

ARTICLE

Macrosystems Ecology

Disentangling spatial and environmental effects: Flexible methods for community ecology and macroecology

 Duarte S. Viana^{1,2}  | Petr Keil^{1,3,4}  | Alienor Jeliaskov^{1,3,5} 

¹German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Leipzig, Germany

²Leipzig University, Leipzig, Germany

³Institute of Computer Science, Martin Luther University Halle-Wittenberg, Halle (Saale), Germany

⁴Faculty of Environmental Sciences, Czech University of Life Sciences Prague, Praha-Suchdol, Czech Republic

⁵University of Paris-Saclay, INRAE, HYCAR, Antony, France

Correspondence

Duarte S. Viana

Email: duarte.viana@idiv.de

Funding information

Deutsche Forschungsgemeinschaft, Grant/Award Number: FZT 118; Research Excellence in Environmental Sciences (REES); sDiv, the Synthesis Centre of iDiv

Handling Editor: Meredith K. Steele

Abstract

Community ecologists and macroecologists have long sought to evaluate the importance of environmental conditions in determining species distributions, community composition, and diversity across sites. Different methods have been used to estimate species–environment relationships, but their differences to jointly fit and disentangle spatial autocorrelation and structure remain poorly studied. We compared how methods in four broad families of statistical models estimated the contribution of the environment and space to variation in species occurrence and abundance. These methods included redundancy analysis (RDA), generalized linear models (GLMs), generalized additive models (GAMs), and three types of tree-based machine learning (ML) methods: boosted regression trees (BRT), random forests, and regression trees. The spatial component of the model consisted of Moran's eigenvector maps (MEMs; in RDA, GLM, and ML), smooth spatial splines (in GAM), or tree-based nonlinear modeling of spatial coordinates (in ML). We simulated typical site-by-species data to assess the methods' performance in (1) fitting environmental and spatial models, and (2) partitioning the variation explained by environmental and spatial predictors. We observed marked differences in performance among methods generally caused by their different performances in fitting spatial structures. Generalized linear model and BRT with MEMs were generally the most reliable methods for partitioning the variation explained by environmental and spatial effects across a wide range of simulated scenarios. The remaining methods tended to underestimate pure spatial effects, because of either underfitting of simulated spatial structures or overestimation of environmental effects compared to spatial effects when jointly estimated. Performing variation partitioning on nine different empirical datasets using these methods yielded contrasting results, especially in the estimation of the spatial fraction of variation. Our results suggest that previously overlooked methods for performing variation partitioning, especially tree-based ML, offer flexible approaches to analyze site-by-species matrices. We provide general guidelines on the usefulness of different models under different ecological and sampling scenarios, for species distribution modeling, community ecology, and macroecology.

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2022 The Authors. *Ecosphere* published by Wiley Periodicals LLC on behalf of The Ecological Society of America.

KEYWORDS

canonical analysis, Gaussian niche, metacommunity, random forest, spatial autocorrelation, species distribution models, variation partitioning

INTRODUCTION

The environment is a major driver of species occurrence and abundance, shaping many facets of biodiversity, from fine-scale community composition to large-scale species distributions and co-occurrence (Chase & Leibold, 2003; Townsend Peterson et al., 2011). Consequently, the environment is central in ecological theory, including coexistence theory (Chesson, 2000; Chesson & Warner, 1981), modern niche theory (Chase & Leibold, 2003), metacommunity theory (Leibold et al., 2004; Thompson et al., 2020), and biogeographical and macroecological theory (Ricklefs & Jenkins, 2011; Townsend Peterson et al., 2011). Species–environment relationships (SERs) have been widely estimated to (1) characterize the species' niches and model species distributions and community composition (Bar-Massada, 2015; Norberg et al., 2019; Townsend Peterson et al., 2011), (2) explore the importance of environmental filtering as an ecological process (Cottenie, 2005; Soininen, 2014), and (3) correct for environmental effects when studying biotic interactions and other community assembly processes (e.g., D'Amen et al., 2018; Ovaskainen et al., 2017).

One of the major challenges when studying SERs is the correct estimation of environmental effects while accounting for spatial autocorrelation of species distributions at different spatial scales, caused by both environmental autocorrelation and spatial processes such as dispersal limitation. In particular, a popular aim in community ecology has been to disentangle, using variation partitioning, the relative importance of environmentally driven (niche) processes from spatial processes often associated with neutral theory (Cottenie, 2005; Leibold & Chase, 2017; Peres-Neto et al., 2006; Soininen, 2014). However, estimating SERs while accounting for spatial autocorrelation and/or performing variation partitioning has been done differently in different research fields.

In biogeography and macroecology, generalized linear and additive models (GLM and GAM) as well as tree-based machine learning (ML) methods are widely used to estimate SERs (e.g., Elith & Graham, 2009; Norberg et al., 2019). These methods have been, however, overlooked for disentangling spatial dependence in the context of variation partitioning, particularly in a community context with multispecies responses. Machine learning methods such as random forest and boosted regression trees (BRT) (Elith & Graham, 2009), and their multivariate versions (Nieto-

Lugilde et al., 2018), have several advantages over classical regression methods, particularly because they are usually bound by fewer statistical assumptions and are inherently suited to model complex interactions and nonlinear relationships. Although comparisons of different models to estimate species distributions exist (Norberg et al., 2019), the usefulness of methods such as GAM and ML to model different spatial structures while estimating environmental effects, and to partition explained variation, has been less explored.

In community ecology, multivariate methods such as constrained ordination (e.g., canonical correspondence analysis and redundancy analysis [RDA]) or distance-based methods have been widely used to partition explained variation in community composition according to environmental and spatial effects, and to infer community assembly processes (Cottenie, 2005; Peres-Neto et al., 2006; Soininen, 2014; Tuomisto & Ruokolainen, 2006). However, these methods have been criticized, since their performance depends on the strength of effects, the spatial structure of the environmental variables, and model specification (e.g., Gilbert & Bennett, 2010; Smith & Lundholm, 2010).

Here, we compare the virtues and drawbacks of methods based on constrained ordination, GLMs, GAMs, and ML (Table 1) to model and disentangle environmental and spatial effects. These methods can all, in theory, model environmental and spatial effects but use different techniques to achieve it. They differ particularly in the estimation of spatial effects, due to their different abilities to approximate nonlinear relationships such as typical “wobbly” spatial surfaces. Methods such as RDA (a popular constrained ordination method) and GLM typically use Moran's eigenvector maps (MEMs) as linear predictors to model spatial autocorrelation and structure (Dray et al., 2012; Griffith & Peres-Neto, 2006); GAMs can fit complex nonlinear relationships via semiparametric splines (Wood, 2017); and tree-based ML can model nonlinear spatial surfaces by recursively splitting the response along spatial coordinates. These differences among methods also affect the estimation of environmental effects, but SERs such as consumer–resource curves (e.g., Fussmann et al., 2005) or thermal performance curves (e.g., Krenek et al., 2011) are typically less complex and less wiggly than spatial surfaces, and can be often approximated by simpler parametric functions (e.g., linear, logarithmic, or unimodal). Further, some of these methods assume symmetrical or normal error distributions, such as RDA and some ML algorithms, but may be suboptimal for frequency distributions

TABLE 1 The methods compared in this study and respective model fitting settings and parameters

Method family	Method name	Method full name	Type	Environmental response	Spatial modeling	Specific settings in R
Constrained ordination	RDA	Redundancy analysis (based on linear regression)	P	Linear, polynomial	MEM (based on XY coordinates)	<ul style="list-style-type: none"> - R package: <i>base</i> (function “lm”; similar to the <i>vegan</i> approach) - Species data: Hellinger-transformed - Distribution: normal
Generalized linear models	GLM	Generalized linear model	P	Linear, polynomial	MEM (based on XY coordinates)	<ul style="list-style-type: none"> - R package: <i>base</i> - Species data: raw - Distribution: normal (abundance), quasi-Poisson (counts), binomial (binary)
Generalized additive models	GAM	Generalized additive model	SP	Linear, spline ($k = 3$) (potentially any)	Splines on XY coordinates	<ul style="list-style-type: none"> - R package: <i>mgcv</i> - Species data: raw - Distribution: normal (abundance), Poisson (counts), and binomial (binary)
Machine learning tree-based methods	BRT	Boosted regression trees	NP	Any ^a	Recursive splitting of response along XY coordinates, MEM (based on XY coordinates)	<ul style="list-style-type: none"> - R package: <i>mvboost</i> - Species data: raw - Distribution: normal (abundance), Poisson (counts), and Bernoulli (binary) - No. trees: 1000 - Learning rate: 0.01 or 0.001 - Interaction depth: 2 (1 with MEMs) - Internal CV: no
	UniRF	Univariate random forest	NP	Any ^a	XY coordinates, MEM (based on XY coordinates)	<ul style="list-style-type: none"> - R package: <i>randomForest</i> - Species data: raw (abundance), categorical 0/1 (binary) - No. trees: 500 - Node size: default - Sample size: N or $N/5$
	MVRF	Multivariate random forest	NP	Any ^a	XY coordinates, MEM (based on XY coordinates)	<ul style="list-style-type: none"> - R package: <i>randomForestSRC</i> - Species data: raw (abundance), categorical 0/1 (binary) - No. trees: 500 - Node size: default - Sample size: N or $N/5$
	MVRT	Multivariate regression tree	NP	Any ^a	XY coordinates, MEM (based on XY coordinates)	<ul style="list-style-type: none"> - R package: <i>mvpart</i> - Species data: raw (abundance), numeric 0/1 (binary) - Node size: 5 (no CV) - Interaction depth: 1–20 (CV) - Internal CV: yes or no

Note: All the method choices, such as the type of spatial modeling technique, are justified in the “Methods” and Appendix S1: Table S3

Abbreviations: CV, cross-validation; MEMs, Moran’s eigenvector maps; NP, nonparametric; P, parametric; SP, semiparametric.

