a_{i,j}]$), whose elements are pairwise weights calculated from the pairwise between-sites distances using a distance-weighting function (hereafter referred to as a DWF). The different types of SEF differ by the nature and specific parameters of that DWF. For MEM, Dray, Legendre, and Peres-Neto (2006) provided three DWFs (namely the linear, concave up, and concave down DWFs). Besides these, it may be useful to explore other suitable DWFs in order to further our options for SEF. In particular, four of the common variogram functions used for kriging (namely, the spherical, exponential, Gaussian, and hole effect variogram functions) can be adapted for use within the MEM-based predictive SEF framework.

Objectives

In the present study, we developed the calculations whereby the MEM framework can be adapted to generate SEF that are suitable for making predictions (i.e., informed interpolation) for environmental variables observed in the field, be they abiotic (e.g., temperature, humidity, pH, pressure) or biotic (e.g., species abundance, density, or diversity). We also included new DWF derived from common variogram models. We carried out a simulation study to test their performance at predicting spatial variation in various situations involving various types of randomly-generated spatially-structured (Brownian motion) plots, random sets of sampling locations, and sample sizes for each of the seven DWF under consideration. Lastly, we exemplified spatial modelling in practice by modelling the substrate density and water content of the peat vegetation mat located along the shore of a Canadian Shield bog lake, and the spatial distribution of 35 Oribatid mite species living in that soil.

Materials and methods

MEM: Calculation

MEM calculation, as defined in Dray, Legendre, and Peres-Neto (2006), proceeds from the two matrices that we mentioned previously in the introduction, namely the connectivity $\mathbf{B}$ and the weights $\mathbf{A}$. The next step consists in the Hadamard (element-wise) product of these two matrices, resulting in a weighted connectivity matrix ($\{\mathbf{B} * \mathbf{A}\}$). Matrix $\mathbf{B}$ has values $b_{i,j} = 1$ when any two points i and j are connected and $b_{i,j} = 0$ otherwise. It can be obtained from a list of edges from a connectivity graph, such as that derived, for instance, from a Delauney triangulation, a minimum spanning tree, or simply by truncation, i.e., by applying a distance threshold to a matrix of pairwise distances among locations ($[d_{i,j}]$, e.g., a Cartesian or geodesic two-dimensional space, a three-dimensional Euclidean space, or a one-dimensional transect), or some other type of connectivity matrix among the sites. As stated earlier, the spatial weights matrix $\mathbf{A}$ may be obtained by transforming the elements of $[d_{i,j}]$ using a DWF. Following that, matrix $\{\mathbf{B} * \mathbf{A}\}$ is row- and column-centred to a mean of 0 and submitted to eigenvalue decomposition. By virtue of the centring to 0, the centred weighted connectivity matrix has a rank of at most $n - 1$, where n is the number of different locations. It thus has at most $n - 1$ non-zero eigenvalues and eigenvectors. The whole process can be written in matrix notation as follows:

$$\mathbf{Q}\{\mathbf{B} * \mathbf{A}\}\mathbf{Q} = \mathbf{U}\mathbf{D}_\lambda\mathbf{U}^\mathsf{T}, \tag{1}$$

where $\mathbf{Q} = \mathbf{I}_n - \frac{1}{n}\mathbf{1}_{n \times n}$ is the idempotent centring matrix of dimension $n \times n$ ($\mathbf{I}_n$ is an identity matrix of order n and $\mathbf{1}_{n \times n}$ is an $n \times n$ all-ones matrix), $\mathbf{U}$ is a matrix of eigenvectors of dimensions $n \times k$, where $k \leq (n-1)$, and $\mathbf{D}_\lambda$ is a diagonal matrix of (non-zero) eigenvalues. As shown by Jong, Sprenger, and Veen (2010) there is an algebraic equivalence between these eigenvalues and the Moran's index (I) of their corresponding eigenvectors, had $\{\mathbf{B} * \mathbf{A}\}$ been used during the index calculation. Assuming the values on the diagonal of $\{\mathbf{B} * \mathbf{A}\}$ to be 0, this equivalence is the following:

$$I_{\lambda_k} = n \frac{\lambda_k}{\sum_{\forall i,j} b_{i,j} a_{i,j}}, \tag{2}$$

Three DWF have been proposed by Dray, Legendre, and Peres-Neto (2006) (Table 1). It is noteworthy that these functions do not form an exhaustive set of all possible DWFs; many other such functions can be developed, which may be suitable for specific questions.

Making predictions

Distance-weighting functions

In this paper, we are interested in the behaviour of MEM eigenvectors (SEV) between the sampling locations, in order to assess their potential as bases for predictive Moran's eigenvector maps (hereafter referred to as pMEM). While pMEM has a similar purpose as spatial interpolation methods such as kriging, the former are based on descriptors (i.e., the column vectors of matrix $\mathbf{U}$) rather than on direct calculations on the raw response data. The SEF used for pMEM are continuous functions and defined for any location in the space surrounding the sampling locations. Their values at the sampling locations are exactly those of the column vectors of $\mathbf{U}$, but their values vary at surrounding locations. As such, the SEV are the expression of the SEF at the sampling locations, whereas the sampling sites and the surrounding locations define the set of points in space on which the SEF are mapped. Moreover, the extent and shape of the spatial structure that the SEFs represent are conditioned by the set of sampling locations and the distances among them.

There may be a link between the spatial operator (i.e., the DWF) and the smoothness of the resulting SEF, possibly impacting their adequacy for representing spatial phenomena. Notably, the smoothness of the SEF in the vicinity of the sampling locations entails that they are representative points along continua rather than singularities, around which sharp spatial shifts may be occurring (See Appendix II – Analysis of SEF shape and smoothness, for an in-depth discussion on that subject).

For the sake of simplicity, we will restrict the definition of connectivity to be strictly distance-based and thus, from here, disregard any graph-based definition. This simplification enables us to formalize both the connectivity and distance-weighting into single functions of the distances with parameter d_{max} acting as a truncation distance beyond which points are considered non-connected as follows:

$$w_{i,j} = \begin{cases} d_{i,j} < d_{max}, f(d_{i,j}; d_{max}, \alpha) \\ d_{i,j} \geq d_{max}, 0 \end{cases}, \tag{3}$$

where $f(d_{i,j}; d_{max}, \alpha)$ is a function of the distance with a range parameter d_{max} and a shape parameter α (see Appendix I – Distance weighting function derived from the MEM framework, for details about these functions).

These functions take values 0 for distances above d_{max}, thereby involving a threshold in an implicit, distance-based, manner. For the calculation of pMEM, matrix $\mathbf{W} = [w_{i,j}]$, can therefore replace matrix $\{\mathbf{B} * \mathbf{A}\}$ since it involves an implicit distance threshold $d \leq d_{max}$. On the other hand, this definition implies that the value 1 is consistently found on the diagonal of $\mathbf{W}$, which alters the equivalence between the eigenvalues and corresponding eigenvector's associated to Moran's index (I), which is now calculated as follows:

$$I_{\lambda_k} = n \frac{\lambda_k - 1}{\sum_{\forall i,j} w_{i,j} - n}, \tag{4}$$

Therefore, using a continuous spatial operator has little impact on the interpretation of the eigenvectors in terms of Moran's index.

