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

Duarte S. Viana[1,2]*, Petr Keil[1,3,4], Alienor Jeliazkov[1,3,5]

[1]German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Puschstraße 4, 04103 Leipzig, Germany

[2]Leipzig University, Ritterstraße 26, 04109 Leipzig, Germany

[3]Institute of Computer Science, Martin Luther University Halle-Wittenberg, 06120 Halle (Saale), Germany

[4]Faculty of Environmental Sciences, Czech University of Life Sciences Prague, Kamýcká 129, Praha – Suchdol, 165 00, Czech Republic

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

*Corresponding author; e-mail: duarte.viana@idiv.de

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 binary occurrence and abundance. These methods included distance-based regression, generalized linear models (GLM, and the special case of RDA), generalized additive models (GAM), and tree-based machine learning (ML): regression trees, boosted regression trees (BRT), and random forests. The spatial component of the model consisted of spatial distance (in distance-based regression), Moran's Eigenvector Maps (MEM; in GLM and ML), smooth spatial splines (in GAM), or tree-based non-linear modelling 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 mostly caused by imbalanced performance in estimating environmental and spatial effects. Such differences also manifested when analyzing eight different empirical datasets. GLM 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 underfit simulated spatial structures, causing underestimation of spatial fractions 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 modelling, community ecology, and macroecology.

Keywords: metacommunity, variation partitioning, canonical analysis, multiple regression, random forest, Gaussian niche, spatial autocorrelation

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 and Leibold 2003, Townsend Peterson et al. 2011). Consequently, the environment is central in ecological theory, including coexistence theory (Chesson and Warner 1981, Chesson 2000), modern niche theory (Chase and Leibold 2003), metacommunity theory (Leibold et al. 2004, Thompson et al. 2020), as well as biogeographical and macroecological theory (Ricklefs and Jenkins 2011, Townsend Peterson et al. 2011). Species-environment relationships (SERs) have been widely estimated to (i) characterize the species' niches and model species distributions and community composition (Townsend Peterson et al. 2011, Bar-Massada 2015, Norberg et al. 2019), (ii) explore the importance of environmental filtering as a biodiversity process (Cottenie 2005, Soininen 2014), and (iii) correct for environmental effects when studying biotic interactions and other community assembly processes (e.g. Ovaskainen et al. 2017, D'Amen et al. 2018).

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, Peres-Neto et al. 2006, Soininen 2014, Leibold and Chase 2017).

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 and Graham 2009, Norberg et al. 2019). In community ecology multivariate methods, in particular constrained ordination (such as CCA and RDA) or distance-based methods, have been widely used to partition explained variation in community composition according to environmental and spatial effects and infer community assembly processes (Cottenie 2005, Peres-Neto et al. 2006, Tuomisto and Ruokolainen 2006, Soininen 2014). 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 and Bennett 2010, Smith and Lundholm 2010).

On the other hand, methods used in biogeography and macroecology such as GLM, GAM, and ML, have been widely overlooked for disentangling spatial dependence in the context of variation partitioning, particularly in a community context with multi-species responses. Machine learning methods such as Random Forest and Boosted Regression Trees (Elith and 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 non-

linear relationships. Although comparisons of different models to estimate species distributions exist (Norberg et al. 2019), the usefulness of methods such as GAM and machine learning to model different spatial structures while estimating environmental effects, and to partition variation, has been less explored. Here, we compare the virtues and drawbacks of methods based on generalized linear models, generalized additive models, and machine learning (Table 1) to model and disentangle environmental and spatial effects. We do this comparison by using both simulated and empirical data. We specifically focus on site-by-species matrices of occurrence or abundance as response variables, but this study also applies to any other kind of response variable affected by spatial autocorrelation (e.g., species richness and beta-diversity).

Methods

All the analyses were carried out in R (R Development Core Team 2020), and all the computer code is available at https://github.com/duarte-viana/iVarPart.

Data simulation

We simulated site-by-species community data where variation in abundance or binary occurrence corresponded to pre-defined 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 (here fixed at A=N/2), 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

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

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, s.d.=2). Alternatively,

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

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 S1.1.

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 S1.2.

Multi-species 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

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

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 predicted 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 S1.3). 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:

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

$$r^2_X = r(Y_i, X_i)^2, \qquad (5)$$

$$[X] = r^2_X, \qquad (6)$$

$$[W] = r^2_{XW} - r^2_X. \qquad (7)$$

Although W is independent from and thus uncorrelated with X, some spurious collinearity between them can arise and jointly explain variation in Y. Because W is a pure spatial variable, the collinear effect is the effect of spatially correlated X. Note that 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 $\bar{Y}_i$, i.e., $\beta_X \propto r_X^2$ and $\beta_W \propto r_W^2$ (Appendix S1: Figure S1.4). 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.

We considered different scenarios in the type of data (normal, counts, or binary), the size of the grid (N = 25, 100, or 400 sites), the type of spatial variogram model (exponential or Gaussian), the autocorrelation range of W (A = 0.01N, 0.5N, 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, 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 also considered a scenario where a random subsample (50 out of 400 sites) was taken to simulate a sampling effect, and another scenario where either 3 or 6 random environmental variables orthogonal to the response (i.e. noise) were added to inspect overfitting propensity. For each combination of simulation parameters, 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 β_W =1-β_X. Each combination of parameters and β_X was replicated 5 times, resulting in a total of 3465 simulation runs.

Statistical methods

We considered methods in four broad families of statistical models (Table 1): distance-based regression, generalized linear models (GLM), generalized additive models (GAM), and tree-based machine learning (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. Because considering all the possible approaches with each method would result in an intractable list, we opted for the most widely used ones, as described below. Our goal was to assess the performance of these different approaches for fitting environmental and spatial effects (Exercise 1), and partitioning explained variation (Exercise 2).

1. *Distance-based regression*

Applying regression on compositional distances between sites derived from site-by-species abundance or occurrence (i.e. pairwise beta-diversity) (Tuomisto and Ruokolainen 2006) has often been used to investigate the relative importance of environment and space in determining community assembly patterns (see Tuomisto and Ruokolainen, 2008, and the discussion initiated with Legendre et al. 2005; Weinstein et al., 2014; Robroek et al., 2017). The most common distance-based regression method is the Multiple Regression on distance Matrices (MRM), which is the adaptation of the Mantel regression to multiple regression analysis (Lichstein 2007). It consists of regressing ecological distances (i.e. compositional dissimilarities between sites; the response matrix) against the matrices of environmental and geographical distances (the predictor matrices). We used Bray-Curtis dissimilarities for the species data (R package *vegan*; Oksanen et al. 2017), and Euclidean distances for both the environment and the spatial coordinates. A simple linear regression was performed on the pairwise distances using function 'lm' of the R *stats* package (R Development Core Team 2020).

2. *Generalized linear models*

We fitted GLMs using the *stats* R package (R Development Core Team 2020). We fitted either a linear effect or a second-degree polynomial to model the environmental effect (depending on the simulated environmental component *X*). To model the spatial effects, we used distance-based Moran's Eigenvector Maps (MEM), calculated using the function "dbmem" (with default settings) of the R package *adespatial* (Dray et al. 2016). MEM variables represent the spatial autocorrelation across sampled sites at different spatial scales (i.e., different grains of autocorrelation). Also more broadly known as Principal Coordinates of Neighbourhood Matrices (PCNM), MEMs are widely used to model spatial structure and autocorrelation. 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 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 also included the linear models underlying RDA (van den Wollenberg 1977) in our comparison. The difference to GLM models is that the species data are usually first transformed and subsequently used to estimate a linear model for each species – note, therefore, that the error distribution is always assumed to be normal, independently of the type of data. Accordingly, prior to GLM, the response data were Hellinger-transformed (as recommended in Legendre and Gallagher 2001, Legendre and Legendre 2012) and, when the response to the environment was bell-shaped, log-transformed to improve model fitting.

3. *Generalized additive models*

GAMs 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) or fixed k=10 on the spatial coordinates of the sites to model the spatial effect. The different k allowed us to compare different initial degrees of "wiggliness" of the basis functions. The smoothing parameters were estimated by restricted maximum likelihood (REML). 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.

4. *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 machine learning 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 the method by defining a different value for the algorithm parameter that mostly affected the smoothness of the fitted model, according to *a priori* trial-error explorations.

Multivariate Regression Trees (MVRT). This method fits a single multivariate regression tree (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 or without internal cross-validation. The model with MEMs was estimated without cross-validation.