^aAny response shape, independently of what was simulated in this study, can be fitted, including complex interactions and nonlinearities.

of typical data such as species occurrence (binary data) and abundance (counts data). Our goal was to assess the performance of different statistical approaches for fitting environmental and spatial effects (Exercise 1), and partitioning explained variation (Exercise 2), using simulated data. In addition, we explored the impacts of choosing different methods to partition explained variation in nine empirical datasets. We specifically focus on site-by-species matrices of occurrence or abundance as response variables, but this study can apply to any kind of (spatialized) ecological response, such as species richness and beta-diversity.

METHODS

Data simulation

We simulated site-by-species community data where variation in abundance or binary occurrence corresponded to predefined species' responses to environmental conditions, spatial gradients, or both. We used a simple simulation consisting of a grid of N cells (hereafter sites), where each site was occupied by a different community and was environmentally homogeneous. The occurrence or abundance (\bar{Y}_{ij}) of each species i at a given site j depended on an environmental response (X_{ij}), a spatial response (W_{ij}), or their linear combination.

Environmental response (X)

We simulated one spatially autocorrelated environmental variable E ($0 \leq E \leq 1$) on the grid by simulating a random Gaussian field in which the autocorrelation level was set by the range parameter (A) of an exponential variogram model, using the R package *gstat* (Pebesma, 2004). Species responded to E either linearly or according to a Gaussian curve (i.e., a unimodal, bell-shaped response). As such, the response to the environment (X_{ij}) of species i in cell j was

$$X_{ij} = \beta_i E_j, \quad (1)$$

where E_j is the environmental value in cell j and β_i is the slope of the linear response (β_i was taken randomly from a normal distribution of β values; mean = 10, SD = 2). Alternatively,

$$X_{ij} = X_{\max} e^{-\frac{(E_j - \mu_i)^2}{2\sigma^2}}, \quad (2)$$

where μ_i is the mean of the Gaussian response (i.e., the optimal environmental condition for species i , μ being

equally spaced along the environmental variable from 0.05 to 0.95), σ is the standard deviation (i.e., the fundamental niche breadth), and X_{\max} ($=20$) is a constant defining the maximum abundance. For examples of the linear and Gaussian-shaped responses, see Appendix S1: Figure S1.

Spatial response (W)

The spatial response W was defined by simulating another random Gaussian field S on the same grid, independently from E , and setting the abundance (or occurrence) W_{ij} to be directly proportional to S_{ij} ($W_{ij} = X_{\max} S_{ij}$). X_{\max} ($=20$) sets W_{ij} to the same scale as X_{ij} . We simulated different types of spatial structures S by defining two different spatial models (an exponential or Gaussian variogram model) with varying ranges (A). For examples of spatial patterns of S , and thus W , see Appendix S1: Figure S2.

Multispecies responses (\bar{Y}_i)

In order to vary the relative contributions of the environment (X) and space (W) to variation in species' abundance or occurrence, X_i and W_i were given different weights (β_X and β_W , respectively; $\beta_X + \beta_W = 1$), so that

$$\bar{Y}_{ij} = \beta_X X_{ij} + \beta_W W_{ij}. \quad (3)$$

The simulated site-by-species matrix was obtained as follows. Occurrence (i.e., presence-absence) data Y_{ij} were obtained by first transforming \bar{Y}_{ij} (centered around 0) into probabilities P_{ij} using an inverse logit function, and then randomly sampling from a Bernoulli distribution with parameter P_i . Two types of abundance data Y_{ij} were generated: data with normally distributed errors ($\mu = 0$ and $\sigma = 2$) and data with Poisson distributed errors using the deterministic abundances \bar{Y}_{ij} as the mean abundances of the Poisson distribution (see an example of the spatial pattern of the resulting Y in Appendix S1: Figure S3). The relative contribution of X and W to variation in Y was measured by partitioning the squared correlation coefficient r^2 between Y_i and \bar{Y}_i (equivalent to R^2 in a linear regression) into the total fraction $[X]$ of variation attributable to X and the pure fraction $[W]$ of variation attributable to W , according to:

$$r_{XW}^2 = r(Y_i, \bar{Y}_i)^2, \quad (4)$$

$$r_X^2 = r(Y_i, X_i)^2, \quad (5)$$

$$[X] = r_X^2, \quad (6)$$

$$[W] = r_{XW}^2 - r_X^2. \quad (7)$$

Note that the difference between Y_i and \bar{Y}_i is the noise added by the random sampling, and thus, the fraction of unexplained variation (i.e., noise) is $1 - r_{XW}^2$. Although W is independent from X , some collinearity between them can arise by chance and jointly explain variation in Y . Because W is a pure spatial variable, the collinear effect is the effect of spatially correlated X . The fraction $[X]$ contains this shared fraction of variation, unlike the pure fraction of variation $[W]$. As expected, β was directly related to the partial correlation r between X_i or W_i and \bar{Y}_i , that is, $\beta_X \propto r_X^2$ and $\beta_W \propto r_W^2$ (Appendix S1: Figure S4). The r^2 values were Pearson correlation for normal data, Spearman rank correlation for counts data, and point-biserial correlation (as defined in the R package *ltm*; Rizopoulos, 2006) for binary data. $[X]$ and $[W]$ represent the reference fractions against which we compared the values of the variation fractions according to the different methods in Exercise 2 (see below).

We considered different scenarios by varying the type of data (normal, counts, or binary), the size of the grid ($N = 25, 100, \text{ or } 400$ sites), the type of spatial variogram model of W (exponential or Gaussian), the autocorrelation range of both W and X ($A = 0.01N, 0.5N, \text{ or } N$), and the type of response to the environment, either linear or bell-shaped with varying niche breadth ($2\sigma^2 = 0.002, 0.02, \text{ or } 0.2$) (see also Appendix S1: Table S1). The reference scenario, while one target parameter varied, was $N = 100, A = 0.5N$ and $2\sigma^2 = 0.02$. In addition, we considered a scenario where a random subsample (50 out of 400 sites) was taken to simulate a sampling effect, and another scenario where either three or six random environmental variables orthogonal to the response (i.e., noise) were added to inspect overfitting propensity. The number of species was 20 in all simulations. For each combination of simulation parameters (Appendix S1: Table S2), we simulated a range of predictive weights β_X from 0 to 1 increasing by 0.1, determining the relative importance of environment and space, since $\beta_W = 1 - \beta_X$. Each combination of parameters and β_X was replicated five times, resulting in a total of 3960 simulation runs.

Statistical methods

We considered methods in four broad families of statistical models (Table 1): constrained ordination, GLMs, GAMs, and tree-based ML. These methods can fit species–environment and species–space relationships in a

variety of ways as different environmental and spatial predictors can be used to estimate the same environmental and spatial effects. Some, but not all, techniques can be used with more than one method to model space and environment (see Appendix S1: Table S3)—for example, recursive splitting of response variables along spatial coordinates can only be used with ML, and not with RDA, GLM, or GAM, but MEMs can be used in RDA, GLM, and ML; thus, we could not consider all combinations. Further, each of these methods can be used in a variety of ways by tuning the respective algorithms, but these fine-tunings should be chosen according to the type of data at hand and goals of the study. For our comparison, we tuned the methods based on the simulated types of data and our own experience while performing preliminary analyses and explorations.

Constrained ordination

We used RDA (Rao, 1964) as the method representing the broad family of multivariate methods. In RDA, the response variables are modeled with ordinary linear regressions (one for each species), and then, an ordination is performed on the predictions, although we did not need to perform the ordination step for our purposes. The difference to GLMs is that in RDA only ordinary linear regression is used, and although not necessary, the species data are usually first transformed to allow for an unbiased analysis of community composition gradients (Legendre & Gallagher, 2001). We used the Hellinger transformation (RDA-Hel) as recommended in Legendre and Gallagher (2001) and Legendre and Legendre (2012) to downweight the effect of rare species. Still, because this data transformation consists in dividing each value by the sum of abundances of the respective site (i.e., the row sum) and taking its square root, which might somehow distort the simulated data and break the underlying simulated relationships, we also performed RDA on raw data to make it comparable to all the remaining methods (RDA-raw). The response to the environment was modeled as either a linear or a second-degree polynomial, depending on the simulated data, and when the simulated response to the environment was bell-shaped, we log-transformed the data prior to the Hellinger transformation to better approximate the shape of the response (see Appendix S1: Figure S5). The spatial effects were modeled with distance-based MEM variables using the function “dbmem” (with default settings) of the R package *adespatial* (Dray et al., 2016). Moran’s eigenvector map variables represent the spatial autocorrelation across sampled sites at different spatial scales (i.e., different grains of autocorrelation). We used the R function “lm” from the R *stats*

package (R Development Core Team, 2020) to fit the linear models underlying RDA.

Generalized linear models

All GLMs were fitted as single models to each species, and the predictions were then stacked into a matrix of predicted species abundances or occurrences. The error distribution of the model was defined according to the type of simulated data: a normal (Gaussian) distribution for normal errors, quasi-Poisson for counts data (due to best estimation convergence compared to Poisson and negative binomial), and binomial/Bernoulli distribution for binary data. We fitted either a linear or a second-degree polynomial to model the response to the environment (depending on the simulated environmental component X). We used MEMs (calculated as described for RDA) to model the spatial effects. The GLMs were fitted using the *stats* R package (R Development Core Team, 2020).