Variogram models

As stated earlier, the DWFs proposed by Dray, Legendre, and Peres-Neto (2006) are but a subset of all possible such functions. For this paper, we propose the addition of four DWFs derived from variogram models commonly used for kriging (Legendre and Legendre 2012). These functions are the spherical, exponential, Gaussian, and hole effect DWFs. For kriging, these variogram functions $f(d)$ describe how the spatial variance ($\gamma(d)$) increases from a local variance value (γ_n, i.e., the nugget) towards a theoretical maximum variance value (γ_s, i.e., the sill) as the distance increases as follows:

$$\gamma(d) = \gamma_n + (\gamma_s - \gamma_n) f(d), \tag{5}$$

where $f(d)$ is the variogram model function. The distance at which $\gamma(d)$ reaches γ_s is called the range of the variogram. For pMEM, the DWF has a maximum value of $w_i = 1$ at $d_i = 0$ and a minimum value of $w_i = 0$ at $d_i = d_{max}$, which corresponds to the range of the variogram function. Therefore, the variogam-based DWF are defined as $w_i = 1 - f(d_i)$ (Table 2).

These functions were studied, alongside the linear, power, and hyperbolic DWFs presented earlier, and inspired by the ones proposed by Dray, Legendre, and Peres-Neto (2006), as DWFs for spatial modelling or plain spatial interpolation using pMEM (Figure 1).

It is noteworthy that parameter d_{max} in the exponential, Gaussian, and hole effect DWF do not involve a threshold making $w_i = 0$ when $d_i \geq d_{max}$. Also, note that the common definitions for the exponential and Gaussian DWFs would involve multiplying $d_{i,j}/d_{max}$ (or $(d_i/d_{max})^2$) by 3 within the equations. We regarded that multiplication as superfluous since its only notable effect is to make the shape of these two DWFs differ more markedly from that of the other five DWFs; we thus avoided it.

Spatial eigenfunctions

One can represent the spatial eigenvectors of the centred weight matrix by performing an algebraic reorganization of the eigensystem equation presented earlier (Eq. 1), as follows:

$$\left(\mathbf{I}_n - \frac{1}{n}\mathbf{1}_{n\times n}\right)\mathbf{W}\left(\mathbf{I}_n - \frac{1}{n}\mathbf{1}_{n\times n}\right) = \mathbf{U}\mathbf{D}_\lambda\mathbf{U}^\top \tag{6.1}$$

$$\mathbf{I}_n\mathbf{W}\mathbf{I}_n - \frac{1}{n}\mathbf{I}_n\mathbf{W}\mathbf{1}_{n\times n} - \frac{1}{n}\mathbf{1}_{n\times n}\mathbf{W}\mathbf{I}_n + \frac{1}{n^2}\mathbf{1}_{n\times n}\mathbf{W}\mathbf{1}_{n\times n} = \mathbf{U}\mathbf{D}_\lambda\mathbf{U}^\top \tag{6.2}$$

$$\mathbf{W} - \frac{1}{n}\mathbf{W}\mathbf{1}_{n\times n} - \frac{1}{n}\mathbf{1}_{n\times n}\mathbf{W} + \frac{1}{n^2}\mathbf{1}_{n\times n}\mathbf{W}\mathbf{1}_{n\times n} = \mathbf{U}\mathbf{D}_\lambda\mathbf{U}^\top \tag{6.3}$$

$$\left(\mathbf{W} - \frac{1}{n}\mathbf{W}\mathbf{1}_{n\times n} - \frac{1}{n}\mathbf{1}_{n\times n}\mathbf{W} + \frac{1}{n^2}\mathbf{1}_{n\times n}\mathbf{W}\mathbf{1}_{n\times n}\right)\mathbf{U}\mathbf{D}_{\lambda^{-1}} = \mathbf{U} \tag{6.4}$$

,

where $\mathbf{W}$ is the pairwise weight matrix between the observations. Let $\mathbf{W}^*$ be the weight matrix calculated for m new locations using the same DWF as for $\mathbf{W}$ and the distances between the new locations and the original ones found in $\mathbf{W}$ (hence, the dimensions of $\mathbf{W}^*$ are $m \times n$). The values of these new locations on the SEF defined previously are obtained as follows:

$$\mathbf{U}^* = \left(\mathbf{W}^* - \frac{1}{n}\mathbf{W}^*\mathbf{1}_{n\times n} - \frac{1}{n}\mathbf{1}_{m\times n}\mathbf{W} + \frac{1}{n^2}\mathbf{1}_{m\times n}\mathbf{W}\mathbf{1}_{n\times n}\right)\mathbf{U}\mathbf{D}_{\lambda^{-1}}, \tag{7}$$

where the matrix of SEF values $\mathbf{U}^*$ has dimensions $m \times k$, with k being the number of non-zero eigenvalues in the eigensystem. Using that approach, it is possible to calculate the values of the SEF at any location, and thus make spatially-explicit predictions. However, we have yet to provide an assessment of the adequacy of the seven DWF defined previously for such a purpose.

Estimating parameters

The choice of a DWF, as well as the estimation of parameters d_{max} and α, can be carried out using different global search methods. For the present study we propose to select the most suitable DWF by trying them all, while estimating the most suitable DWF parameters separately for each of the functions using the directed evolution approach described by Ardia et al. (2011) and implemented in **R** language function **DEoptim** (Mullen et al. 2011). The objective criterion to be minimized during the DEoptim global search procedure was the mean squared prediction error (MSE). By default, function **DEoptim** uses a population size ten times the number of parameters (i.e., 20 individuals in our two-parameter case), and 200 generations.

Numerical simulations

We performed a simulation study assessing SEF ability for making predictions. For that purpose, we generated 25 two-dimensional maps. Each of these maps contained 5 184 points regularly spaced over a 72×72 staggered-row triangular grid pattern with neighbouring points located at distances 1 (in arbitrary spatial units) from one another. The data were generated at each point of that grid following a randomly-seeded Wiener process (i.e., Brownian motion) whose implementation is described in the appendices (Appendix I – Algorithm to generate the spatially-structured random maps).

To simulate the effect of sampling variation and sample size (n), 25 sets of 500 vertices were randomly selected. From each of these sets, pairs of subsets of n = 10, 20, 50, 100, 200, and 500 were picked as the training data sets, and all other $5\,184 - n$ data points were used as the testing data sets. This procedure resulted in 3 750 simulated data sets (25 maps $\times$25 subsets $\times$6 sample sizes). SEF were calculated from each training simulated data set using each of the seven DWF, for a grand total sample size of 26 250 ($3\,750 \times 7$) trials. For each of these trials, a **DEoptim** global search for estimating parameter values for d_{max} and α minimizing the MSE was carried out using the default population size and 50 generations (lower than the default in order to mitigate computational time given the large number of simulations). Values of the lower and upper bounds for d_{max} were 1 and 1 000, respectively, whereas the ones for α were 0.25 and 1.75, respectively.

Simulations results were analyzed on the basis of the predictions quality factor Q, which the log ratio of the mean square deviation MSD and the mean squared error MSE, whereas the coefficient of prediction (P^2) was used to display the results (see Appendix I – Calculations on the simulation results, for a justification of using Q and for P^2 and details on their calculation).

Simulation results were analyzed using the analysis of variance (ANOVA). Two such analyses were performed. A first ANOVA was carried out on all 26 250 trials using four variables, one quantitative: the base-10 logarithm of the sample size ($\log_{10} n$), and three qualitative (or factors): DWF, Map, and $Sample$, as well as their six second-order interaction and four third-interaction terms. A second restrained ANOVA was performed on the subset of the best-DWF trial for each of the 3 750 simulated data sets. The latter was a three-variable design with variables: $\log_{10} n$, Map, and $Sample$, and their four second-order and single third-order interaction terms.

Application example

Data set

The SEF prediction approach described in the present study was applied to a well-studied data example. The chosen data set involves the distribution of 35 taxa of Oribatid mite (class: Arachnida) in a peat bog mat located on the shore of Lac Geai, a small lake located on the territory of « Station de Biologie des Laurentides, Université de Montréal », in the conurbation of St. Hippolyte, Quebec, Canada. This data set was first described by Borcard and Legendre (1994) and various copies are available, notably from `R` packages **ade4** (`oribatid`), **codep** (`mite`), and **vegan** (`mite`) as well as from the Borcard, Gillet, and Legendre (2018) book. Sampling was carried out in June of 1989.