Univariate Random Forest (UniRF). This fits a univariate random forest (Breiman 2001, Hastie et al. 2009) to each species individually, using 'randomForest' R function (package *randomForest*). A random forest consists of a set of regression trees fitted to bootstrapped (i.e. sampled with replacement) data, 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 used the default settings of the 'randomForest' function, as we expect the defaults to be most often adopted by users – specifically, the random forest consists of 500 regression trees, each with a sample size of either N (the sample size of the data) or $N/5$ (i.e. 20% of the data), 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 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 (CNMSE), 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 et al. 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 sqrt(p), rounded up, where p is the total number of predictors. Such as for UniRF, we varied the sample size to be either N (the sample size of the data) or $N/5$ (i.e. 20% of the data) in order to assess different "smoothness" settings. Here again, the model with MEMs used a sample size of N.

Boosted Regression Trees (BRT). 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*) and fitted two variants of the algorithm – shrinkage (or "learning rate") set to either 0.01 or 0.1. We set the tree depth to two 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, where we used a tree depth of 1 to avoid modelling unnecessary interactions between MEMs. The model with MEMs used a sample learning rate of 0.01.

Assessment of method performance on simulated data

Our study comprised two exercises. In Exercise 1 we assessed the flexibility of the methods in Table 1 to approximate the simulated linear or bell-shaped responses of species to the environment, as well as their responses to spatial structures which are typically more complex and non-linear. In Exercise 2 we partitioned the variation in multi-species abundance or occurrence caused by the spatial and environmental effects, and compared these to the simulated 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 values predicted by the method $\hat{Y}$. However, deviations from the true model might be due to under- or overfitting: if data are overfitted, the predictions will be closer to the observed data Y (as the added error will be also fitted to some extent) and the correlation between Y and $\hat{Y}$ will be higher than the correlation between $\bar{Y}$ and $\hat{Y}$; if data are underfitted, the correlation between Y and $\hat{Y}$ will be lower than that between $\bar{Y}$ and $\hat{Y}$. 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 $\beta_W = 0$), and spatial effects alone (i.e., when $\beta_X = 0$ and $\beta_W = 1$). RDA and MRM were excluded from this exercise, because the transformations of the response data make these models incomparable with the others.

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 [$\hat{W}$]. 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 simulated scenarios was assessed by comparing the estimated vs. 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 GLM methods (including RDA) we adjusted the r^2 to account for the number of predictors (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 different methods (Table 1) by performing variation partitioning 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 3 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 (Fig. 1). Most methods tended to underfit the data when both environmental and spatial effects were included, except GLM, BRT, and the MVRF variant with lower smoothness (MVRF-SS0 and MVRF-

MEM) (Fig. 1a-c). The environmental effects were generally better fitted than the spatial effects, as the predictions match the true (simulated) response more closely (Fig. 1d-i). In particular, parametric or semi-parametric methods (GLM, GAM) showed the best fitting performance for the environmental effects (Fig.1d-f), except for data with normal error distribution because these models failed to accurately fit the bell-shaped response, even if the response variable was log-transformed (Appendix S1, Fig. S1.5b). Methods using MEMs tended to misfit the spatial component by either overfitting (GLM, BRT, MVRF) or underfitting it (UniRF, MVRT), but in general to a lesser extent than the other methods. The spatial splines in GAM and non-linear modelling of spatial coordinates in ML tended to underfit the spatial component, except BRT with the higher learning rate (Fig.1g-i).

The type of environmental response and spatial structure also caused fitting differences (Fig. 2). The parametric and semi-parametric methods (GLM and GAM) showed decreasing fitting performance as the shape of the bell-shaped environmental response narrowed, whereas BRT (and the other ML methods) could better fit these types of response, but could also easily overfit if not sufficiently smoothened. For the spatial effects, the fitting performance increases with the smoothness of the spatial pattern (Fig. 2). For a nearly random spatial structure (A=0.01N), for which we expected that the models did not fit the data (i.e. we expected simulated-predicted correlations to be close to 0, and thus a score closer to -1 in Fig. 2d, e, f), we observed that MEMs overfitted the data, BRT overfitted the data to some extent (unless learning rate was reduced), whereas GAM correctly did not overfit into this random variation (Fig. 2, lower panels).

Exercise 2. Variation partitioning

The variation partitioning performance, i.e. the difference between the estimated ([$\hat{X}$], [$\hat{W}$]) and simulated ([X], [W]) fractions, varied considerably across methods, but was consistent across the different types of data (Fig. 3, 4). None of the methods perfectly recovered both the spatial and environmental fractions from the simulated reference variation partitioning. GLMs and BRT with MEMs (BRT-MEM) were the most consistent with the simulated variation partitioning (Fig. 3, 4). In contrast, the distance-based regression (MRM) severely underestimated both fractions of variation (Fig. 3). The results were qualitatively similar when using R^2 or deviance-based pseudo-R^2 metrics for partitioning the variation (Appendix S1; Fig. S1.7, S1.8).

For GLMs, the tendency to generally overfit the spatial component led to the overestimation of the spatial fraction of variation [$\hat{W}$] (Fig. 3). The environmental fraction was generally correctly estimated, but was underestimated for the narrowest bell-shaped responses – variation that was instead fitted by the spatial component (Fig. 4). RDA generally largely underestimated the environmental and spatial fractions of variation – note, however, that the spatial fraction was systematically higher when using the classical R^2 metric, although its trend did not generally follow that of the simulated data (Appendix S1; Fig. S1.7 and S1.8).

GAMs underestimated the spatial fraction of variation but generally correctly estimated the environmental fraction. Still, and even though we used splines to fit the environmental effect, the narrowest responses to the environment were also underestimated. Because we let the smoothing parameter to be optimized, we did not find differences between the number of dimensions of the basis functions (default or k=10).

The performance of machine learning differed to a large extent 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 model smoothness. All ML methods tended to

overestimate the environmental fraction of variation, except UniRF and MVRT with cross-validation (MVRT-CV) (Fig. 3). On the contrary, the spatial fraction was generally underestimated, though only slightly for BRT with MEMs (Fig. 3, 4). The less smoothened versions of BRT (BRT-lr0.1) and MVRF (MVRF-SS0) largely overestimated the environmental fraction, but had only minor effects on the estimation of the spatial fraction (Fig. 3). BRT with MEMs was the method that performed consistently better across all simulated scenarios (Fig. 4). However, BRT was sensitive to sample size and the number of environmental covariates, noticeably overestimating the environmental fraction when we added environmental covariates unrelated to the response.

Empirical data

The choice of method had a clear influence on the results of the variation partitioning on empirical data (Fig. 5). Fitting GLM and MVRT with cross-validation (MVRT-CV) led to convergence problems for some data sets, likely because of the large number of spatial predictors relative to sample size in GLMs, and the too few observations for cross-validation in MVRT. Consistently with the simulation exercises, the environmental and spatial fractions in RDA were systematically underestimated relative to GAM and BRT. The results for GAM were more similar to BRT, in particular for the environmental fraction of variation. The results for BRT with spatial coordinates (BRT-lr0.01) and MEMs (BRT-MEM) as spatial predictors had also the same trend, but the spatial fraction was systematically higher in BRT-MEM.

Discussion

Our simulations show that it is challenging to recommend a single universal method for modelling 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 effects. GLMs and BRT with MEMs can provide more balanced fits that more consistently disentangle the spatial and environmental sources of variation in 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. GLMs are particularly useful when specific hypotheses about species-environment relationships are considered, as we can limit the scope of modelled responses to, for instance, linear and unimodal. In contrast, BRT allows us to fit any kind of response. The other methods can also be useful under certain situations.

GAMs are a good compromise, as modelled species-environment relationships 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 fitted using flexible splines on spatial coordinates. Because GAMs tend to fit smoother spatial surfaces in comparison to other methods, they might be more transferable, for example if the goal is to predict the response in unsampled sites. Note, however, that these models underestimate spatial effects when spatial autocorrelation acts over short distances. To allow GAMs to be flexible enough as to model wobblier effects, we can for example perform model selection according to *k* and smoothness of the splines (Wood 2007).

Tree-based ML methods have the advantage of being flexible and user-friendly, and are especially useful when we lack specific hypotheses about the functional forms of species-

environment relationships, even though ML output can be interpreted and tested for effect significance (Lucas 2020). Furthermore, tree-based ML inherently models interactions between predictors, a desirable property when modelling complex ecological systems. However, it was challenging to parameterize ML algorithms that worked well for jointly fitting the environmental and spatial components. Simply using spatial coordinates to model space turned out to underfit most spatial structures, but forcing the model 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, for reasons that remain elusive. Thus, although ML showed potential to be used for variation partitioning purposes, we still need to learn more about how to fine-tune the ML parameters affecting spatial structure.