Generalized additive models

Generalized additive models were fitted using the R package *mgcv* (Wood, 2017). We fitted either a linear effect or a thin-plate spline with $k = 3$ dimensions (for the case of unimodal responses) to model the environmental effect, and 2D thin-plate splines with either a default k (that depends on the number of predictors; GAM-kdef) or fixed $k = 10$ on the spatial coordinates of the sites to model the spatial effect (GAM-k10). The different k allowed us to compare different initial degrees of “wiggleness” of the basis functions. The smoothing parameters were estimated by restricted maximum likelihood. Similarly to GLM, the error distributions were tailored to the type of data, but unlike in GLM, the Poisson distribution was preferred for counts data due to good convergence.

Machine learning: Tree-based methods

Tree-based methods (Hastie et al., 2009) recursively split the response along a set of predictor variables, resulting in one tree or multiple trees (i.e., a “forest”). In contrast to the previous methods, tree-based ML makes no assumption regarding the functional form of the SERs; instead, the relationship is learned from the data. As such, the environmental predictor was E and the spatial predictors were either the spatial coordinates (XY) or MEMs (see model specifications and settings in Table 1).

When spatial coordinates were used, we tried two variants of each method by defining two different values for the algorithm parameter that mostly affected the smoothness of the fitted model (see below), according to a priori trial-error explorations.

Boosted regression trees

This method uses a gradient boosting algorithm of Friedman (2001), which fits, to each species separately, a sequence of regression trees, where each new tree is applied to the residuals from the previous tree (Miller et al., 2016). We used the R function “mvtb” (package *mvtboost*), which is fitted according to the algorithm presented in the R package *gbm* (and documentation therein; Greenwell et al., 2020), and fitted two variants of the algorithm—shrinkage (or “learning rate”) set to either 0.01 (BRT-lr0.01) or 0.1 (BRT-lr0.1). We set the tree depth to 2 to allow the model to fit complex spatial structures resulting from the interaction between the spatial coordinates (X , Y), except for the model with MEMs (BRT-MEM), where we used a tree depth of 1 to avoid interactions between MEMs. The model with MEMs used a sample learning rate of 0.01.

Univariate random forest (UniRF)

This fits a UniRF (Breiman, 2001; Hastie et al., 2009) to each species individually, using the “randomForest” R function (package *randomForest*). A RF consists of a set of regression trees fitted to bootstrapped data (i.e., sampled with replacement), each tree fitted to a random fraction of predictors. Predictions of the individual trees are then averaged to get the overall prediction. The UniRF method is identical to the method called Gradient Forest implemented in package *gradientForest* (Ellis et al., 2012). We set the algorithm to 500 regression trees, each with a sample size of either N (the sample size of the data; UniRF-SS0) or $N/5$ (i.e., 20% of the data; UniRF-SS20), each fitted to data resampled randomly with replacement, using a random subset (1/3) of the predictors. The sample size was observed to influence the smoothness of the fitted model, thus being tightly related to the control of overfitting. The model with MEMs (UniRF-MEM) used a sample size of N .

Multivariate random forest (MVRF)

Similarly to the univariate version, this multivariate version fits a regression tree to each species, but now all happens within a single function call, and the split rule is the composite normalized mean-squared error, where each component (species) of the composite is normalized so that the mean abundance of the species does not influence the split rule. We used the R function “rfsrc” (package *randomForestSRC*; Ishwaran & Kogalur, 2019). We used the minimum number of observations in any terminal (“leaf”) node fixed to 5 and followed the default settings of the R

function so that each tree in the forest was fitted to data resampled randomly with replacement, and the number of randomly chosen predictors in each tree was the square root of the total number of predictors. Such as for UniRF, we varied the sample size to be either N (the sample size of the data; MVRF-SS0) or $N/5$ (i.e., 20% of the data; MVRF-SS20) in order to assess different “smoothness” settings. Here again, the model with MEMs used a sample size of N .

Multivariate regression trees (MVRT)

This method fits a single MVRT (De’ath, 2002) to explain the abundance or occurrence data. We used its implementation in R function “mvpart” (package *mvpart*). We fitted a tree with the minimum number of observations in any terminal (“leaf”) node fixed to 5. We fitted the models with (MVRT-CV) or without (MVRT-noCV) internal cross-validation. The model with MEMs (MVRT-MEM) was estimated without cross-validation.

Assessment of method performance with simulated data

Our study comprised two exercises. In Exercise 1, we assessed the ability of the methods presented in Table 1 to approximate the simulated responses of species to the environment, as well as the spatial structure of their distributions. In Exercise 2, we partitioned the variation in multispecies abundance or occurrence caused by the spatial and environmental effects, and compared these to the simulated (i.e., reference) spatial and environmental fractions of explained variation.

Exercise 1. Model fitting performance

We applied each of the methods (Table 1) on simulated data and obtained the model predictions (\hat{Y}). The fitting performance was calculated by the correlation between the simulated abundance or occurrence values before adding the error structure (\bar{Y} ; i.e., the true model) with the model predictions \hat{Y} . However, deviations from the true model could be due to under- or overfitting: If data are overfitted, the predictions will be (wrongly) closer to the simulated response values Y (as the added error will also be fitted to some extent) and the correlation between Y and \hat{Y} will be higher than the correlation between Y and \bar{Y} ; if data are underfitted, the correlation between Y and \hat{Y} will be lower than that between Y and \bar{Y} . We explicitly assessed these causes of underperformance. The performance was assessed for models estimating environmental and spatial effects jointly, environmental effects alone (i.e., when $\beta_X = 1$

and $\beta_W = 0$; but note that it includes the effect of spatially autocorrelated environment), and spatial effects alone (i.e., when $\beta_X = 0$ and $\beta_W = 1$). RDA-Hel was excluded from this exercise, because the transformation of the response data makes this model incomparable with the simulated data and thus with the other methods (though we assessed its performance in *Exercise 2*, as the variation fractions are always on the same 0–1 scale).

Exercise 2. Variation partitioning performance

We performed variation partitioning with each of the methods according to the same procedure as for the simulated data (Equations (4)–(7)), but now we used the model predictions to calculate the r^2 and obtain the fractions of explained variation $[\hat{X}]$ and $[\hat{W}]$:

$$r_{XW}^2 = r\left(Y_i, \hat{Y}_{XW,i}\right)^2, \quad (8)$$

$$r_X^2 = r\left(Y_i, \hat{Y}_{X,i}\right)^2, \quad (9)$$

$$[\hat{X}] = r_X^2, \quad (10)$$

$$[\hat{W}] = r_{XW}^2 - r_X^2. \quad (11)$$

The predictions $\hat{Y}_{XW,i}$ correspond to the full model with both environment and space as predictors, and the predictions $\hat{Y}_{X,i}$ correspond to the model with the environmental predictor alone. Note that the spatial predictors in the statistical models also estimate the effect of the environment that is spatially autocorrelated; thus, $[\hat{X}]$ also contains the variation attributed to the effect of spatially correlated environment. As such, the shared fraction of variation estimated from the statistical models was systematically higher than the shared fraction of the simulated data. This is why we compare the total (pure + shared) environmental fraction of variation. The performance of each method under the different simulation scenarios was assessed by comparing the estimated versus simulated fractions of variation. To keep the comparison fair, we also used squared correlation coefficients (r^2) for the estimated variation fractions. To check possible effects of the choice of goodness-of-fit metric, we also performed variation partitioning using R^2 or deviance-based pseudo- R^2 metrics depending on the type of data and method. For RDA and GLM, we adjusted the r^2 to account for the number of predictors (as recommended in Peres-Neto et al., 2006). Negative fractions of variation were set to 0.

Comparison of methods using empirical data

To explore the impacts of method choice on empirical results, we compared the results of the variation partitioning performed by the different methods (Table 1) on nine empirical datasets (see Appendix S2). Each empirical dataset consisted of a site-by-species abundance matrix, some environmental variables, and geographical coordinates of the sites. For GLMs, we included both linear and quadratic effects for the environmental predictors. Because the sample size of these data tended to be small ($29 < N < 138$), we limited the number of environmental predictors by randomly choosing three continuous predictors that were the same across the different models.

RESULTS

Exercise 1. Model fitting performance

We observed clear differences in fitting performance among the different methods (Figure 1). Most methods tended to underfit the data when both environmental and spatial responses were included, including GAM, UniRF, MVRF-SS20, and MVRT (Figure 1a–c), which was mostly caused by underfitted simulated spatial structures (Figure 1g–i). Environmental responses were better approximated in general (Figure 1d–f) than spatial responses, but UniRF and MVRF with more smoothness (UniRF-SS20 and MVRF-SS20), the MVRT methods, RDA, GLM, and GAM for normal data did not accurately fit the environmental responses. RDA, as well as GLM and GAM for normal data, could not fit the bell-shaped responses adequately (see also Appendix S1: Figure S5b), and BRT with lower smoothness (BRT-lr0.1) tended to slightly overfit the environmental response. The spatial responses were generally underfitted, particularly in GAMs, UniRF, MVRF-SS20, MVRT-noCV, and MVRT-CV (Figure 1g–i). The other models were better fitted, although GLM, BRT-lr0.1, BRT-MEM, and MVRF-MEM tended to overfit the simulated spatial structures. The methods using MEMs were generally more efficient in modeling spatial structure, except UniRF-MEM.