The sampling area was 10 m long by 2.6 m wide, with the long axis stretching from the forest to the open water of the lake. The coordinates of its centre were approximately (45.99549, −73.99370). Further details on the lake, its water, and its surroundings are found in Borcard and Legendre (1994).

Core samples of peat were taken and the Oribatid mites inhabiting them were extracted, sorted, identified, and classified into 35 morphospecies and genera. These taxa are chiefly based on morphology, since relatively little was known about the ecology and physiology of these small animals. The Oribatid community structure was analyzed using a principal component analysis (PCA) of the Hellinger-transformed abundance data (Legendre and Gallagher 2001), keeping the first two principal components.

In addition to the Oribatid mite counts by species, the data set includes six environmental predictors, namely, (1) the substrate density (quantitative; the mass of an unpacked volume dry substrate, $\mathrm{g} \cdot \mathrm{L}^{-1}$), (2) the water content (quantitative; the mass of water by volume of wet substrate, $\mathrm{g} \cdot \mathrm{L}^{-1}$), (3) the substrate type (qualitative; represented by six non mutually exclusive binary-coded classes: « Sphagn1 », « Sphagn2 », « Sphagn3 », « Sphagn4 », « Litter », « Bare peat »), shrub density (semi-quantitative; three levels: « None »,

« Few », « Many »), topography (qualitative; two mutually exclusive classes: « Hummock » and « Blanket »), and a binary variable indicating flooded areas. This last variable was obtained from the maps in Borcard and Legendre (1994) (their figure 1) and is not available in the data sets in `R` packages **ade4**, **codep**, and **vegan**.

We assembled the data points into a single point geometry stored as a geopackage file and added polygon geometries for the substrate type, shrub density, topography, and flooded areas, which we outlined manually at a resolution of roughly 0.01 m from the three images obtained from figure 1 in Borcard and Legendre (1994) using software QGIS https://qgis.org (Figure 2). The species and environmental data matrix contains the variables on which spatial modelling will be carried out in this example.

Modelling

Two continuous environmental variables, namely substrate density and water content, were not available from geographic information layers, but measurements had been taken at the sampling point locations. To be able to use them for predicting the density of the different mite species at any location over the sampling area, a continuous map of these variables was needed. We took this need as an opportunity to illustrate single-variable prediction using SEF exclusively. We began be generating a point grid over the sampling area with a resolution of 5 cm. This grid was used as a basis for generating GIS rasters for the different variables involved in this example. Variables substrate density and water content were modelled using an L_1-regularized (LASSO) linear regression model calculated using `R` package **glmnet** (Friedman, Tibshirani, and Hastie 2010) using the Gaussian family of Generalized Linear Models (GLM) and predicted values were computed over the grid points. Values of parameters d_{max} and α were estimated by **DEoptim** (default parameters), using d_{max} values between 1 and 5 m and α values between 0.15 and 1.85. Seven cross-validation folds were used for estimating the predictive power. Assignment to cross-validation folds was carried out in a systematic manner following the order in which the data appear in the data set by selecting data points with indices $i + 7 * j$ where i is the cross-validation fold (1–7) and $j = 0, 1, 2, ..., 9$.

Variables *substrate type*, *shrub density*, *topography*, and *flooded* were available directly from the polygon geometries. Variable *substrate type* was a set of non mutually exclusive classes, since cores had sometimes purposefully been taken at the boundaries of areas with different substrates in order to study ecological transitions. Therefore, this variable was available as a six-column matrix of binary (or dummy) variables rather than as a single factor with mutually exclusive levels. Each element of that binary matrix was divided by the sum of the row in order to make all the rows of the resulting transformed matrix sum to 1. This treatment made the effects of the substrate types additive. Variable *shrub density* was semi-quantitative and treated using polynomial contrasts, whereas variable *topography*, which has two levels was transformed into a binary variable and centred to a mean value of 0. Finally, variable *flooded* was used as is.

We modelled species distributions from the individual count data using a Poisson-family L_1-regularized multivariate generalized regression model (GLM), which was also calculated using the R package **glmnet** (Friedman, Tibshirani, and Hastie 2010; Tay, Narasimhan, and Hastie 2023). We used customized R language code to allow **glmnet** to handle a multivariate response (i.e., the 35 mite species) since the package does not support multivariate models natively. The model's quality of fit was estimated separately for each species using the likelihood-based R^2 coefficient for the Poisson-family of GLM proposed by Guénard et al. (2017). Values of parameters d_{max} and α were estimated by **DEoptim** under identical conditions as for the aforementioned substrate density and water content models.

Results

Simulations

The analysis of variance computed over all simulation results reveals that Q_{pred} was most affected by the sample size of the training set ($\log_{10} n$, Table 3). This result was expected; it is well know that the potential of a model at generalizing its target data is a function of the sample size of its training data, as it is a consequence of Hoeffding's inequality (Hoeffding 1963). The second most important factor was Map, which showed that maps generated during the simulation had various levels of predictability by spatial modelling. The third factor was DWF, followed by *Sample* (see Appendix III – Figures AIII 1–3 for details on these results). The marginal effects of these factors were all statistically significant. All but one of the second-order interaction terms were also statistically significant, the notable exception being interaction term $DWF \times Sample$. All second order interaction terms involving $\log_{10} n$ and Map were statistically significant. Two of the four third-order interaction terms were statistically significant: interaction term $\log_{10} n \times DWF \times Map$, indicating that the manner by which $\log_{10} n$ affects Q_{pred} varies among various DWF-Map combinations, and interaction term $\log_{10} n \times Map \times Sample$, indicating that the effect of $\log_{10} n$ is also modulated in various ways among the Map-Sample combinations.

The distance weighting function that was the most frequently associated with the best model was the power function (1633 instances, 43.5%), followed by the Gaussian, (684, 18.2%), the hole effect (499, 13.3%), the exponential, (359, 9.6%), the hyperbolic (320, 8.5%), the spherical (147, 3.9%), and, finally, the linear DWF (108, 2.9%). During the simulations, the Q_{pred} of the best-DWF models was also mainly affected by the sample size (Table 4). The mean P^2 was 0.2959 when $n = 10$ ($Q_{pred} = 0.1524$), and increased to 0.4538 when $n = 20$ ($Q_{pred} = 0.2627$), to 0.6096 when $n = 50$ ($Q_{pred} = 0.4085$), to 0.6972 when $n = 100$ ($Q_{pred} = 0.5188$), to 0.7651 when $n = 200$ ($Q_{pred} = 0.6292$), and finally to 0.8321 when $n = 500$ ($Q_{pred} = 0.775$).

The Q_{pred} of the best-DWF models also varied among the maps and, but to a much lesser extent, among the subsets. The significant among-map variation in the Q_{pred} entails that some of the maps are more or less predictable than others as a consequence of their random origin from sets of sporadically spread initial points (See appendix III – Supplementary figures – Simulation results). Interaction term $\log_{10} n \times Map$ was also statistically significant, indicating that an increase in the size of the training sample improves predictions for some of the maps more than for some others.

The among-sample variation of the Q_{pred} was smaller than that of Map, and interaction term $\log_{10} n \times Sample$ was also significant (See appendix III – Supplementary figures – Simulation results). It thus appears that some of the randomly-generated training samples were more suitable than some others to properly sample the maps, and that this suitability was increased in different ways as the sample size was increased.