This brings us to the challenge of modelling spatial structures. Among the spatial structures that we considered, those generated with exponential variograms (Appendix S1: Figure S1.2) tended to be underfitted, whereas nearly random spatial structures tended to be overfitted by the methods that better fitted the former (Fig. 2). Therefore, it is critical to choose the type of spatial structure depending on the processes involved (e.g. dispersal limitation or connectivity) and goals of the study. MEMs are useful for interpretation purposes, for example to infer spatial scales of variation (Dray et al. 2012, Murakami and Griffith 2019). GAMs, on the other hand, can fit smooth spatial surfaces (e.g. Appendix S1: Figure S1.6), such as those simulated via a Gaussian variogram model (Appendix S1: Figure S1.2). And ML can offer flexible solutions, but *a priori* explorations, e.g., through simulations, might be needed in order to know how flexible the spatial structure needs to be. We also warn that increasing numbers of spatially autocorrelated environmental predictors generate spurious correlations (Chapman

2010), something that shall be accounted for when performing variation partitioning (Clappe et al. 2018 and the method therein). Cross-validation is a good solution, but it should be performed carefully in a spatial setting, because fitted spatial surfaces are hardly transferable in space, and training and validation data may not be spatially independent (Roberts et al. 2017). 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), and if sample size is sufficient, then spatially blocked cross-validation is a good option (Roberts et al. 2017).

Our results show that there is margin for improvement. ML 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 modelling, 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 non-linear responses while using JSDM machinery (Harris 2015). Other possibilities include the use of autoregressive models (Dormann et al. 2007), though we are unware of their use for variation partitioning, 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 environment are hard to hypothesize in the context of community or multi-species 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 species-environment relationships are considered, GLM as a parametric method is a reasonable choice for variation partitioning. By choosing appropriate methods to model different responses to the environment, and different spatial structures at different spatial scales, 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.

Acknowledgements

We thank Jonathan Chase and Pedro Peres-Neto 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 REES (Research Excellence in Environmental Sciences) grant from Faculty of Environmental Sciences, Czech University of Life Sciences in Prague.

Literature cited

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 19:831–841.

Chase, J. M., and M. A. Leibold. 2003. Ecological niches: linking classical and contemporary approaches. University of Chicago Press.

Chesson, P. 2000. Mechanisms of maintenance of species diversity. Annual Review of Ecology and Systematics 31:343–366.

Chesson, P. L., and R. R. Warner. 1981. Environmental Variability Promotes Coexistence in Lottery Competitive Systems. The American Naturalist 117:923–943.

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 99:1737–1747.

Cottenie, K. 2005. Integrating environmental and spatial processes in ecological community dynamics. Ecology Letters 8:1175–1182.

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 41:1233–1244.

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 92:169–187.

De'ath, G. 2002. Multivariate regression trees: A new technique for modeling species-environment relationships. Ecology 83:1105–1117.

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

Dray, A. S., G. Blanchet, D. Borcard, G. Guenard, T. Jombart, G. Larocque, P. Legendre, N. Madi, and H. H. Wagner. 2016. adespatial: Multivariate Multiscale Spatial Analysis.

Dray, S., R. Pélissier, P. Couteron, M. J. Fortin, P. Legendre, P. R. Peres-Neto, E. Bellier, R. Bivand, F. G. Blanchet, M. De Cáceres, A. B. Dufour, E. Heegaard, T. Jombart, F. Munoz, J. Oksanen, J. Thioulouse, and H. H. Wagner. 2012. Community ecology in the age of multivariate multiscale spatial analysis. Ecological Monographs 82:257–275.

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 93:156–168.

Friedman, J. H. 2001. Greedy function approximation: A gradient boosting machine. Annals of Statistics 29:1189–1232.

Gilbert, B., and J. R. Bennett. 2010. Partitioning variation in ecological communities: do the numbers add up? Journal of Applied Ecology 47:1071–1082.

Harris, D. J. 2015. Generating realistic assemblages with a joint species distribution model. Methods in Ecology and Evolution 6:465–473.

Hastie, T., R. Tibshirani, and J. Friedman. 2009. The elements of statistical learning: prediction, inference and data mining. Springer-Verlag, New York.

Ishwaran, H., U. B. Kogalur, and M. U. B. Kogalur. 2019. Package ‘randomForestSRC.’

Laliberté, E. 2008. Analyzing or Explaining Beta Diversity? Comment. Ecology 89:3232–3237.

Legendre, P., D. Borcard, and P. R. Peres-Neto. 2005. Analyzing beta diversity: Partitioning the spatial variation of community composition data. Ecological Monographs 75:435–450.

Legendre, P., and E. Gallagher. 2001. Ecologically meaningful transformations for ordination of

species data. Oecologia 129:271–280.

Legendre, P., and L. Legendre. 2012. Numerical ecology. 3rd Englis. Elsevier Science BV, Amsterdam.

Leibold, M. A., and J. M. Chase. 2017. Metacommunity Ecology. Princeton University Press.

Leibold, M. A., M. Holyoak, N. Mouquet, P. Amarasekare, J. M. Chase, M. F. Hoopes, R. D. Holt, J. B. Shurin, R. Law, D. Tilman, M. Loreau, and A. Gonzalez. 2004. The metacommunity concept: a framework for multi-scale community ecology. Ecology Letters 7:601–613.

Lichstein, J. W. 2007. Multiple regression on distance matrices: A multivariate spatial analysis tool. Plant Ecology 188:117–131.

Lucas, T. C. D. 2020. A translucent box: interpretable machine learning in ecology. Ecological Monographs n/a:ecm.1422.

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 9:834–848.

Norberg, A., N. Abrego, F. G. Blanchet, F. R. Adler, B. J. Anderson, J. Anttila, M. B. Araújo, T. Dallas, D. Dunson, J. Elith, S. D. Foster, R. Fox, J. Franklin, W. Godsoe, A. Guisan, B. O'Hara, N. A. Hill, R. D. Holt, F. K. C. Hui, M. Husby, J. A. Kålås, A. Lehikoinen, M.

Luoto, H. K. Mod, G. Newell, I. Renner, T. Roslin, J. Soininen, W. Thuiller, J. Vanhatalo, D. Warton, M. White, N. E. Zimmermann, D. Gravel, and O. Ovaskainen. 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 20:561–576.

Pebesma, E. J. 2004. Multivariable geostatistics in S: The gstat package. Computers and Geosciences 30:683–691.

Peres-Neto, P. R., P. Legendre, S. Dray, and D. Borcard. 2006. Variation partitioning of species data matrices: Estimation and comparison of fractions. Ecology 87:2614–2625.

R Development Core Team. 2017. R: A language and environment for statistical computing.

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 366:2438–2448.

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, J. J. Lahoz-Monfort, B. Schröder, W. Thuiller, D. I. Warton, B. A. Wintle, F. Hartig, and C. F. Dormann. 2017. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography 40:913–929.

Robroek, B. J. M., V. E. J. Jassey, R. J. Payne, M. Martí, L. Bragazza, A. Bleeker, A. Buttler, S. J. M. Caporn, N. B. Dise, J. Kattge, K. Zajac, B. H. Svensson, J. Van Ruijven, and J. T. A.

Verhoeven. 2017. Taxonomic and functional turnover are decoupled in European peat bogs. Nature Communications 8:1161.

Smith, T. W., and J. T. Lundholm. 2010. Variation partitioning as a tool to distinguish between niche and neutral processes. Ecography 33:648–655.

Soininen, J. 2014. A quantitative analysis of species sorting across organisms and ecosystems. Ecology 95:3284–3292.

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 23:1314–1329.

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 University Press.

Tuomisto, H., and K. Ruokolainen. 2006. Analyzing or explaining beta diversity? Understanding the targets of different methods of analysis. Ecology 87:2697–2708.

Tuomisto, H., and K. Ruokolainen. 2008. Analyzing or Explaining Beta Diversity? Reply. Ecology 89:3244–3256.

van den Wollenberg, A. L. 1977. Redundancy analysis an alternative for canonical correlation analysis. Psychometrika 42:207–219.

Weinstein, B. G., B. Tinoco, J. L. Parra, L. M. Brown, J. A. McGuire, F. Gary Stiles, and C. H. Graham. 2014. Taxonomic, phylogenetic, and trait beta diversity in south American hummingbirds. American Naturalist 184:211–224.