The type of environmental response and spatial structure also caused fitting differences (Figures 2 and 3). The methods RDA, GLM, GAM, and MVRF-SS20, and to a lesser extent UniRF and MVRT, showed decreasing fitting performance as the shape of the bell-shaped environmental response narrowed, whereas BRT and MVRF with lower smoothness (MVRF-SS0 and MVRF-MEM) could better fit these narrow responses, despite their propensity to overfit

the narrowest environmental response in some situations (Figure 2). For the spatial effects, the fitting performance increased with the smoothness of the spatial pattern (Figure 3). For a nearly random spatial structure ($A = 0.01N$), for which we expected that the models would not fit the data (i.e., we expected simulated–predicted correlations to be close to 0, and thus a score closer to -1 in Figure 3), we observed that the models with MEMs overfitted the data (except UniRF-MEM), BRT and MVRF-SS0 overfitted the data to some extent, whereas GAM and UniRF correctly did not overfit into this random variation.

Exercise 2. Variation partitioning

The variation partitioning performance, that is, the difference between the estimated ($[\hat{X}]$, $[\hat{W}]$) and simulated ($[X]$, $[W]$) fractions, varied considerably across methods, but was generally consistent across the different types of data (Figures 4 and 5). None of the methods perfectly recovered both the spatial and environmental fractions of the simulated reference variation partitioning. Overall, GLMs and BRT-MEM were the most consistent with the simulated variation partitioning, particularly in the estimation of the spatial fraction (Figures 4 and 5; RDA-raw performed as well as GLM and BRT-MEM for the spatial fraction, but it was worse for the environmental fraction). For the spatial effects, consistently with the fitting performance assessed in Exercise 1, variation fractions were in general underestimated. The results were qualitatively similar when using R^2 or deviance-based pseudo- R^2 metrics for partitioning the variation (Appendix S1: Figures S7 and S8). The performance for each individual combination of simulation parameters can be seen in Appendix S1: Figures S9–S20 (equivalent to Figure 5 but separately plotted for each case). Even though the shared variation fraction was not directly evaluated (as it was attributed to the environmental fraction), the methods were robust to the amount of environmental autocorrelation, as variation in performance was low within methods (compared to that among methods) when the level of spatial autocorrelation in the environment varied (Appendix S1: Figure S21).

Redundancy analysis underestimated the environmental variation fraction (Figure 4), especially for low sample size ($N = 25$), narrow niche breadth, and when random environmental predictors were added (Figure 5). Redundancy analysis also generally underestimated the spatial fraction, but estimated it better than the environmental fraction. We saw similar results in GLM (Figures 4 and 5), but the GLM could better deal with bell-shaped environmental responses in both counts and binary data (see also Figure 1). The main difference between GLM and RDA was

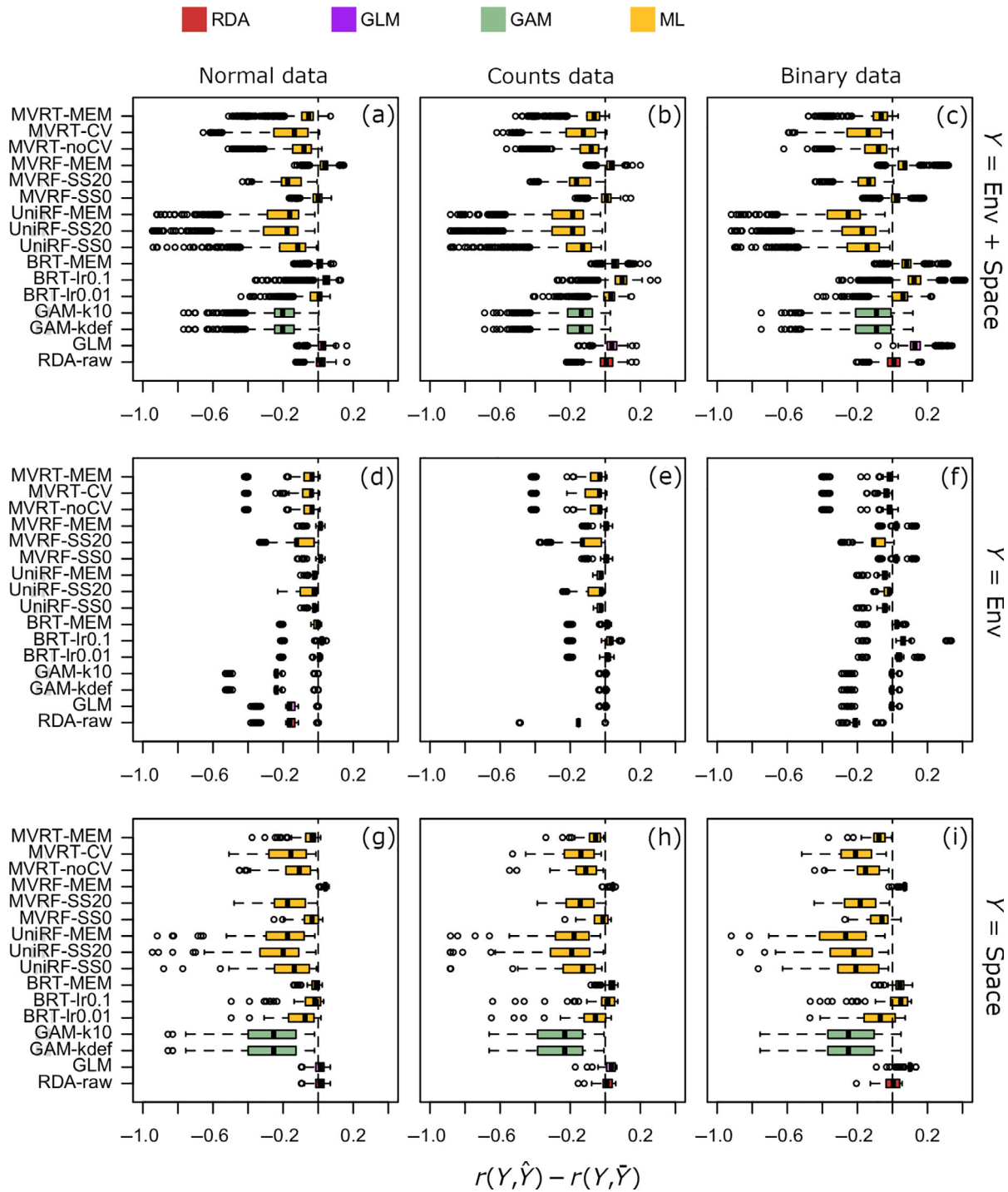


FIGURE 1 Fitting performance of the different statistical models for the additive effect of environment and space (a–c, each boxplot based on 1265 simulations, cases 1–6, and 8–13; Appendix S1: Table S2), the environmental effect alone (d–f, $\beta_X = 1$; each boxplot based on 100 simulations, cases 1–10, and 13), and the spatial effect alone (g–i, $\beta_W = 1$; each boxplot based on 115 simulations, cases 1–6, and 8–13), fitted to normal, counts, and binary data, across the different generated data (see Appendix S1: Tables S1 and S2). The performance is given by the correlation between the simulated response values (Y) and the model predictions (\hat{Y}) minus the correlation between the simulated response values (Y) and the simulated deterministic response corresponding to the true model (\bar{Y}). The dashed line represents a perfect fit (i.e., $\hat{Y} = \bar{Y}$), positive values indicate overfitting, and negative values indicate underfitting. BRT, boosted regression trees; CV, internal cross-validation (with, “CV,” or without, “noCV”); GAM, generalized additive models; GLM, generalized linear models; k , dimension of the spline basis function (default, “def,” or with fixed $k = 10$, “k10”); lr, learning rate (0.01 or 0.1); MEMs, Moran’s eigenvector maps; ML, machine learning; MVRF, multivariate random forest; MVRT, multivariate regression trees; RDA, redundancy analysis; SS, resample size (0 for $SS = N$ and 20 for $SS = N/5$); UniRF, univariate random forest