Finally, interaction terms $Map \times Sample$ and $\log_{10} n \times Map \times Sample$ were also statistically significant, highlighting that the different random training samples had varying suitability at representing the different maps, and that this suitability also increased in different ways with increasing training sample size.

Oribatid mite example

The best subordinate model predicting substrate density was found to use the power DWF (Appendix Eq. A2) with a range of $d_{max} = 1.14\,\mathrm{m}$ and a shape parameter value of $\alpha = 0.67$. This model was made of six SEF; the square root of the mean squared error (RMS) was $11.3\,\mathrm{g\ L^{-1}}$ ($P^2 = 0.088$; Figure 3). This model was thus only slightly better than taking the mean value substrate density ($39.28\,\mathrm{g\ L^{-1}}$) as the predicted value. For the water content model, the best DWF was the Gaussian DWF (Eq. T2 3 from Table 2) with a range of $d_{max} = 1.12\,\mathrm{m}$, comprising 11 SEF; the $RMSE$ was $122.5\,\mathrm{g\ L^{-1}}$ ($P^2 = 0.25$).

The best DWF for predicting Oribatid mite species distribution was the power DWF (Appendix Eq. A2), with a range of $d_{max} = 2.34\,\mathrm{m}$, a shape parameter value of $\alpha = 1.68$, and deviance value ($-2 \log L$) of 4.137. The model's likelihood-based R^2 varied from 0.073 for species *Hyporufu* to 0.878 for species *Limncfci* (median: 0.548; Appendix II Table A-II 1). The ability of the model to predict mite species counts was proportional to the mean abundance of the species in the sampling area ($F_{1,33} = 18.32$, $P < 0.001$; with log-transformed mean abundance and predictability estimated as $Q = -\log_{10}(1 - R^2)$). For instance, the expected R^2 is 0.357 for a mean count of 0.157 ind. core^{-1} (the minimum value observed), 0.574 for a mean count of 1 ind. core^{-1}, 0.745 for a mean count of 10 ind. core^{-1}, and 0.807 for a mean count of 35.26 ind. core^{-1} (the maximum value observed). Also consistent with this result is the observation that species absent from a large number of sites (e.g., *Hyporufu*, which is absent from 60 of the 70 sites) tend to have a small R^2 compared to species present in many sites (e.g., *Limncfci*, which is absent from only 15 of the 70 sites).

Community structure

The two axes of the PCA carried out on the transformed species composition matrix accounted for approximately 25% of the variance of the data matrix (Figure 2). The first PCA axis was driven chiefly by the preponderance of *Tectvela Oppiniva*, and *Suctobsp*, which are associated with negative loading, with respect to that of *Limncfci*, *Limncfru*, and, to a lesser extent, *Trhyposp* and *Trimalsp*, which are associated with positive loading. The second PCA axis was driven by the preponderance of *Limncfru*, *Hoplcfpa*, and *Suctobsp*, which are associated with negative loading, with respect to that of *Limncfci*, *Trhyposp*, and *Tectvela*, which are associated with positive loading.

The components of the Oribatid community structure described by the PCA axes followed their own particular distribution spatial patterns (Figure 4). For the first PCA axis, large negative values were observed close to the forest line, at a distance of approximately 1 m from the lower end of the plot, whereas large positive values were observed near the waterline at a distance of approximately 1 m from it. The most extreme values of the second PCA axis (positive) were observed close the the forest and in and around the flooded areas. Negative second PCA axis loading values were observed on the right of the map at around a third of the distance from the waterline and forest line.

Discussion

In the present study, we developed the predictive Moran's Eigenvector Maps, a computational framework for making spatially-explicit spatial predictions at arbitrary locations about sampling points bearing known values. This goal is similar to that of common spatial interpolation methods such as kriging. However, whereas interpolation methods are non-parametric and thus based on the direct involvement of the data points, pMEM is a parametric method involving explanatory descriptors. That property entails that pMEM is a method that does not provide direct interpolation estimates of the variable it seeks to estimate. Instead, it provides descriptors, in the form of SEF, to be used later during analyses and model development. These descriptors are usable as is (e.g., when predicting substrate density or water content in the mite example) or in combination with additional descriptors (e.g., when predicting Oribatid mite species distributions in our example). Furthermore, any suitable model estimation approach can be used during the subsequent steps of the modelling workflow (e.g., an L_1 regularized generalized linear model in the oribatid mite examples). Besides the more common linear model estimation methods such as the one we used in the example, alternative machine learning methods can also be used. These methods include regression trees, gradient boosted trees (Mason et al. 1999; Chen and Guestrin 2016), Bayesian additive regression trees (Chipman, George, and McCulloch 2010), support vector machines (Cortes and Vapnik 1995), artificial neural networks (Goodfellow, Bengio, and Courville 2016), among others. In machine learning parlance, pMEM is referred to as a « feature engineering » approach (Chollet and Allaire 2018). This preliminary step involves the introduction of a numerical representation of the spatial coordinates in the model, in the form of latent variables. The addition of this numerical representation helps the model in modelling the response(s) on the basis of estimated spatial variation patterns.

Since pMEM involve descriptors, error estimation on the predictions is handled by the method that uses them for modelling. For instance, the handling of prediction error is well-established for multiple linear regression (but see Zhang (1993) for a caveat on using that approach). At the price of more computational power, cross-validation or other random sampling approaches (e.g., bootstrap, jackknife) can be used to obtain numerical estimates of the prediction error for virtually any modelling method. The details about the estimation of prediction error belongs to the particular method using the pMEM eigenfunctions and are thus outside the scope of the present study.

The simulation study we performed indicates that any of the DWF may at times yield sets of SEF that were the most appropriate to model the simulated data, which were samples from two-dimensional maps generated by Brownian motion simulations. When applied to real data, the three models built involved SEF from two DWF: the power DWF and the Gaussian DWF. These observations indicate that SEF with different orders of continuity may be equally suitable for spatial modelling and that having multiple DWF is a beneficial aspect of the pMEM toolbox, as it is presently developed. Actually, other DWF besides the ones described in the present study may be proposed in future developments of the pMEM method.

Simulation results indicated that pMEM were able to model and predict spatially structured variables with various degrees of success, depending primarily on the sample size and secondarily on a suite of other factors related to sampling and DWF selection, albeit to a lesser extent (Table 3). Simulation results highlighted that the data generation procedure was also successful at producing maps with various degrees of predictability using pMEM. Some of the DWF were more often selected than others as the best-suited one for a given set of conditions (in terms of spatial context, sample, and so on). For instance, the power DWF was the most commonly selected and the linear DWF was the least commonly selected, yet every DWF was found to be the most adequate at making spatially-explicit predictions on given $Map \times Sample$ combinations. On the one hand, picking the most suitable DWF was not as important for spatial predictability as the sample size, and its effect was relatively small with respect to the among-map variability, yet more important than the among-sample variability. On the other hand, choosing the most suitable DWF incurs no supplementary cost, unlike increasing the sample size, or altering the sampling approach.

The present study exemplified the use of pMEM using a modest-sized data set involving 70 observations. Using SEF-only models and regularized regressions, we were nevertheless able to predict the spatial distributions of two environmental variable, the substrate's density and water content, with some success ($P^2 > 0$). Then, using complex models involving environmental variables, we have been able to predict the distribution of 35 mite species with various degrees of success. For instance, substrate density was predicted with a modest accuracy (P^2 of 0.088), with an $RMSE$ of $11.3\,\mathrm{g\ L^{-1}}$, which was only slightly above the variable's standard deviation ($11.9\,\mathrm{g\ L^{-1}}$). Substrate water content was slightly more accurately predicted (P^2 of 0.25), with an $RMSE$ of $122.5\,\mathrm{g\ L^{-1}}$ for a standard deviation of $142.4\,\mathrm{g\ L^{-1}}$. Model accuracy for mite species distribution was mainly influenced by the observed species counts, with P^2 values from a minimum of 0.073 to a maximum of 0.878. This result was not unexpected as the rare species were absent from most cores and only found at low frequencies in a few other cores, thereby making the determination of their preferred conditions more uncertain. On the other hand, the more prevalent species were observed in most of the cores with low to higher frequencies, a situation that makes it easier to determine the preferred conditions sought after by the species, provided that relevant descriptors are available. Finally, we found no statistically significant differences between the Oribatid mite β-diversity, in terms of the total β-diversity and $LCBD$ indices, calculated from the observed and the predicted species frequencies.