Wood, S. N. 2017. Generalized additive models: An introduction with R, second edition. Chapman and Hall/CRC.

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 modelling	Specific settings in R
Distance-based regression	MRM	Multiple regression on distance matrices	NP[§]	Any/Linear[¶]	XY Euclidean distances	- R package: *base* - Dissimilarity index: Bray-Curtis (from R package *vegan*)
Generalized linear models	RDA	Redundancy analysis (based on linear regression)	P	Linear, polynomial	MEM	- R package: *base* (similar to the *vegan* approach) - Species data: Hellinger-transformed - Distribution: normal
	GLM	Generalized linear model	P	Linear, polynomial	MEM	- R package: *base* - 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	- R package: *mgcv* - Species data: raw - Distribution: normal (abundance), Poisson (counts), binomial (binary)
Machine learning tree-based methods	BRT	Multivariate boosted regression trees	NP	Any	XY coordinates, MEM	- R package: *mvtboost* - Species data: raw - Distribution: normal (abundance), Poisson (counts), Bernoulli (binary) - # trees: 1000 - Learning rate: 0.01 or 0.001 - Interaction depth: 2 (1 with MEMs) - Internal CV: no
	MVRT	Multivariate regression tree	NP	Any	XY coordinates, MEM	- R package: *mvpart* - Species data: raw (abundance), numeric 0/1 (binary) - Node size: 5 (no CV) - Interaction depth: 1-20 (CV) - Internal CV: yes or no
	UniRF	Univariate random forest	NP	Any	XY coordinates, MEM	- R package: *randomForest* - Species data: raw (abundance), categorical 0/1 (binary)

					- # trees: 500 - Node size: default - Sample size: N or N/5
MVRF	Multivariate random forest	NP	Any	XY coordinates, MEM	- R package: *randomForestSRC* - Species data: raw (abundance), categorical 0/1 (binary) - # trees: 500 - Node size: default - Sample size: N or N/5

*P=parametric, NP=non-parametric, SP=semi-parametric
§Depending on the method used to model the link between biological and environmental distances
¶Depending on what is considered as the environmental response, either the raw data (Any) or the environmental distances (Linear)
CV: cross-validation

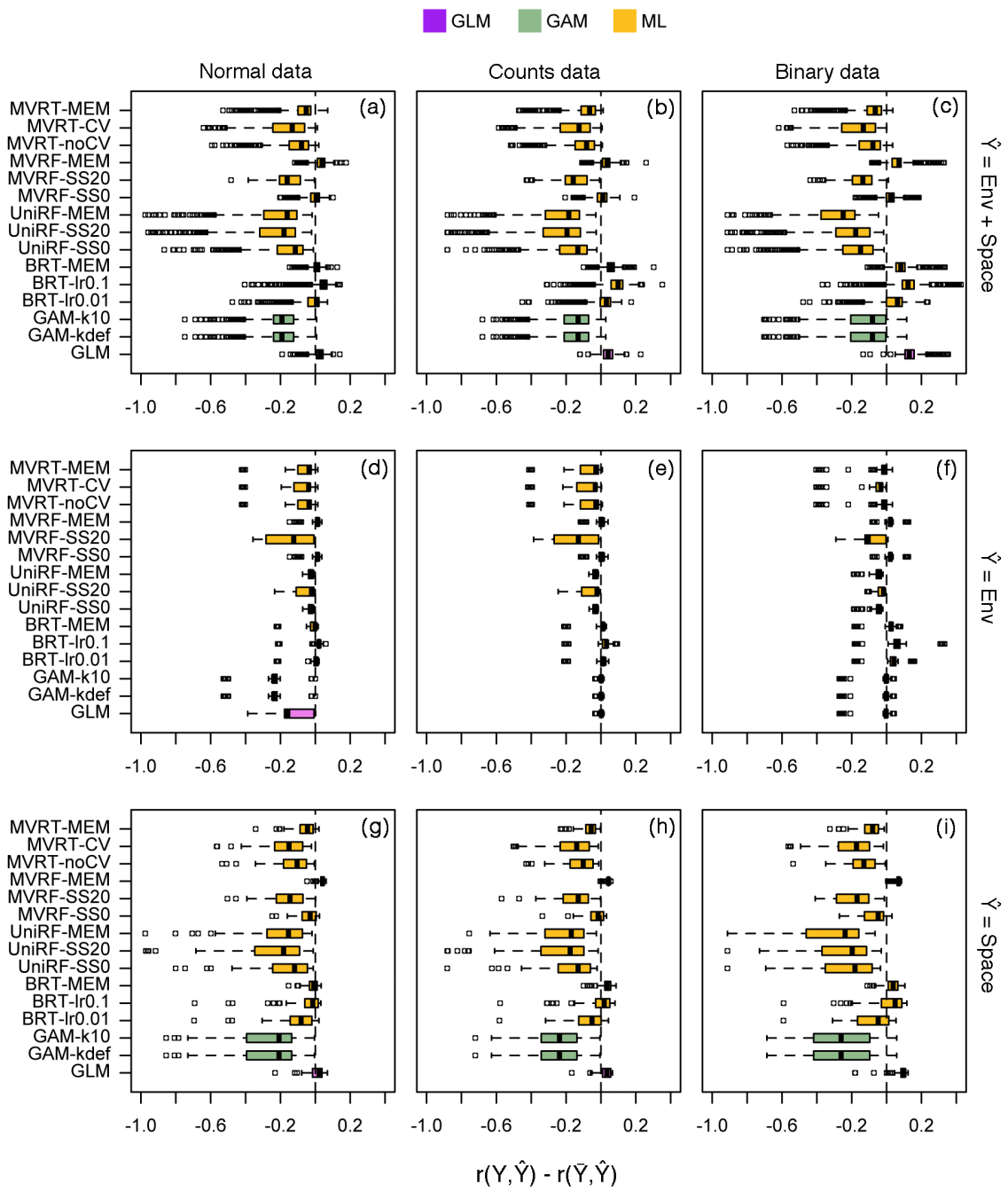


Figure 1. Fitting performance of the different statistical models for the additive effect of environment and space (a-c), the environmental effect alone (d-f), and the spatial effect alone (g-i), fitted to normal, counts, and binary data, across the different generated data (see Appendix S1: Table S1). 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 ($\bar{Y}$) and the model predictions ($\hat{Y}$). The dashed line represents a perfect fit (i.e., $\hat{Y}=\bar{Y}$), positive values indicate overfitting, and negative values indicate underfitting. MEM: Moran's Eigenvector Maps, k: dimension of the spline basis function (default, "def", or with fixed k=10, "k10"), lr: learning rate (0.01 or 0.1), SS: resample size (0 for SS=N and 20 for SS=N/5), CV: internal cross-validation (with, "CV", or without, "noCV").