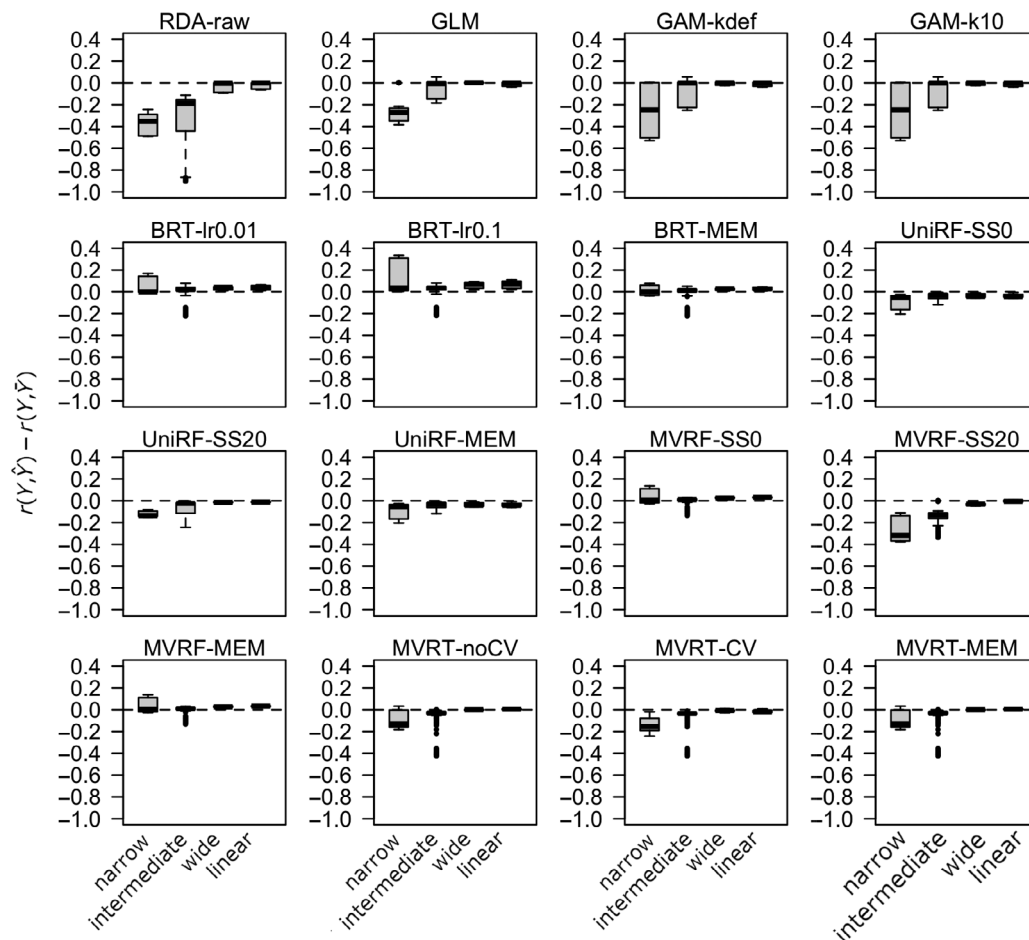


FIGURE 2 Fitting performance of the different statistical models for the different environmental effects: unimodal with different niche breadths (narrow = 0.002, intermediate = 0.02, and broad = 0.2) and linear (see also Appendix S1: Figure S1). The results are pooled for all types of response data (i.e., normal, counts, and binary data), each boxplot based on 360 simulations ($\beta_x = 1$). The performance is given by the correlation between the observed response values (Y) and the model predictions (\hat{Y}) minus the correlation between the simulated response values (Y) and the simulated deterministic response corresponding to the true model (\bar{Y}). The dashed line represents a perfect fit (i.e., $\hat{Y} = \bar{Y}$), positive values indicate overfitting, and negative values indicate underfitting. BRT, boosted regression trees; CV, internal cross-validation (with, “CV,” or without, “noCV”); GAM, generalized additive models; GLM, generalized linear models; k , dimension of the spline basis function (default, “def,” or with fixed $k = 10$, “k10”); lr, learning rate (0.01 or 0.1); MEMs, Moran’s eigenvector maps; MVRF, multivariate random forest; MVRT, multivariate regression trees; RDA, redundancy analysis; SS, resample size (0 for SS = N and 20 for SS = $N/5$); UniRF, univariate random forest

that GLM tended to overestimate the spatial fraction for binary data (Figures 4 and 5).

Generalized additive models underestimated the spatial fraction but generally correctly estimated the environmental fraction (Figures 4 and 5). Even though GAMs used flexible splines for the environmental response, the narrowest responses were still underestimated (Figure 5). We did not find differences between the number of dimensions of the basis functions of the splines (we compared the default or $k = 10$).

The performance of ML differed across its different methods, as well as between the variants of each method depending on the type of spatial variables used and the parameters controlling the smoothness of the models. All

ML methods tended to overestimate the environmental fraction, except for UniRF and MVRT with cross-validation (MVRT-CV) (Figure 4). The tendency of BRT and MVRF to overfit the environmental responses (see above and Figure 1) was also reflected in overestimation of the environmental fraction and underestimation of the spatial fraction (Figure 4). In addition, BRT was particularly sensitive to sample size and the number of environmental covariates, overestimating the environmental fraction when we added environmental covariates unrelated to the response. These problems were less severe for BRT-MEM. The spatial fraction was generally underestimated with all ML methods, although only slightly for BRT with MEMs (Figures 4 and 5). Boosted

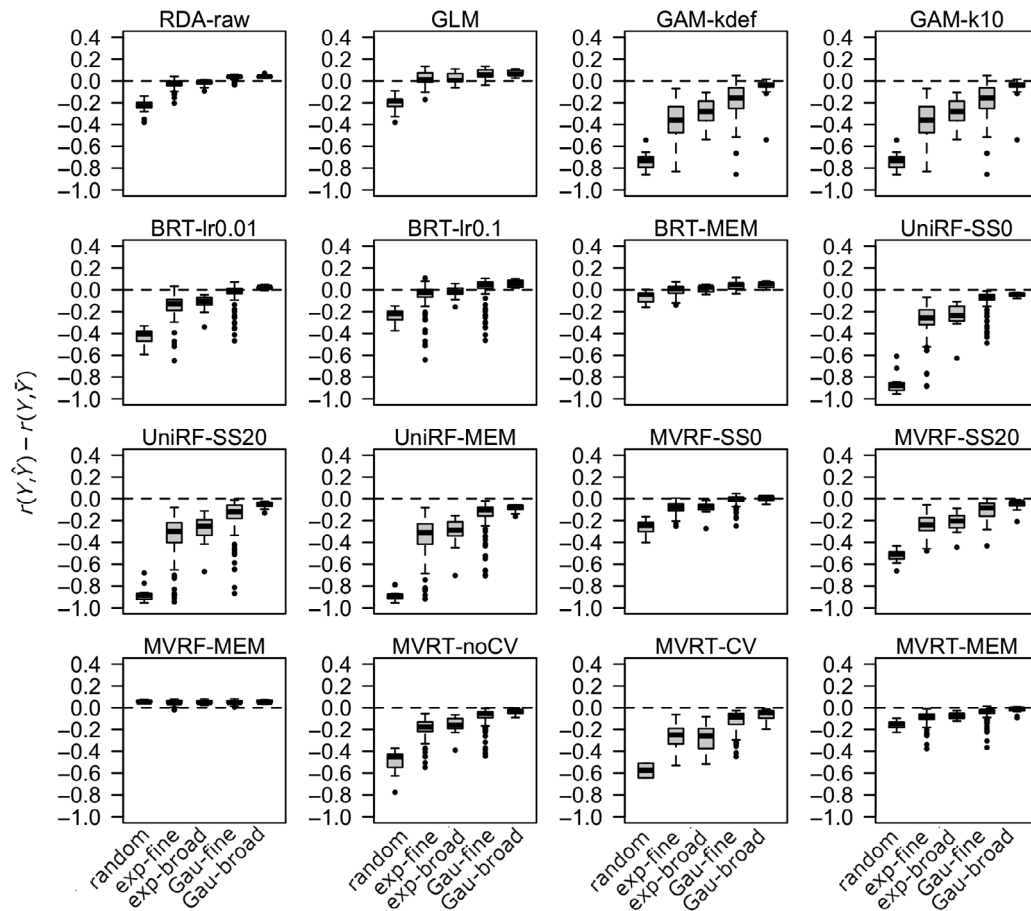


FIGURE 3 Ability of the different statistical models to fit different spatial structures: exponential (exp) and Gaussian (Gau) variogram models with $A = 0.01$, that is, nearly random spatial structure, $A = 0.5N$ or $A = N$. The smoothness of the spatial structure increases to the right (see also Appendix S1: Figure S2). The results are pooled for all types of response data (i.e., normal, counts, and binary data), each boxplot based on 360 simulations ($\beta_w = 1$). The performance is given by the correlation between the observed response values (Y) and the model predictions (\hat{Y}) minus the correlation between the simulated response values (Y) and the simulated deterministic response corresponding to the true model (\bar{Y}). The dashed line represents a perfect fit (i.e., $\hat{Y} = \bar{Y}$), positive values indicate overfitting, and negative values indicate underfitting. BRT, boosted regression trees; CV, internal cross-validation (with, “CV,” or without, “noCV”); GAM, generalized additive models; GLM, generalized linear models; k , dimension of the spline basis function (default, “def,” or with fixed $k = 10$, “k10”); lr, learning rate (0.01 or 0.1); MEMs, Moran’s eigenvector maps; MVRF, multivariate random forest; MVRT, multivariate regression trees; RDA, redundancy analysis; SS, resample size (0 for $SS = N$ and 20 for $SS = N/5$); UniRF, univariate random forest

regression trees with MEMs was the method that performed consistently better across all simulated scenarios (Figures 4 and 5).

Empirical data

The choice of method had a clear influence on the results of the variation partitioning on empirical data (Figure 6). Fitting GLM, MVRF-SS20, and MVRT-CV led to convergence or estimation problems for some datasets, likely because of the large number of spatial predictors relative to sample size in GLMs, and because of insufficient sample size for recursively splitting data and/or performing cross-validation in MVRF-SS20 and

MVRT-CV. The trends in the estimation of environmental variation fractions across the different datasets were mostly consistent across the different methods (high pairwise correlations among methods), except for RDA and BRT-lr0.1, which were less correlated with the other methods. Regarding the estimation of the spatial variation fraction, the methods were highly inconsistent. The methods that provided more consistent results were BRT with MVRF and to a lesser extent GAM with BRT. Even though the total variation explained is overall similar across the different methods, there are strong differences in the variation partitioning outputs, namely, in the estimation of the spatial effects and how they are attributed to autocorrelated environment or pure space.