The computation of the pMEM relies on square matrices for storing distances and the weights and on eigenvalue decomposition, which is a computationally demanding method. While it is not a problem for small data sets such as the ones shown in the present study, requirements in term of computer memory storage and computation time become prohibitive on large data set (a few thousand data points) even for state of the art computer systems. A straightforward solution to adapting pMEM to large data set would be to consider using the pairwise distances between the n data points and a set of k representative spatial kernels disseminated over the study area. The resulting $n \times k$ rectangular distance matrix could be transformed into a spatial weights matrix, submitted to centering and then to singular value decomposition (SVD). By choosing a parsimonious number of kernels, the kernel-based pMEM thus obtained would remain applicable to large data sets (in the tens or hundreds of thousand of data points). Given that k would be much smaller than n, the number of SEF would be equal to k (or, perhaps, slightly smaller). This property might also help in simplifying model building to some extent. That proposal opens other matters that we did not have to ponder with while studying pMEM. For instance, the approach for choosing the number of kernels and their locations would have to be considered (e.g., using medoids *vs* centroids). Also, the linear algebra linking the Moran's I index to the SEF thus defined would need to be demonstrated, since that link is helpful in interpreting the spatial scale associated with the SEF. These matters, and possibly other unanticipated matters that may likely spring up while developing kernel-based pMEM, are clearly beyond the scope of the present study.

The pMEM framework may be useful for other purposes besides our resolutely machine-learning oriented objective of using it for making spatially-explicit predictions. For instance, one may consider using it to correct the confounding effect of spatial autocorrelation when carrying out statistical inference testing. However, it is worth recalling that pMEM are identical to MEM when only considering the sampling points. To what extent the four variogram-based DWF we introduced would improve MEM performance in correcting spatial confounding will remain an unanswered questions until a thorough simulation study addressing that matter is carried out. In the meantime, we consider it safer to assume that actual knowledge about spatial confounding still holds, and thus would direct the reader to the competent literature on that particular subject matter (Thaden and Kneib 2018; Dupont, Wood, and Augustin 2022; Marques, Kneib, and Klein 2022; Mäkinen et al. 2022).

We are hoping that the findings highlighted in the present study will entice scientists to use pMEM to model spatial variation and for making predictions. Also, we look forward for numerical ecologists to further the development of pMEM from its actual enactment, and for software developers to expand the implementation of the approaches to other computer languages and software.

Conflict of interest statement

We declare no conflict of interest.

Data availability

An R package called pMEM and all the data used for this study (computer simulations, example calculations, Appendices) are available through the following anonymous.4open.science link.

References

Antunes, N., W. Schiefenhövel, F. d'Errico, W. E. Banks, and M. Vanhaeren. 2020. "Quantitative Methods Demonstrate That Environment Alone Is an Insufficient Predictor of Present-Day Language Distributions in New Guinea." *PLOS ONE* 15 (10): e0239359. https://doi.org/10.1371/journal.pone.0239359.

Ardia, D., K. Boudt, P. Carl, K. M. Mullen, and B. G. Peterson. 2011. "Differential Evolution with DEoptim." *The R Journal* 3 (1): 27–34. https://doi.org/10.32614/RJ-2011-005.

Blanchet, F. G., P. Legendre, and D. Borcard. 2008. "Modelling Directional Spatial Processes in Ecological Data." *Ecol. Model.* 215: 325–36. https://doi.org/10.1016/j.ecolmodel.2008.04.001.

Borcard, D., F. Gillet, and P. Legendre. 2018. *Numerical Ecology with R, 2^{nd} Edition.* Springer International Publishing AG.

Borcard, D., and P. Legendre. 1994. "Environmental Control and Spatial Structure in Ecological Communities: An Example Using Oribatid Mites (Acari, Oribatei)." *Environ. Ecol. Stat.* 1 (1): 37–61. https://doi.org/10.1007/BF00714196.

———. 2002. "All-Scale Spatial Analysis of Ecological Data by Means of Principal Coordinates of Neighbour Matrices." *Ecol. Model.* 153: 51–68. https://doi.org/10.1016/S0304-3800(01)00501-4.

Chen, T., and C. Guestrin. 2016. "XGBoost: A Scalable Tree Boosting System." In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–94. KDD '16. New York, NY, USA: Association for Computing Machinery.

Chipman, H. A., E. I. George, and R. E. McCulloch. 2010. "BART: Bayesian Additive Regression Trees." *Ann. Appl. Stat.* 4 (1): 266–98. https://doi.org/10.1214/09-AOAS285.

Chollet, F., and J. J. Allaire. 2018. *Deep Learning with R.* Manning Publications.

Cortes, C., and V. Vapnik. 1995. "Support-Vector Networks." *Mach. Learn.* 20 (3): 273–97. https://doi.org/10.1007/BF00994018.

Dray, S., P. Legendre, and P. Peres-Neto. 2006. "Spatial Modelling: A Comprehensive Framework for Principal Coordinate Analysis of Neighbour Matrices (Pcnm)." *Ecol. Modelling* 196: 483–93.

Dupont, E., S. N. Wood, and N. H. Augustin. 2022. "Spatial+: A Novel Approach to Spatial Confounding." *Biometrics* 78 (4): 1279–90.

Forman, R. T. T. 1995. *Land Mosaics: The Ecology of Landscapes and Regions.* Cambridge, UK.: Cambridge University Press.

Forman, R. T. T., and M. Godron. 1986. *Landscape Ecology.* New York, NY, USA.: John Wiley; Sons, Inc.

Friedman, J., R. Tibshirani, and T. Hastie. 2010. "Regularization Paths for Generalized Linear Models via Coordinate Descent." *J. Stat. Softw.* 33 (1): 1–22. https://doi.org/10.18637/jss.v033.i01.

Goodfellow, I., Y. Bengio, and A. Courville. 2016. *Deep Learning.* MIT Press.

Griffith, D. A., and P. R. Peres-Neto. 2006. "Spatial Modeling in Ecology: The Flexibility of Eigenfunction Spatial Analyses." *Ecology* 87: 2603–13.

Guénard, G., G. Lanthier, S. Harvey-Lavoie, C. J. Macnaughton, C. Senay, M. Lapointe, P. Legendre, and D. Boisclair. 2016. "A Spatially-Explicit Assessment of the Fish Population Response to Flow Management in a Heterogeneous Landscape." *Ecosphere* 7 (5): e01252.

———. 2017. "Modelling Habitat Distributions for Multiple Species Using Phylogenetics." *Ecography* 40 (9): 1088–97.

Guénard, G., and P. Legendre. 2018. "Bringing Multivariate Support to Multiscale Codependence Analysis: Assessing the Drivers of Community Structure Across Spatial Scales." *Meth. Ecol. Evol.* 9: 292–304. https://doi.org/10.1111/2041-210X.12864.

Hoeffding, W. 1963. "Probability Inequalities for Sums of Bounded Random Variables." *J. Am. Stat. Assoc.* 58 (301): 13–30. https://doi.org/10.1080/01621459.1963.10500830.