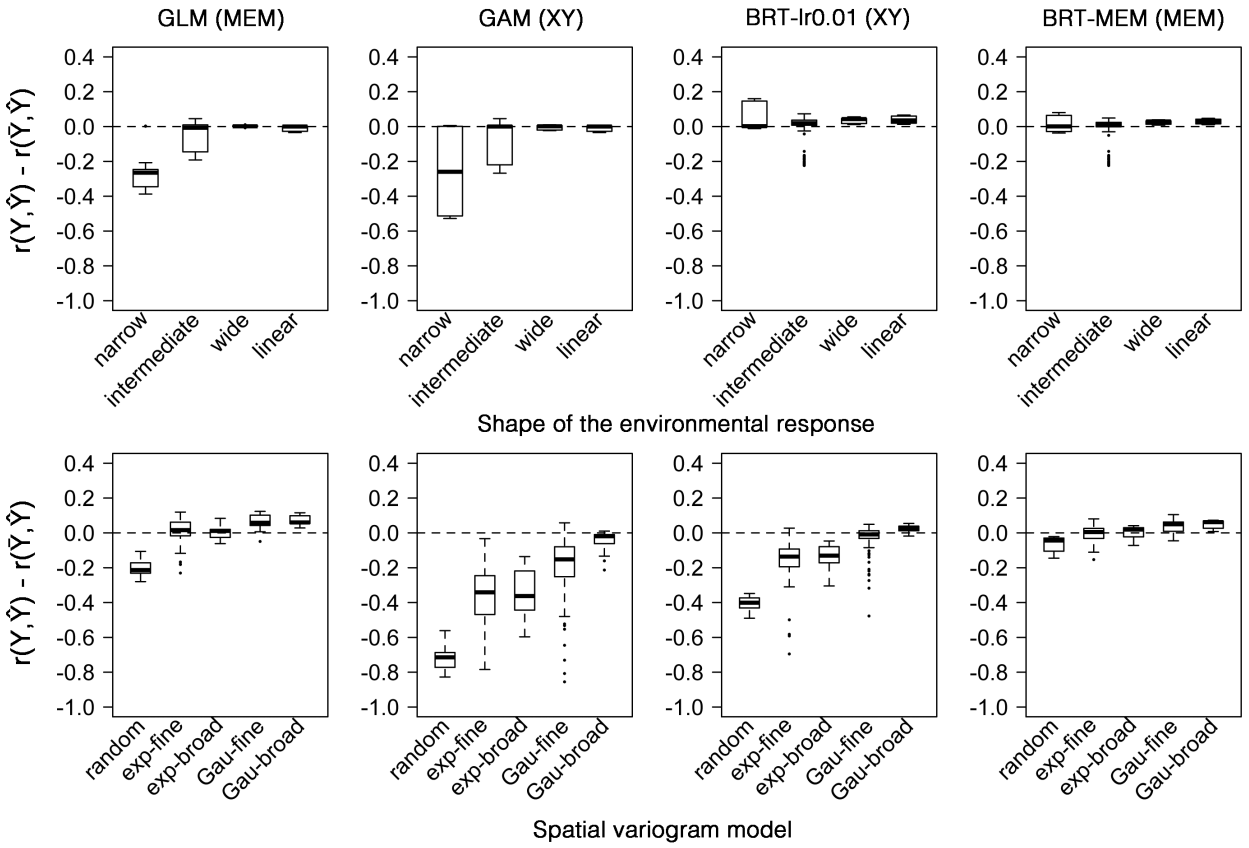


Figure 2. Fitting performance of different statistical models (GLM, GAM-k5, BRT-lr0.01) for the different environmental effects (a, b, c; unimodal with different niche breadths and linear; see also Appendix S1: Figure S1.1) and the different spatial effects (d, e, f; exponential and Gaussian variogram models with A=0.01, i.e., 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 S1.2). The results are pooled for all types of response data (i.e., normal, counts, and binary data). 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 ($\bar{Y}$) and the model predictions ($\hat{Y}$). The dashed line represents a perfect fit (i.e., $\hat{Y}=\bar{Y}$), positive values indicate overfitting, and negative values indicate underfitting.

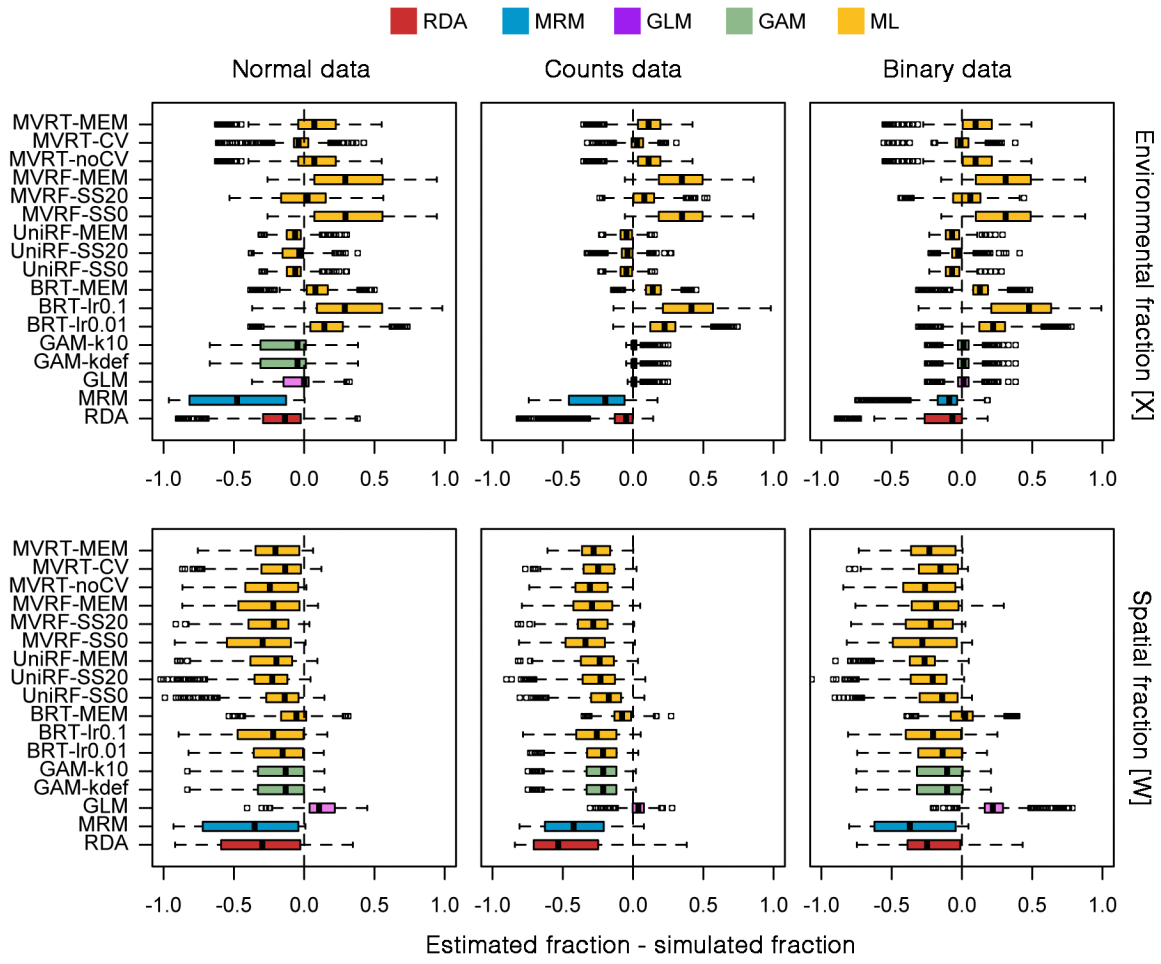


Figure 3. Variation partitioning performance of the different methods for normal, counts, and binary data across the different generated data. The performance is given by the difference between the estimated and simulated fractions of variation explained by the environment ([$\hat{X}$] - [X]; upper panels) and space ([$\hat{W}$] - [W]; lower panels). The dashed line represents equality (simulated = estimated), positive values indicate overestimation, and negative values indicate underestimation.