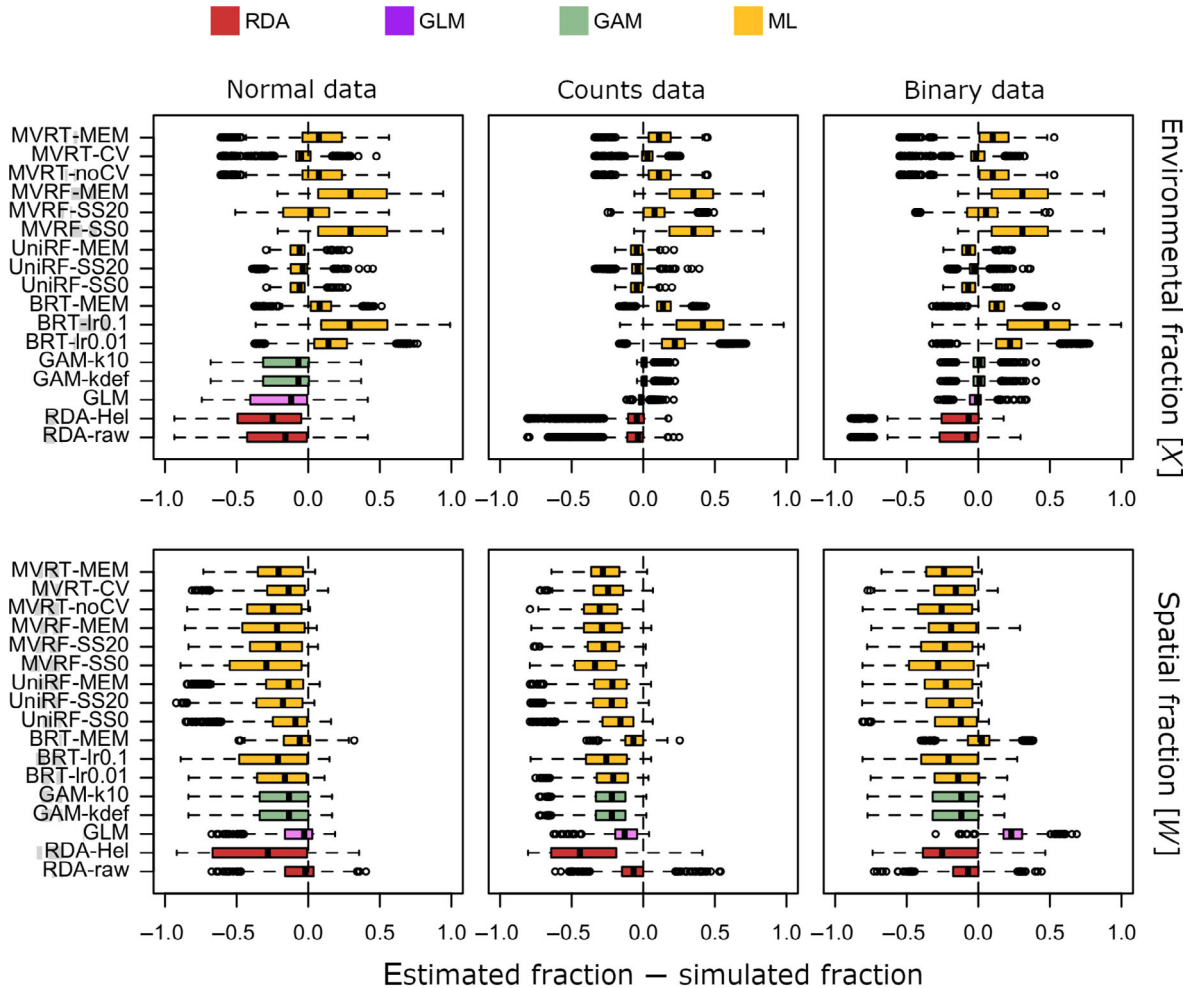


FIGURE 4 Variation partitioning performance of the different methods for normal, counts, and binary data across the different generated data (each boxplot based on 1265 simulations, cases 1–6, and 8–13; Appendix S1: Table S2). The performance is given by the difference between the estimated and simulated fractions of variation explained by the environment ($[\hat{X}] - [X]$; upper panels) and space ($[\hat{W}] - [W]$; lower panels). The dashed line represents equality (simulated = estimated), positive values indicate overestimation, and negative values indicate underestimation. BRT, boosted regression trees; CV, internal cross-validation (with, “CV,” or without, “noCV”); GAM, generalized additive models; GLM, generalized linear models; k , dimension of the spline basis function (default, “def,” or with fixed $k = 10$, “k10”); lr, learning rate (0.01 or 0.1); MEMs, Moran’s eigenvector maps; ML, machine learning; MVRF, multivariate random forest; MVRT, multivariate regression trees; RDA, redundancy analysis; SS, resample size (0 for $SS = N$ and 20 for $SS = N/5$); UniRF, univariate random forest

DISCUSSION

Our simulations show that it is challenging to recommend a single universal method for modeling spatial and environmental effects that can be used under any circumstances. All methods assessed here have advantages and drawbacks, and none of the methods provides a perfect model that always correctly fits both the environmental and spatial responses. Generalized linear models and BRT with MEMs provided balanced fits that consistently disentangled the spatial and environmental sources of variation in simulated species abundance and occurrence. This highlights the potential of GLMs as an alternative to popular constrained ordination methods, and of

BRT combined with powerful spatial predictors (MEMs) as a flexible, yet not widely deployed, method for ecologists aiming at disentangling environmental and spatial effects. Generalized linear models are particularly useful when specific hypotheses about SERs are considered, as we can limit the scope of modeled responses to, for instance, linear and unimodal. In contrast, BRT can fit any kind of response, with its shape completely learned from the data, and is an effective solution if there is no specific hypothesis about SERs.

Generalized additive models are a good compromise, as modeled SERs can be limited to parametric response shapes (e.g., polynomials or splines with a limited number of basis functions, k), while spatial effects can be

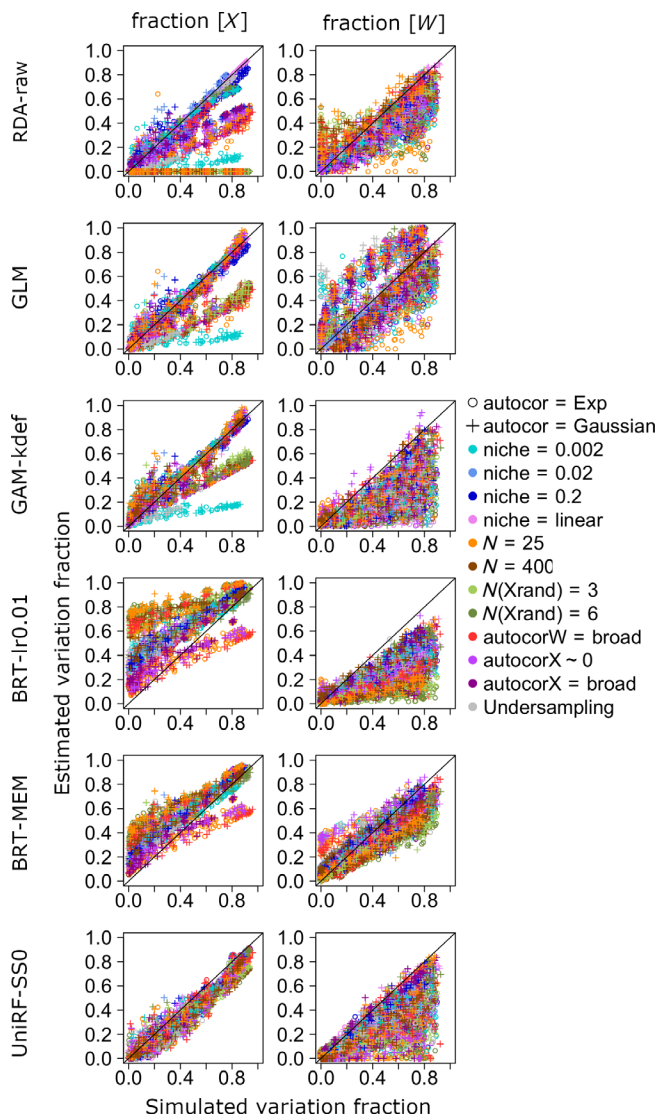


FIGURE 5 Plots of simulated versus estimated fractions of variation explained by the environment ($[X]$ vs. $[\hat{X}]$; left panels) and space ($[W]$ vs. $[\hat{W}]$; right panels) across different methods (different rows) and generated data (different colors). Each case (in color) is replicated five times in each value of β_X (left panels) or β_W (right panels). The black line represents the 1:1 line, where points should fall if the estimated fraction is equal to the simulated fraction. BRT, boosted regression trees; GAM, generalized additive models; GLM, generalized linear models; k , dimension of the spline basis function (default, “def”); lr, learning rate (0.01 or 0.1); MEMs, Moran’s eigenvector maps; RDA, redundancy analysis; SS, resample size (0 for $SS = N$); UniRF, univariate random forest

fitted using flexible splines on spatial coordinates. Because GAMs tend to fit smoother spatial surfaces in comparison with other methods, they might be more transferable, for example, if the goal is to predict the response in unsampled sites. Note, however, that GAMs underestimated spatial fractions when the simulated spatial responses were derived from short variogram ranges

and thus “wobblier” than the fitted GAM splines. To allow GAMs to be flexible enough to model complex responses, the k and spline smoothness can be adjusted prior to the modeling, for example, by comparing information criteria of models with different k (Wood, 2017).