Jong, P., C. Sprenger, and F. Veen. 2010. "On Extreme Values of Moran's I and Geary's c." *Geogr. Anal.* 16 (1): 17–24. https://doi.org/10.1111/j.1538-4632.1984.tb00797.x.

Legendre, P. 1993. "Spatial Autocorrelation: Trouble or New Paradigm?" *Ecology*, no. 6: 1659–73. https://doi.org/10.2307/1939924.

Legendre, P., and M. J. Fortin. 1989. "Spatial Pattern and Ecological Analysis." *Vegetatio* 80 (2): 107–38.

Legendre, P., and E. D. Gallagher. 2001. "Ecologically Meaningful Transformations for Ordination of Species Data." *Oecologia* 129: 271–80.

Legendre, P., and L. Legendre. 2012. *Numerical Ecology, 3rd English Edition.* Amsterdam, The Netherlands: Elsevier Science B.V.

Mahecha, M. D., and S. Schmidtlein. 2008. "Revealing Biogeographical Patterns by Nonlinear Ordinations and Derived Anisotropic Spatial Filters." *Global Ecology and Biogeography* 17 (2): 284–96. https://doi.org/10.1111/j.1466-8238.2007.00368.x.

Mäkinen, J., E. Numminen, P. Niittynen, M. Luoto, and J. Vanhatalo. 2022. "Spatial Confounding in Bayesian Species Distribution Modeling." *Ecography* 33 (11): e06183.

Marques, I., T. Kneib, and N. Klein. 2022. "Mitigating Spatial Confounding by Explicitly Correlating Gaussian Random Fields." *Environmetrics* 33 (5): e2727.

Mason, L., J. Baxter, P. Bartlett, and M. Frean. 1999. "Boosting Algorithms as Gradient Descent." In *Advances in Neural Information Processing Systems*, MIT Press. Vol. 12. Boston, MA, USA.

Matheron, G. 1962. *Traité de Géostatistique Appliquée. Tomes i Et II.* Paris: Éditions Technip.

Moran, P. A. P. 1950. "Notes on Continuous Stochastic Phenomena." *Biometrika* 37 (1/2): 17–23. https:

//doi.org/10.2307/2332142.

Mullen, K. M., D. Ardia, D. Gil, D. Windover, and J. Cline. 2011. “DEoptim: An R Package for Global Optimization by Differential Evolution.” *J. Stat. Soft.* 40 (6): 1–26. https://doi.org/10.18637/jss.v040.i06.

Myers, D. E. 1984. “Co-Kriging — New Developments.” In *Geostatistics for Natural Resources Characterization: Part 1*, edited by G. Verly, M. David, A. G. Journel, and A. Marechal, 295–305. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-009-3699-7_18.

Pascoe, E. L., S. Pareeth, D. Rocchini, and M. Marcantonio. 2019. “A Lack of ‘Environmental Earth Data’ at the Microhabitat Scale Impacts Efforts to Control Invasive Arthropods That Vector Pathogens.” *Data* 4 (4): 133. https://doi.org/10.3390/data4040133.

Pebesma, E. J. 2004. “Multivariable Geostatistics in S: The Gstat Package.” *Comput. Geosci.* 30 (7): 683–91. https://doi.org/10.1016/j.cageo.2004.03.012.

Tay, J. K., B. Narasimhan, and T. Hastie. 2023. “Elastic Net Regularization Paths for All Generalized Linear Models.” *J. Stat. Softw.* 106 (1): 1–31. https://doi.org/10.18637/jss.v106.i01.

Thaden, H., and T. Kneib. 2018. “Structural Equation Models for Dealing with Spatial Confounding.” *Am. Stat.* 72 (3): 239–52.

Zhang, P. 1993. “On the Estimation of Prediction Errors in Linear Regression Models.” *Ann. Inst. Stat. Math.* 45 (1): 105–11. https://doi.org/10.1007/BF00773671.

Figures and tables

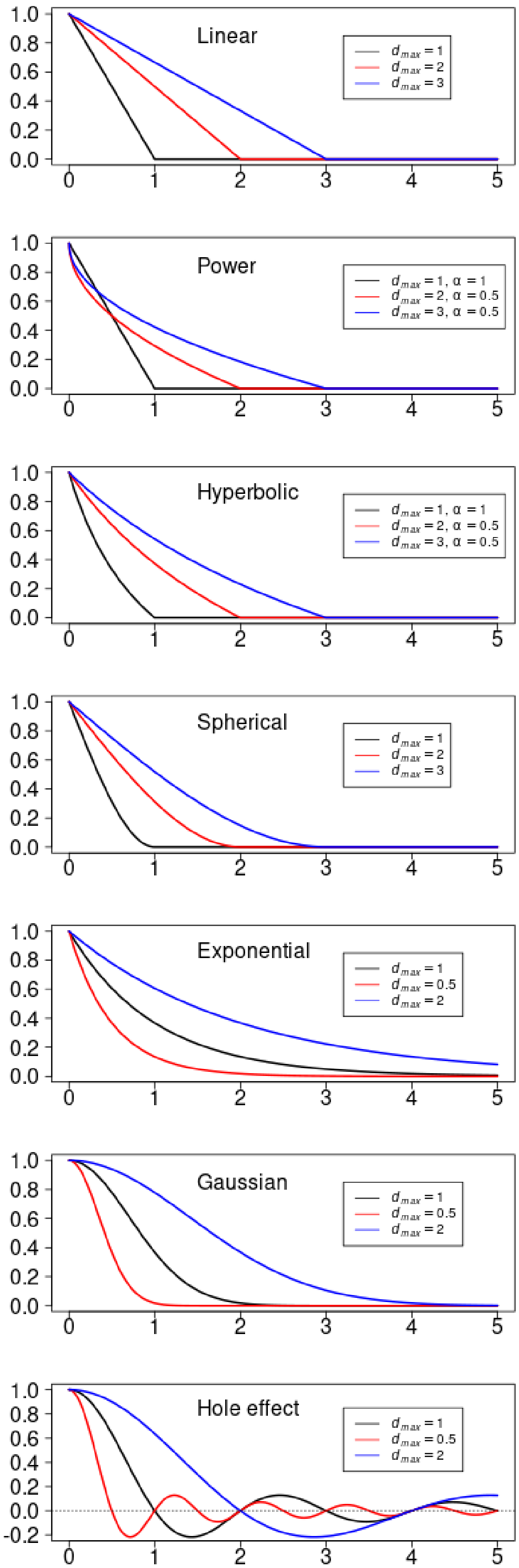

Figure 1. Distance-weighting functions whose potential for spatially-explicit modelling was assessed in this study.