Figure 4

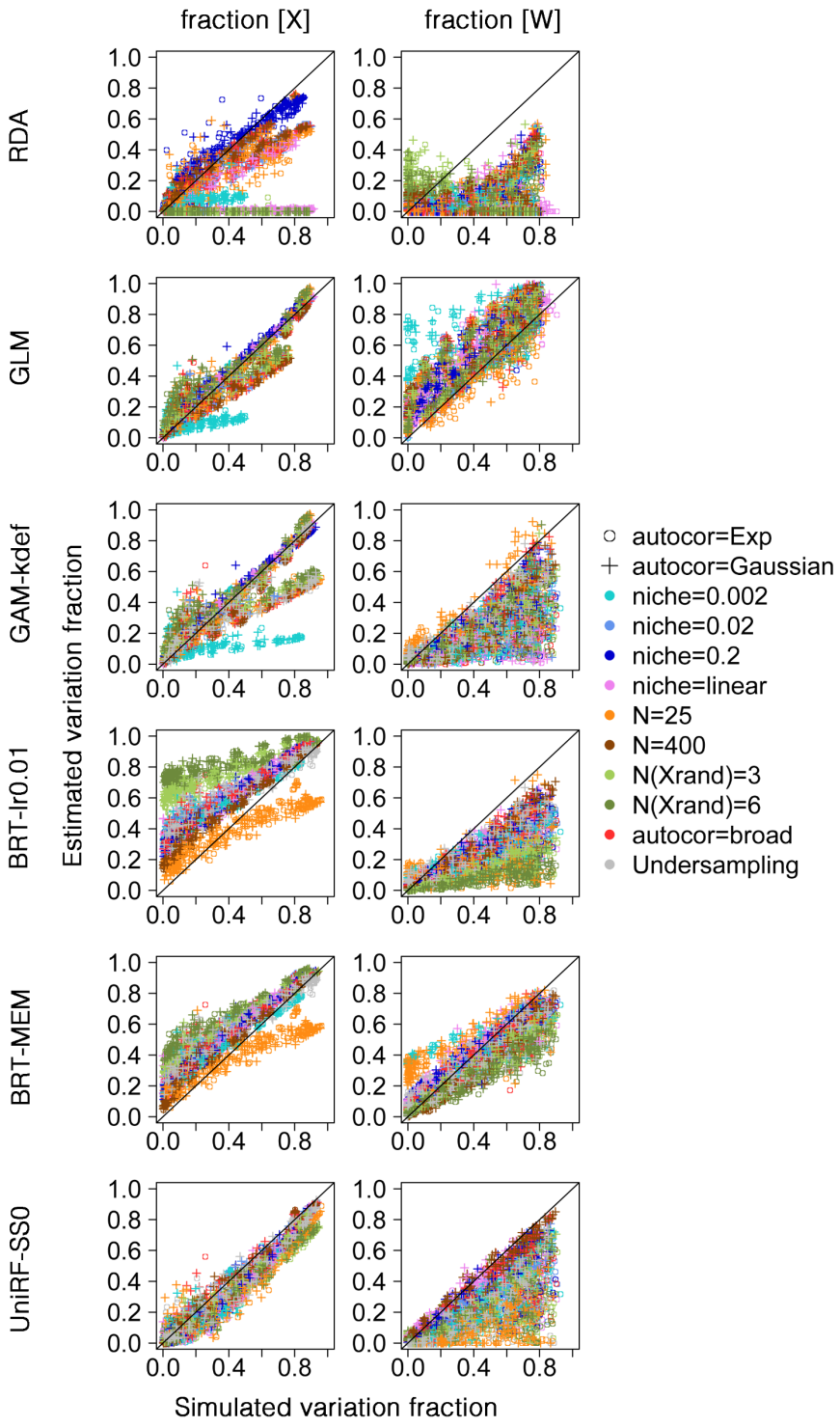


Figure 4. Plots of simulated vs. estimated fractions of variation explained by the environment ([X] vs. [$\hat{X}$]; left panels) and space ([W] vs. [$\hat{W}$]; right panels) across different methods (different rows) and generated data (different colours). The black line represents the 1:1 line, where points should fall if the estimated fraction is equal to the simulated fraction.