Tree-based ML methods are flexible and user-friendly, and are useful when we lack specific hypotheses about the functional forms of SERs. Furthermore, tree-based ML inherently models interactions between predictors, a desirable property when applied to complex ecological systems. However, it was challenging to parameterize ML algorithms to work well both for fitting the environmental and spatial components. Simply using spatial coordinates to model space turned out to underfit most spatial structures, but forcing the algorithms to be less smooth often led to overfitting of the environmental component. The use of MEMs in BRT models can offer a more balanced solution for jointly estimating environmental and spatial effects. However, this was not the case for other ML methods, possibly due to a failure to control for interactions between MEMs in our tested settings (note that BRT can be used with no interactions). Overall, although ML showed potential to be used for variation partitioning, we still need to learn more about how to fine-tune the ML parameters for the purpose of fitting spatial structures and environmental responses within the same model, in particular parameters that affect the smoothness of the model such as sample size in RF, or learning rate in BRT.

In general, modeling spatial structure was a challenge. Among the spatial structures that we considered, those generated with exponential variograms (Appendix S1: Figure S2) tended to be underfitted, whereas nearly random spatial structures tended to be overfitted by the methods that better fitted the former (Figure 2). We suggest to choose the type of spatial model depending on the processes involved (e.g., dispersal limitation or connectivity) and goals of the study. For example, MEMs are useful for interpretation purposes and infer spatial scales of variation (Dray et al., 2012; Murakami & Griffith, 2019). Generalized additive models, on the other hand, can fit smooth spatial surfaces (e.g., Appendix S1: Figure S6), such as those simulated via a Gaussian variogram model (Appendix S1: Figure S2), which can be suitable for interpolation. And ML can offer flexible solutions, but a priori explorations, for example, through simulations, might be needed to determine how flexible the spatial model needs to be.

We also warn that the models considered in our study are suited for modeling environmental and spatial effects only within the study area, as fitted spatial surfaces are hardly transferable in space. If the goal is to predict outside the study area (i.e., extrapolation), as is often the case in species distribution models, the use of

autoregressive models (Dormann et al., 2007) might be a better alternative, though we are unaware of their use for variation partitioning. Alternatively, if the goal is to estimate the total fraction of variation explained by the environment independently of space (i.e., if spatial effects are

not of interest), then spatially blocked cross-validation is a good alternative (Roberts et al., 2017). Spatial blocking can effectively account for spatial autocorrelation, even though the spatial effect is not modeled. If the goal is to perform variation partitioning, note that spurious environmental effects arise from spatial autocorrelation as the number of spatially autocorrelated environmental predictors increases (Chapman, 2010). These spurious effects should be accounted for and attributed to pure space rather than to the shared fraction of variation (Clappe et al., 2018 and the methods therein).

Our results show that there is margin for improvement. For example, machine learning methods are diverse, and more methods can be tested and even developed specifically for the purposes outlined here (e.g., D’Amen et al., 2017; Nieto-Lugilde et al., 2018). Also, joint species distribution models (JSDM), which are based on generalized linear modeling, are becoming popular in community ecology (e.g., Ovaskainen et al., 2017), but their performance to model different spatial structures remains poorly explored. It is also possible to combine ML techniques with JSDMs, which offer a more flexible approach to model nonlinear responses while using JSDM machinery (Harris, 2015). Other possibilities include the development of better cross-validation procedures to partition explained variation, and the development of R^2 adjustment procedures, in particular for ML methods such as BRT, since the addition of irrelevant covariates in the models resulted in the overestimation of the environmental fraction of variation.

In conclusion, we provide a comprehensive assessment of different methods to use in ecology when we need to jointly model environmental and spatial effects. As the particular shapes of species responses to the

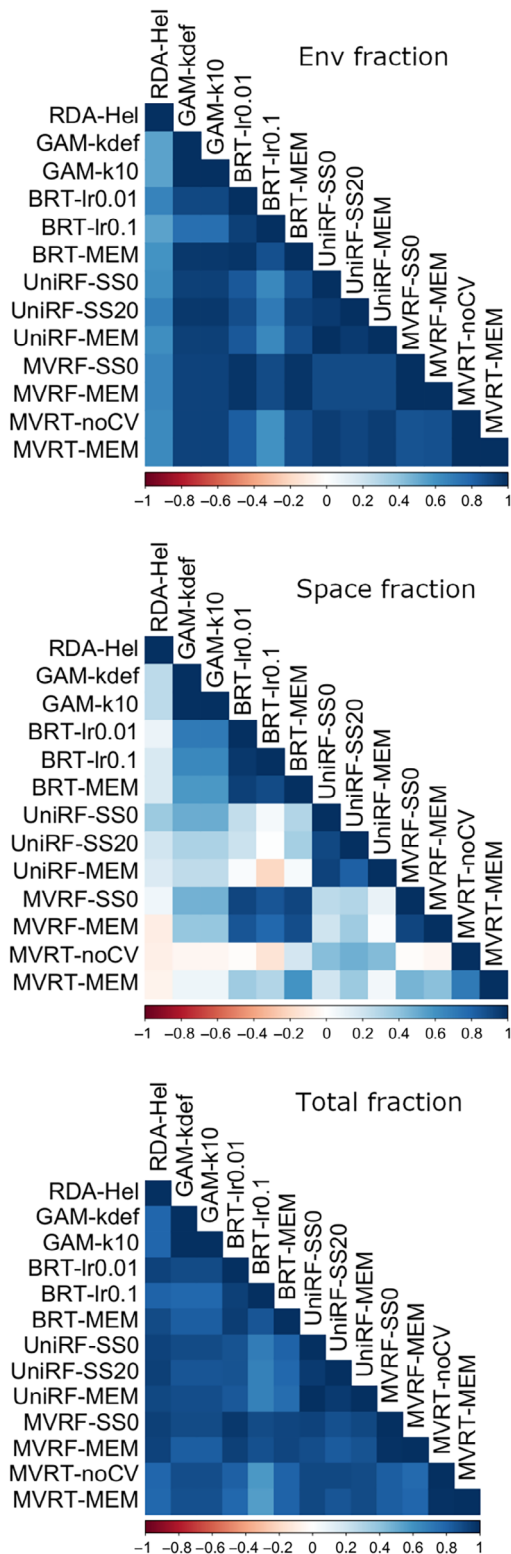


FIGURE 6 Consistency of the different methods in estimating fractions of variation explained by the environment (left panel), space (middle panel), and the combined, total effect (right panel) applied on nine empirical datasets. The pairwise comparison consists of the correlation of the estimated variation fraction values by each pair of methods (the color gradient indicates the correlation value, from -1 in dark red to 1 in dark blue). See Appendix S2 for references of the datasets. BRT, boosted regression trees; CV, internal cross-validation (with, “CV,” or without, “noCV”); GAM, generalized additive models; k , dimension of the spline basis function (default, “def,” or with fixed $k = 10$, “k10”); lr, learning rate (0.01 or 0.1); MEMs, Moran’s eigenvector maps; MVRF, multivariate random forest; MVRT, multivariate regression trees; RDA, redundancy analysis; SS, resample size (0 for $SS = N$ and 20 for $SS = N/5$); UniRF, univariate random forest

environment are hard to hypothesize in the context of community or multispecies data, we introduce tree-based ML as a flexible method that can be widely used with both abundance and occurrence data. If a priori hypotheses about SERs are considered, GLM as a parametric method is a reasonable choice for variation partitioning. The most important message of this study is not to recommend one “best” method, but to bring up a whole suite of possible methods that are often overlooked, such as ML methods. The simulations and statistical approaches covered in this study can be adapted to explore the performance of different models under more specific research questions, study systems, and data types. We hope that this will inspire a new generation of analyses that are better tailored to the data and questions at hand. By choosing appropriate methods to model different responses to the environment and different spatial structures, with typical ecological data such as species abundances and distributions, species diversity, and community composition, our recommendations apply to community ecology, biogeography, and macroecology studies.

ACKNOWLEDGMENTS

We thank Jonathan Chase, Pedro Peres-Neto, and three anonymous reviewers for important discussions and comments. The work was supported by the German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig funded by the German Research Foundation (FZT 118). DSV was supported by sDiv, the Synthesis Centre of iDiv. PK was supported by Research Excellence in Environmental Sciences (REES) grant from Faculty of Environmental Sciences, Czech University of Life Sciences in Prague.

CONFLICT OF INTEREST

The authors declare no conflict of interest.


DATA AVAILABILITY STATEMENT

The code (Viana, 2022) utilized to perform the analyses and generate the results is novel and is available from Zenodo: <https://doi.org/10.5281/zenodo.6240353>. The datasets utilized for this research are publicly available and cited in Appendix S2.