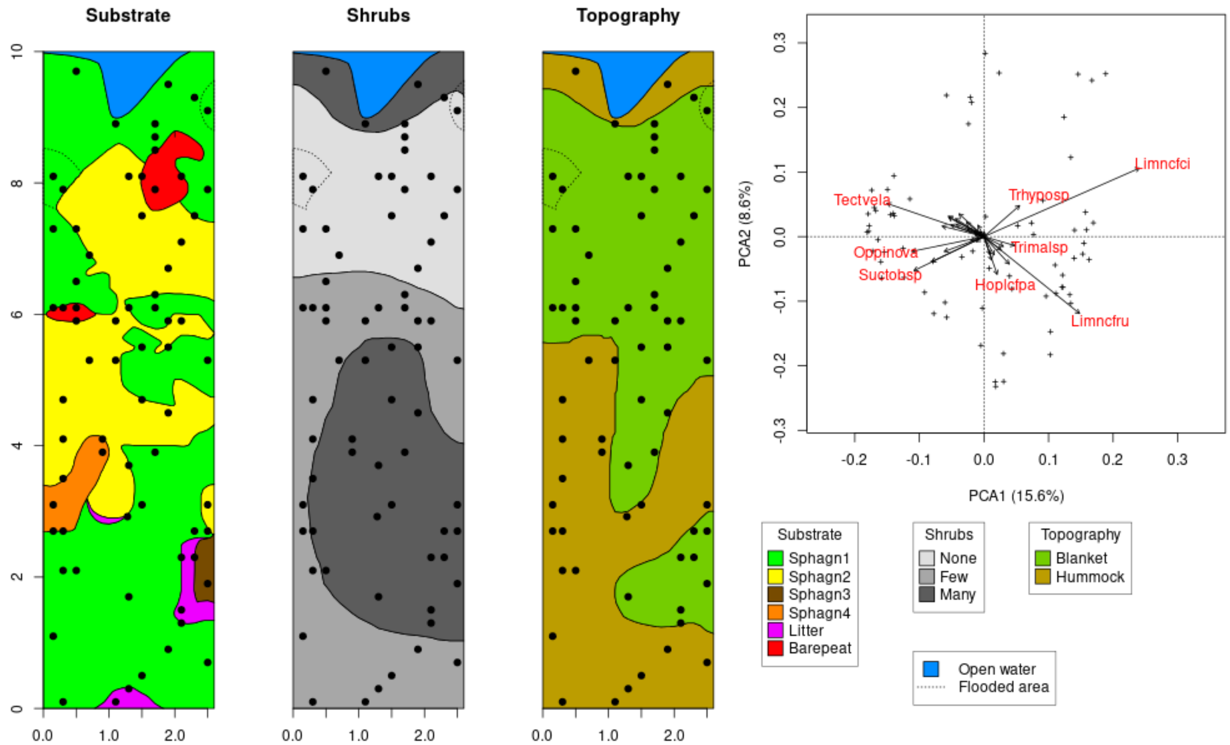

Figure 2. Maps of substrate types, shrub density, and topography outlined from Borcard and Legendre (1994, their Figure 1), together with a principal component analysis (PCA) biplot showing the sampling sites (markers) and Oribatid mite species (arrows). The maps also feature the sampling points (black dots), open waters (blue area), and flooded areas (circumscribed by dotted curves). The labels at the tip of the PCA biplot arrows are the names of the eight species with the largest axis loadings in their vicinity. The two PCA axes represent approximatly a fourth of the total species variation among the sites.

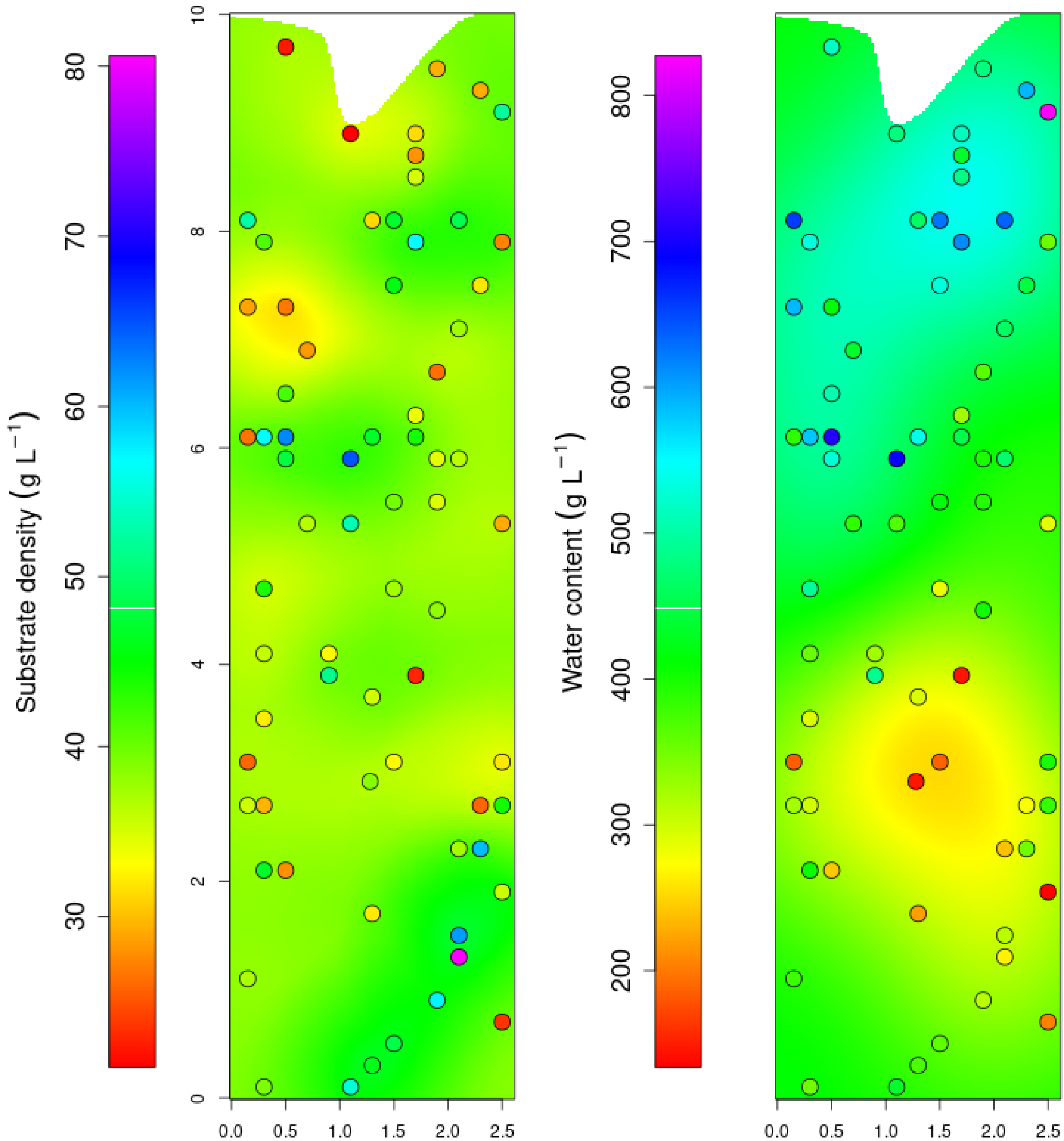

Figure 3. Results of the spatially-explicit models predicting substrate density and water content of the peat, which are defined as the mass (in grams) of solids and water per litre of uncompacted peat. Predictions are presented on the maps as rainbow colors and observed values at the sampling locations are presented with dots using the same rainbow color scale as for the model predictions. The substrate density model ($P^2 = 0.088$) is much weaker than the water content model ($P^2 = 0.25$).