Figure 5

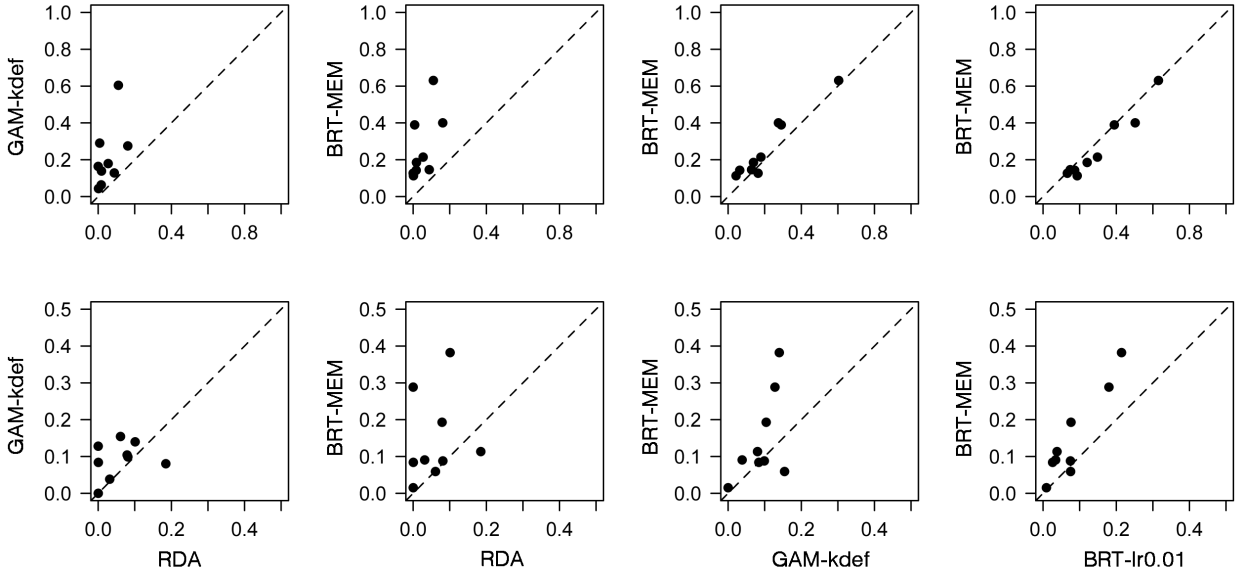
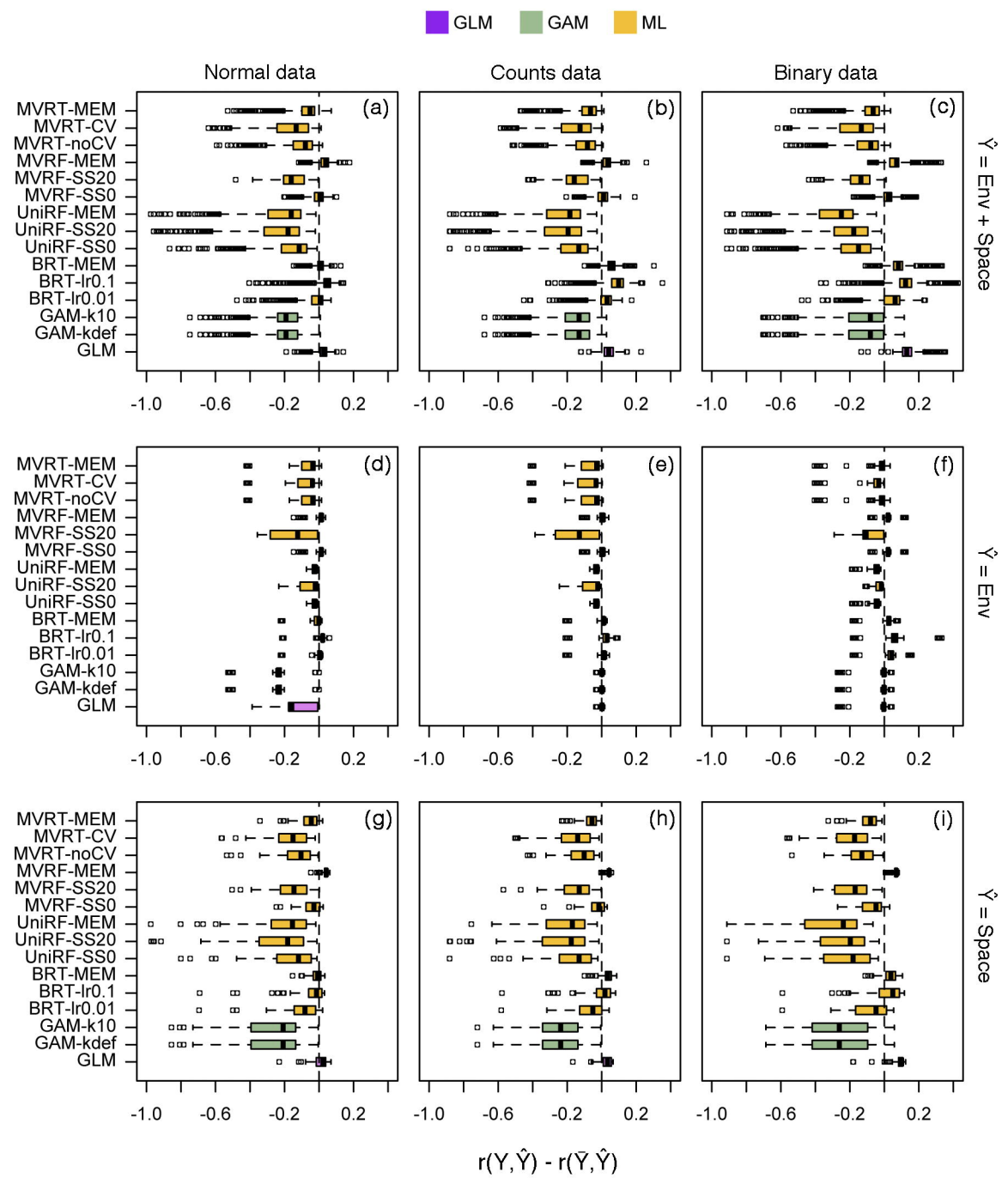
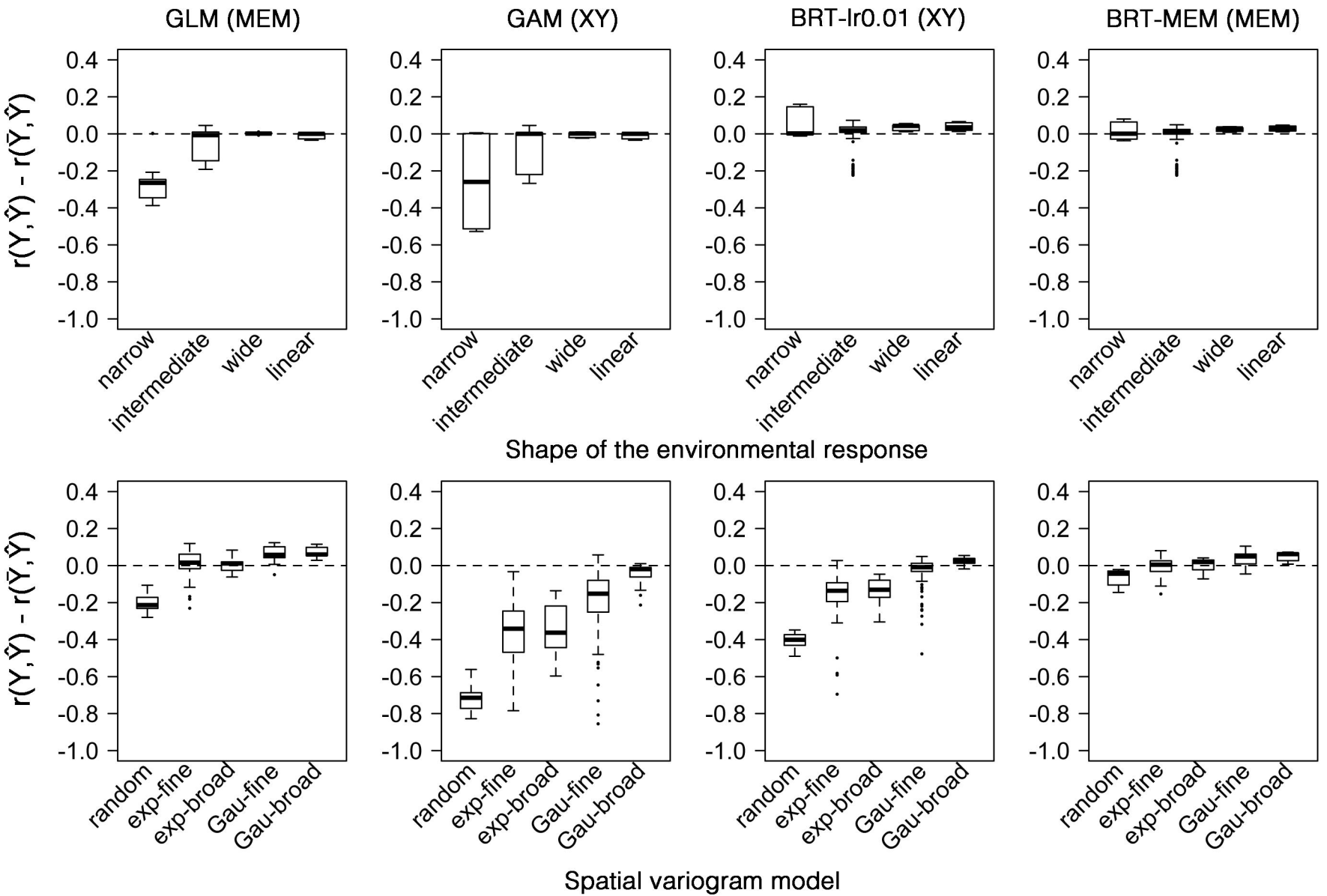
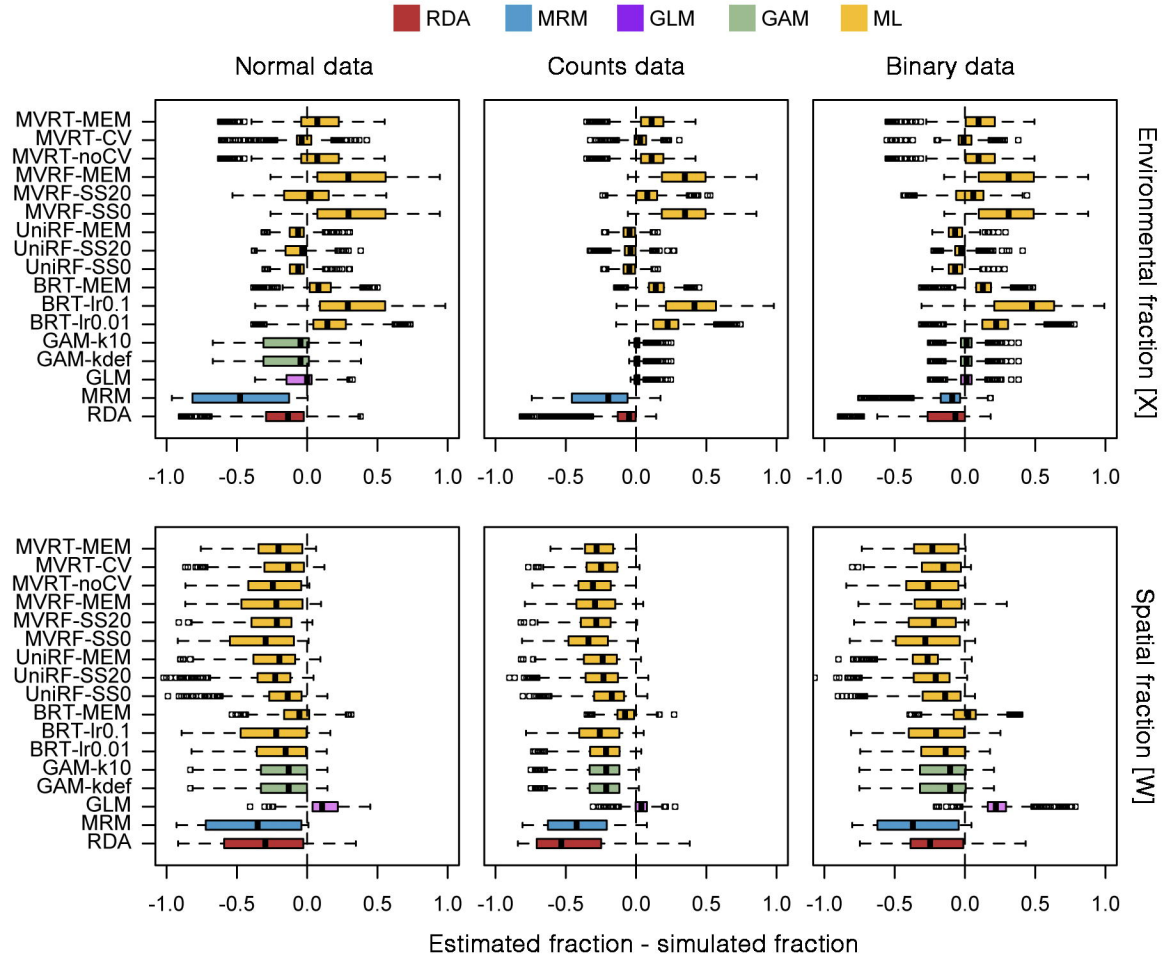
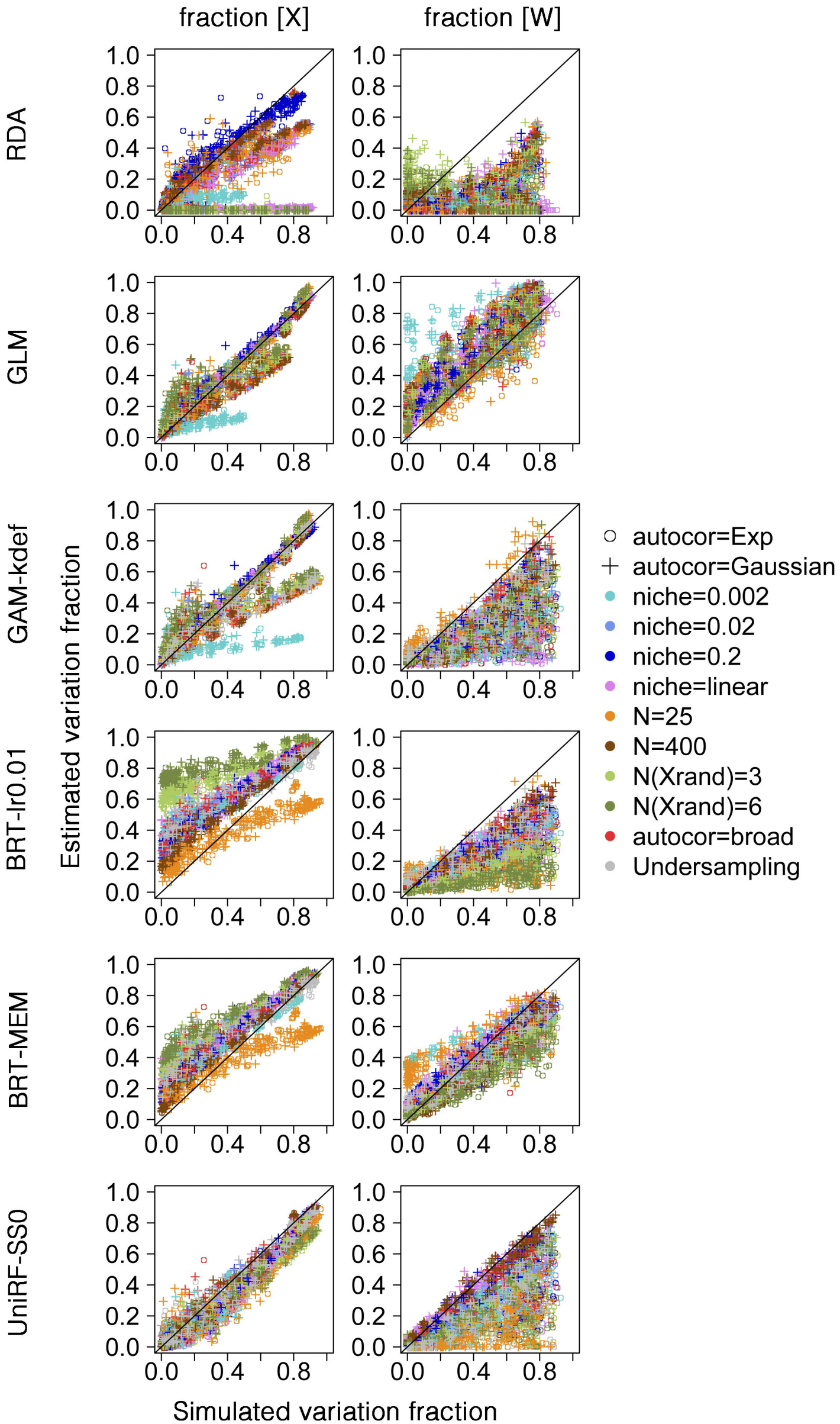