ORCID

Duarte S. Viana  <https://orcid.org/0000-0002-7864-0871>

Petr Keil  <https://orcid.org/0000-0003-3017-1858>

Alienor Jeliązkov  <https://orcid.org/0000-0001-5765-3721>

REFERENCES

- Bar-Massada, A. 2015. “Complex Relationships between Species Niches and Environmental Heterogeneity Affect Species Co-Occurrence Patterns in Modelled and Real Communities.” *Proceedings of the Royal Society B: Biological Sciences* 282: 20150927.
- Breiman, L. 2001. “Random Forests.” *Machine Learning* 45: 5–32.
- Chapman, D. S. 2010. “Weak Climatic Associations among British Plant Distributions.” *Global Ecology and Biogeography* 19: 831–41.
- Chase, J. M., and M. A. Leibold. 2003. *Ecological Niches: Linking Classical and Contemporary Approaches*. Chicago: University of Chicago Press.
- Chesson, P. 2000. “Mechanisms of Maintenance of Species Diversity.” *Annual Review of Ecology and Systematics* 31: 343–66.
- Chesson, P. L., and R. R. Warner. 1981. “Environmental Variability Promotes Coexistence in Lottery Competitive Systems.” *The American Naturalist* 117: 923–43.
- Clappe, S., S. Dray, and P. R. Peres-Neto. 2018. “Beyond Neutrality: Disentangling the Effects of Species Sorting and Spurious Correlations in Community Analysis.” *Ecology* 99: 1737–47.
- Cottenie, K. 2005. “Integrating Environmental and Spatial Processes in Ecological Community Dynamics.” *Ecology Letters* 8: 1175–82.
- D’Amen, M., H. K. Mod, N. J. Gotelli, and A. Guisan. 2018. “Disentangling Biotic Interactions, Environmental Filters, and Dispersal Limitation as Drivers of Species Co-Occurrence.” *Ecography* 41: 1233–44.
- D’Amen, M., C. Rahbek, N. E. Zimmermann, and A. Guisan. 2017. “Spatial Predictions at the Community Level: From Current Approaches to Future Frameworks.” *Biological Reviews* 92: 169–87.
- De’ath, G. 2002. “Multivariate Regression Trees: A New Technique for Modeling Species-Environment Relationships.” *Ecology* 83: 1105–17.
- Dormann, C. F., J. M. McPherson, M. B. Araújo, R. Bivand, J. Bolliger, G. Carl, R. G. Davies, et al. 2007. “Methods to Account for Spatial Autocorrelation in the Analysis of Species Distributional Data: A Review.” *Ecography* 30: 609–28.
- Dray, A. S., G. Blanchet, D. Borcard, G. Guenard, T. Jombart, G. Larocque, P. Legendre, N. Madi, and H. H. Wagner. 2016. *ade4spatial: Multivariate Multiscale Spatial Analysis*.
- Dray, S., R. Pélissier, P. Couteron, M. J. Fortin, P. Legendre, P. R. Peres-Neto, E. Bellier, et al. 2012. “Community Ecology in the Age of Multivariate Multiscale Spatial Analysis.” *Ecological Monographs* 82: 257–75.
- Elith, J., and C. H. Graham. 2009. “Do They? How Do They? WHY Do They Differ? On Finding Reasons for Differing Performances of Species Distribution Models.” *Ecography* 32: 66–77.
- Ellis, N., S. J. Smith, and C. Roland Pitcher. 2012. “Gradient Forests: Calculating Importance Gradients on Physical Predictors.” *Ecology* 93: 156–68.
- Friedman, J. H. 2001. “Greedy Function Approximation: A Gradient Boosting Machine.” *Annals of Statistics* 29: 1189–232.
- Fussmann, G. F., G. Weithoff, and T. Yoshida. 2005. “A Direct, Experimental Test of Resource vs. Consumer Dependence.” *Ecology* 86: 2924–30.
- Gilbert, B., and J. R. Bennett. 2010. “Partitioning Variation in Ecological Communities: Do the Numbers Add Up?” *Journal of Applied Ecology* 47: 1071–82.
- Greenwell, B., B. Boehmke, J. Cunningham, and GBM Developers. 2020. *gbm: Generalized Boosted Regression Models*. R package version 2.1.8. <https://CRAN.R-project.org/package=gbm>.

- Griffith, D. A., and P. R. Peres-Neto. 2006. "Spatial Modeling in Ecology: The Flexibility of Eigenfunction Spatial Analyses." *Ecology* 87: 2603–13.
- Harris, D. J. 2015. "Generating Realistic Assemblages with a Joint Species Distribution Model." *Methods in Ecology and Evolution* 6: 465–73.
- Hastie, T., R. Tibshirani, and J. Friedman. 2009. *The Elements of Statistical Learning: Prediction, Inference and Data Mining*. New York: Springer-Verlag.
- Ishwaran, H., and U. Kogalur. 2019. Fast Unified Random Forests for Survival, Regression, and Classification (RF-SRC). R package version 2.9.0. <https://cran.r-project.org/package=randomForestSRC>.
- Krenek, S., T. U. Berendonk, and T. Petzoldt. 2011. "Thermal Performance Curves of *Paramecium caudatum*: A Model Selection Approach." *European Journal of Protistology* 47: 124–37.
- Legendre, P., and E. Gallagher. 2001. "Ecologically Meaningful Transformations for Ordination of Species Data." *Oecologia* 129: 271–80.
- Legendre, P., and L. Legendre. 2012. *Numerical Ecology*, 3rd ed. Amsterdam: Elsevier Science BV.
- Leibold, M. A., and J. M. Chase. 2017. *Metacommunity Ecology*. Princeton, NJ: Princeton University Press.
- Leibold, M. A., M. Holyoak, N. Mouquet, P. Amarasekare, J. M. Chase, M. F. Hoopes, R. D. Holt, et al. 2004. "The Metacommunity Concept: A Framework for Multi-Scale Community Ecology." *Ecology Letters* 7: 601–13.
- Miller, P. J., G. H. Lubke, D. B. McArtor, and C. S. Bergeman. 2016. "Finding Structure in Data Using Multivariate Tree Boosting." *Psychological Methods* 21: 583–602.
- Murakami, D., and D. A. Griffith. 2019. "Eigenvector Spatial Filtering for Large Data Sets: Fixed and Random Effects Approaches." *Geographical Analysis* 51: 23–49.
- Nieto-Lugilde, D., K. C. Maguire, J. L. Blois, J. W. Williams, and M. C. Fitzpatrick. 2018. "Multiresponse Algorithms for Community-Level Modelling: Review of Theory, Applications, and Comparison to Species Distribution Models." *Methods in Ecology and Evolution* 9: 834–48.
- Norberg, A., N. Abrego, F. G. Blanchet, F. R. Adler, B. J. Anderson, J. Anttila, M. B. Araújo, et al. 2019. "A Comprehensive Evaluation of Predictive Performance of 33 Species Distribution Models at Species and Community Levels." *Ecological Monographs* 89: e01370.
- Ovaskainen, O., G. Tikhonov, A. Norberg, F. Guillaume Blanchet, L. Duan, D. Dunson, T. Roslin, and N. Abrego. 2017. "How to Make More out of Community Data? A Conceptual Framework and Its Implementation as Models and Software." *Ecology Letters* 20: 561–76.
- Pebesma, E. J. 2004. "Multivariable Geostatistics in S: The gstat Package." *Computers and Geosciences* 30: 683–91.
- Peres-Neto, P. R., P. Legendre, S. Dray, and D. Borcard. 2006. "Variation Partitioning of Species Data Matrices: Estimation and Comparison of Fractions." *Ecology* 87: 2614–25.
- R Development Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>
- Rao, R. C. 1964. "The Use and Interpretation of Principal Component Analysis in Applied Research." *Sankhya – Series A* 26: 329–58.
- Ricklefs, R. E., and D. G. Jenkins. 2011. "Biogeography and Ecology: Towards the Integration of Two Disciplines." *Philosophical Transactions of the Royal Society B: Biological Sciences* 366: 2438–48.
- Rizopoulos, D. 2006. "Itm: An R Package for Latent Variable Modeling and Item Response Theory Analyses." *Journal of Statistical Software* 17: 1–25.
- Roberts, D. R., V. Bahn, S. Ciuti, M. S. Boyce, J. Elith, G. Guillera-Arroita, S. Hauenstein, et al. 2017. "Cross-Validation Strategies for Data with Temporal, Spatial, Hierarchical, or Phylogenetic Structure." *Ecography* 40: 913–29.
- Smith, T. W., and J. T. Lundholm. 2010. "Variation Partitioning as a Tool to Distinguish between Niche and Neutral Processes." *Ecography* 33: 648–55.
- Soininen, J. 2014. "A Quantitative Analysis of Species Sorting across Organisms and Ecosystems." *Ecology* 95: 3284–92.
- Thompson, P. L., L. M. Guzman, L. De Meester, Z. Horváth, R. Ptacnik, B. Vanschoenwinkel, D. S. Viana, and J. M. Chase. 2020. "A Process-Based Metacommunity Framework Linking Local and Regional Scale Community Ecology." *Ecology Letters* 23: 1314–29.
- Townsend Peterson, A., J. Soberón, R. G. Pearson, R. P. Anderson, E. Martínez-Meyer, M. Nakamura, and M. B. Araújo. 2011. *Niches and Geographic Distributions. Page Ecological Niches and Geographic Distributions (MPB-49)*. Princeton, NJ: Princeton University Press.
- Tuomisto, H. and Ruokolainen, K. 2006. "Analyzing or Explaining Beta Diversity? Understanding the Targets of Different Methods of Analysis." *Ecology* 87: 2697–708.
- Viana, D. 2022. duarte-viana/iVarPart: v1.0. Zenodo. Code. <https://doi.org/10.5281/zenodo.6240353>.
- Wood, S. N. 2017. *Generalized Additive Models: An Introduction with R*, 2nd ed. London: Chapman and Hall/CRC.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Viana, Duarte S., Petr Keil, and Alienor Jeliakov. 2022.

"Disentangling Spatial and Environmental Effects: Flexible Methods for Community Ecology and Macroecology." *Ecosphere* 13(4): e4028. <https://doi.org/10.1002/ecs2.4028>