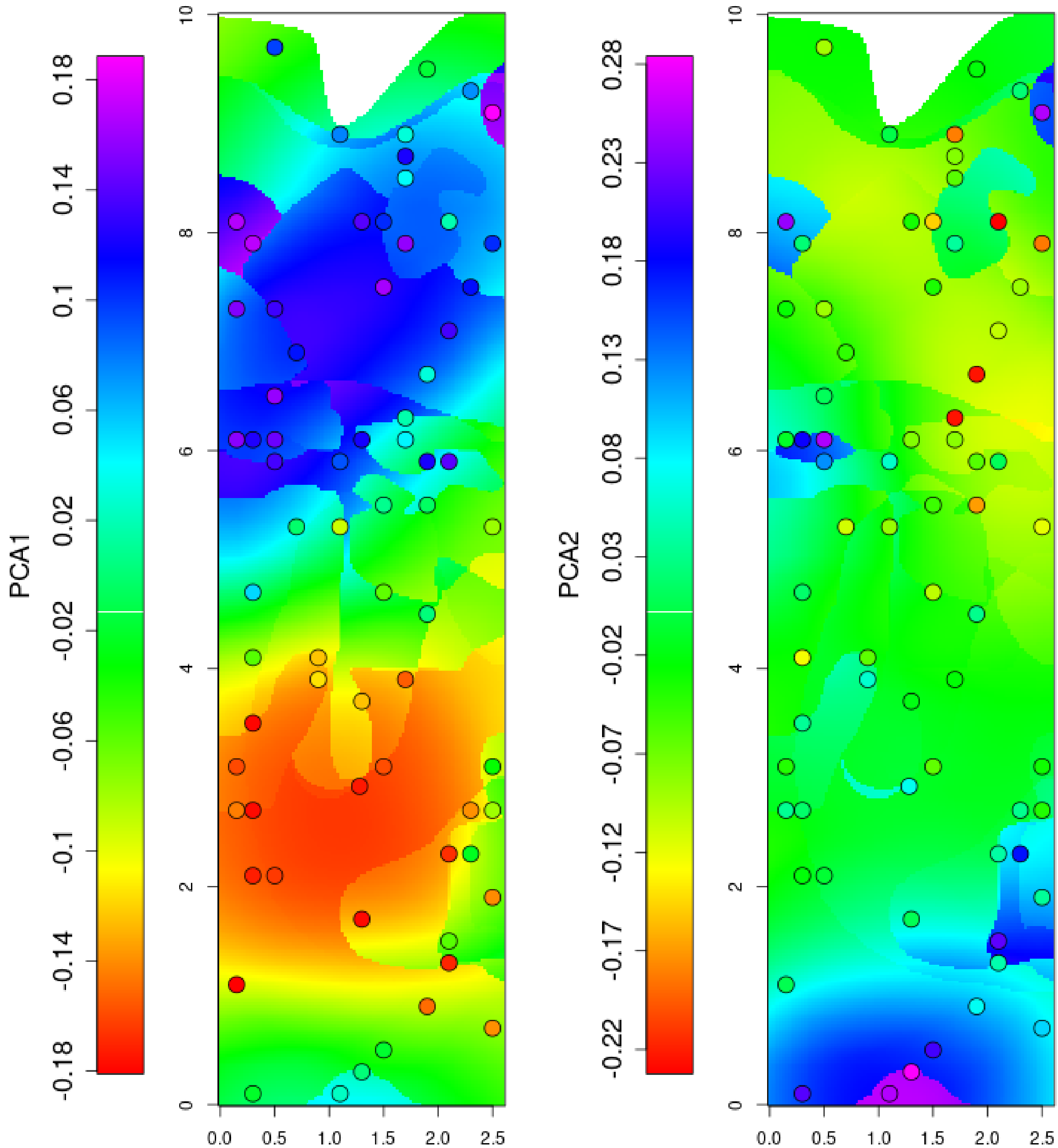

Figure 4. Site loadings of a two-axis PCA over the Oribatid mite study area. These axes represent the two main components of the Oribatid mite community structure (representing approximately 25% of the among-site species variability). Rainbow colors pixels on the surface of the study area are the values obtained from predicted species counts, whereas the background color of the markers correspond to the observed PCA axis loadings.

Table 1. Distance-weighting functions, from Dray, Legendre, and Peres-Neto (2006), commonly used for Moran's eigenvector maps calculation.

Name	Definition	
Linear	$a_{i,j} = 1 - \frac{d_{i,j}}{d_{max}}$	(T1 1)
Concave up	$a_{i,j} = 1 - \left(\frac{d_{i,j}}{d_{max}}\right)^{\alpha}$	(T1 2)
Concave down	$a_{i,j} = \frac{1}{d_{i,j}^{\alpha}}$	(T1 3)

Notes:

- d_{max}: the maximum distance for two points to be considered neighbours, also referred to as the range parameter
- α: a shape parameter

Table 2. Distance-weighting functions usable for the generation of predictive Moran's Eigenvector Maps, which are based on the classical MEM framework (1–3) of variogram models (4–7).

Name	Definition	
Linear[1]	$w_i = \begin{cases} d_i < d_{max}, 1 - \frac{d_i}{d_{max}} \\ d_i \geq d_{max}, 0 \end{cases}$	T2 1
Power[1]	$w_i = \begin{cases} d_i < d_{max}, 1 - \left(\frac{d_i}{d_{max}}\right)^{\alpha} \\ d_i \geq d_{max}, 0 \end{cases}$	T2 2
Hyperbolic[1]	$w_i = \begin{cases} d < d_{max}, \frac{\left(1+\frac{d_i}{d_{max}}\right)^{-\alpha} - 2^{-\alpha}}{1-2^{-\alpha}} \\ d_i \geq d_{max}, 0 \end{cases}$	T2 3
Spherical	$w_i = \begin{cases} d_i < d_{max}, 1 - 1.5\left(\frac{d_i}{d_{max}}\right) + 0.5\left(\frac{d_i}{d_{max}}\right)^3 \\ d_i \geq d_{max}, 0 \end{cases}$	T2 4
Exponential	$w_i = \mathrm{e}^{-\frac{d_i}{d_{max}}}$	T2 5
Gaussian	$w_i = \mathrm{e}^{-\left(\frac{d_i}{d_{max}}\right)^2}$	T2 6
Hole effect	$w_i = \begin{cases} d_i = 0, 1 \\ d_i > 0, \frac{d_{max}}{\pi d_i} \sin \frac{\pi d_i}{d_{max}} \end{cases}.$	T2 7

Notes:

1. See Appendix I – Distance weighting function derived from the MEM framework for a through presentation of these three DWF.

[Move this table to the appendix on smoothness, including the table header]

Table 3. Results of the analysis of variance of the effect of the sample size ($\log_{10} n$), distance weighting functions (DWF), maps (Map), and samples ($Sample$) on the coefficient of prediction (P^2) of all the models generated during the simulation study. The analysis also included the second and third order interaction terms.

	ν	$F_{\nu,\nu_{res}}$	P
$\log_{10} n$	1	337800	< 0.0001
DWF	6	225.4	< 0.0001
Map	24	3782	< 0.0001
$Sample$	24	38.04	< 0.0001
$\log_{10} n \times DWF$	6	174.5	< 0.0001
$\log_{10} n \times Map$	24	263.4	< 0.0001
$\log_{10} n \times Sample$	24	34.3	< 0.0001
$DWF \times Map$	144	1.936	< 0.0001
$DWF \times Sample$	144	0.4698	> 0.05
$Map \times Sample$	576	8.796	< 0.0001
$\log_{10} n \times DWF \times Map$	144	1.58	< 0.0001
$\log_{10} n \times DWF \times Sample$	144	0.6375	> 0.05
$\log_{10} n \times Map \times Sample$	576	8.13	< 0.0001
$DWF \times Map \times Sample$	3456	0.3894	> 0.05
Residuals	20956		

Table 4. Results of the analysis of variance of the effect of the sample size ($\log_{10} n$), maps (Map), and samples ($Sample$), including their second and third order interaction terms, on the coefficient of prediction (P^2) of the models with the best distance weighting functions; the latter is the one providing the highest value of the prediction quality metric for any map and subset combinations.

	ν	$F_{\nu,\nu_{res}}$	P
$\log_{10} n$	1	55980	< 0.0001
Map	24	669.7	< 0.0001
$Sample$	24	7.49	< 0.0001
$\log_{10} n \times Map$	24	39.11	< 0.0001
$\log_{10} n \times Sample$	24	6.794	< 0.0001
$Map \times Sample$	576	1.515	< 0.0001
$\log_{10} n \times Map \times Sample$	576	1.423	< 0.0001
Residuals	2500		