Partitioning environment and space in species-by-site matrices: a comparison of methods for community ecology and macroecology

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

[1]German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Deutscher Platz 5e, D-04103 Leipzig, Germany

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

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

*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 composition across sites (hereafter species-environment relationship; SER). Different methods have been used to estimate SERs, but their differences and respective reliability remain poorly known. We compared the performance of four families of statistical methods in estimating the contribution of the environment to explain variation in the occurrence and abundance of co-occurring species while accounting for spatial correlation. These methods included distance-based regression (MRM), constrained ordination (RDA and CCA), generalised linear, mixed, and additive models (GLM, GLMM, GAM), and tree-based machine learning (regression trees, boosted regression trees, and random forests). We first used a simple process-based simulation model of community assembly to generate data with a known strength of (i) niche processes driven by environmental conditions and (ii) spatial processes driven by environmental autocorrelation and dispersal limitation. Then we applied the different methods to infer the spatially-explicit SER and compared their performance in partitioning the environmental and spatial fractions of variation. We found that machine learning methods, namely boosted regression trees and random forests, most accurately recreated the true trends of both occurrence and abundance data. GAM was also a reliable method, though likelihood optimisation did not converge for low sample sizes. The latter is a good option if a priori hypotheses on the functional type of individual species-environment relationships are considered. The remaining methods performed worse under virtually all simulated conditions. Our results suggest that tree-based machine learning is a robust and user-friendly approach that can be widely used for partitioning explained variation in species-by-site matrices. The appropriate use of methods to estimate SERs and assess the importance of drivers of community assembly and species distributions across studies, spatial scales, and disciplines will contribute towards synthesis in community ecology and biogeography.

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

Introduction

The environment is a major driver of species occurrence and abundance. The interaction of species with the environment shapes their ecological niche and influences many facets of biodiversity distribution, from fine-scale community composition to large-scale species distributions and co-occurrence (Chase and Leibold 2003). Thus, 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), as well as biogeographical and macroecological theory (Townsend Peterson et al. 2011). Species-environment relationships (SERs) have been widely estimated to (i) characterise the species' niches and model species distributions, (ii) explore the importance of niche selection as an important biodiversity process, and (iii) correct for environmental effects when studying biotic interactions and other community assembly processes.

However, depending on the study objectives and ecological discipline, SERs have been estimated using different methods. In community ecology, a popular topic has been to disentangle the relative importance of environmentally driven (niche) processes from spatial processes often associated with neutral theory (Leibold and Chase 2017). In this context, SERs have most commonly been estimated using multivariate regression methods, usually constrained ordination such as CCA and RDA (Cottenie 2005, Peres-Neto et al. 2006, Soininen 2014). These methods became popular due to their ease of use, as well as their ability to account for spatial correlation and to partition the explained variation according to environmental and spatial effects. However, the ability of these methods to accurately disentangle the contribution of environmental and spatial variables has been criticised (e.g. Gilbert and Bennett 2010, Smith and Lundholm 2010). A new approach has been proposed to deal with criticism, specifically on how to disentangle spatial correlation produced by autocorrelated environment from that produced by other processes such as dispersal limitation (Clappe et al. 2018), but the appropriateness of these methods to model non-linear species responses to the environment has been less explored.

An alternative way to estimate SERs is to use generalised linear, additive, or mixed modelling (GLM, GAM, GLMM). Although these models are widely used to model univariate responses in the context of species distribution modelling, their ability to model multi-species responses is less explored. Even joint species distribution models (JSDM; Pollock et al. 2014, Ovaskainen et al. 2017) are relatively new, and more testing and exploration is needed before evaluating their effectiveness to model niche responses and partitioning sources of variation. Yet another alternative is to deploy machine learning methods such as Random Forest and Boosted Regression Trees (Elith and Graham 2009), and their multivariate versions (Nieto-Lugilde et al. 2018). Tree-based methods 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 thorough comparisons of different models to estimate species distributions have been made, mostly based on generalised linear models and machine learning models (see Norberg et al. 2019), for a comprehensive review and comparison of methods), these have not yet been widely used in community ecology, and as far as we are aware, never in variation partitioning analysis to evaluate SERs.