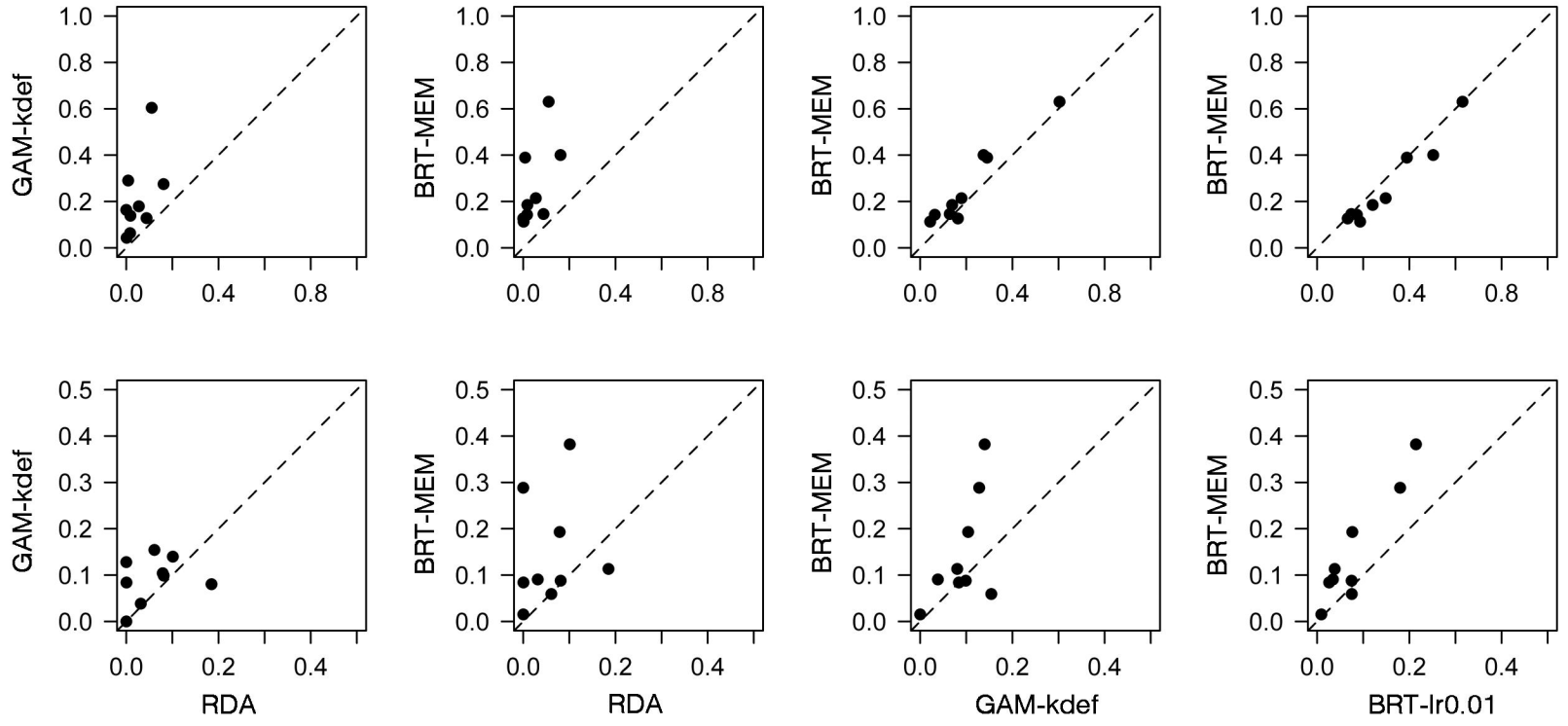
Figure 5. Cross-comparison of the estimated fractions of variation explained by the environment (upper panels) and space (lower panels) between RDA (as a widely used variation partitioning method for multivariate data), GAM, and a tree-based method (BRT; either with MEMs as spatial predictors or with spatial coordinates and a learning rate of 0.01) for the empirical datasets. The dashed line represents the 1:1 line. See Appendix S2 for references.


GLM
GAM
ML
Normal data
Counts data
Binary data
(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
(i)
MVRT-MEM
MVRT-CV
MVRT-noCV
MVRF-MEM
MVRF-SS20
MVRF-SS0
UniRF-MEM
UniRF-SS20
UniRF-SS0
BRT-MEM
BRT-lr0.1
BRT-lr0.01
GAM-k10
GAM-kdef
GLM
-1.0
-0.6
-0.2
0.2
Ŷ = Env + Space
Ŷ = Env
Ŷ = Space
r(Y,Ŷ) - r(Ȳ,Ŷ)


GLM (MEM)
GAM (XY)
BRT-lr0.01 (XY)
BRT-MEM (MEM)
r(Y,Ŷ) - r(Ỹ,Ŷ)
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
narrow
intermediate
wide
linear
Shape of the environmental response
random
exp-fine
exp-broad
Gau-fine
Gau-broad
Spatial variogram model


RDA
MRM
GLM
GAM
ML
Normal data
Counts data
Binary data
Environmental fraction [X]
Spatial fraction [W]
MVRT-MEM
MVRT-CV
MVRT-noCV
MVRF-MEM
MVRF-SS20
MVRF-SS0
UniRF-MEM
UniRF-SS20
UniRF-SS0
BRT-MEM
BRT-lr0.1
BRT-lr0.01
GAM-k10
GAM-kdef
GLM
MRM
RDA
-1.0
-0.5
0.0
0.5
1.0
Estimated fraction - simulated fraction


fraction [X]
fraction [W]
RDA
GLM
GAM-kdef
BRT-lr0.01
BRT-MEM
UniRF-SS0
Estimated variation fraction
Simulated variation fraction
0.0
0.2
0.4
0.6
0.8
1.0
autocor=Exp
autocor=Gaussian
niche=0.002
niche=0.02
niche=0.2
niche=linear
N=25
N=400
N(Xrand)=3
N(Xrand)=6
autocor=broad
Undersampling


GAM-kdef
BRT-MEM
BRT-MEM
BRT-MEM
GAM-kdef
BRT-MEM
BRT-MEM
BRT-MEM
RDA
RDA
GAM-kdef
BRT-lr0.01