Another challenge with the estimation of the SER is that species occurrence and abundance may be autocorrelated due to spatial dependence caused by dispersal limitation and/or environmental autocorrelation. Many methods exist to model spatial structures and account for autocorrelation in the response variable (Dormann et al. 2007, Bivand et al. 2013), but the appropriateness of each method depends on the ecological question and statistical approach used to estimate the SER (more details provided in Methods).

Our aim was to compare the performance of different methods in estimating SERs and partitioning explained variation in species-by-site matrices. To this end, we assessed the performance of each method in quantifying the proportion of variation attributed to spatially correlated and uncorrelated environment, as well as the proportion of variation exclusively attributed to space. We applied the different methods with the single purpose of estimating the relative contributions of environment and space to explaining the occurrence or abundance of multiple species sampled in a given area. Explanation, rather than prediction, has been the goal of numerous studies in community ecology and biogeography, in which the most popular method has been constrained ordination, such as CCA and RDA. However, many other methods based also on regression *sensu lato*, in which the minimisation of a loss or risk function such as mean square error (MSE) or likelihood is used for model fitting, can also be used for partitioning the sources of variation in species-by-sites matrices according to goodness-of-fit measures, namely R^2 or pseudo-R^2.

Methods

The general methodological approach consisted of (1) generating species-by-site matrices using a spatially explicit process-based simulation of community assembly, (2) estimating the SERs according to the different statistical methods (Table 1), and (3) comparing the performance of the different methods to perform variation partitioning. All the computer code is available in https://github.com/duarte-viana/iVarPart.

Simulations of community structure

We used a simple simulation that allowed us to know the true contributions of environment and space to explaining variation in multi-species occurrence and abundance underlying community structure (largely following Gilbert and Bennett 2010). A metacommunity was simulated on a grid where species abundance or occurrence were defined by the additive effect of an environmental and a spatial component. The environment consisted of two environmental gradients, one that was spatially correlated and the other that was randomly distributed. The spatial component was created by simulating independent species distributions and weighting all cells of the grid by a dispersal kernel. Both components produce spatial autocorrelation in the species data, either through the response to the spatially correlated environment or through the pure spatial effect caused by the dispersal limitation. The environmental and spatial components were given different weights (*We* and *Ws*, respectively; *We* + *Ws* = 1) to control their relative contribution to the final species data (abundance or binary occurrence). Thus, these weights represented the true trend in the variation fractions. See the detailed description of the simulation model in Appendix S1.

Statistical methods

We considered four families of statistical methods: distance-based regression, constrained ordination, generalised linear, additive, and mixed models, and tree-based machine learning (Table 1). Note that constrained ordination methods are strictly multivariate, while generalised linear models and tree-based methods stack the predictions for each species. These methods can fit the species-environment, and especially species-space relationships, in a variety of ways, and for some methods more than one approach can be used. Because considering all the possible variations within methods would quickly result in a huge list of methods, we opted for using the most popular and widely used approaches for the sake of tractability (detailed in Table 1 and described below). Independently of how the environmental and spatial components of the models were estimated, we could always partition their contribution for explained variation and compare it with the true contribution to assess the performance.

Distance-based regression

Modelling compositional distance (i.e. pairwise beta-diversity) has been used to investigate the relative importance of environment and space in determining community assembly patterns by converting species-by-site abundance or occurrence into compositional distances between sites (Tuomisto and Ruokolainen 2006). We acknowledge that this method does not directly compare to the other methods considered here, as the response variable consists of compositional distances between communities rather than raw species' abundance or occurrence. Still, since this method is often considered suitable for disentangling sources of variation (see Legendre et al. 2005, and the discussion that followed in, e.g., Laliberté 2008, Tuomisto and Ruokolainen 2008, Legendre et al. 2008), we also assessed its performance. The most common distance-based regression method is the Multiple Regression on distance Matrices (MRM), which is simply the adaptation of the Mantel regression to multiple regression analysis (Lichstein 2007). It consists in 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 both species' abundance and occurrence data (according to the metrics used in the R package *vegan*; Oksanen et al. 2017) and Euclidean distances for the predictors. A simple linear regression was performed on the pairwise distances using function 'lm' of the R stats package (R Development Core Team 2017).

Constrained ordination

We used Canonical Correspondence Analysis (CCA; Ter Braak 1986, Legendre and Legendre 2012), Redundancy Discriminant Analysis (RDA; van den Wollenberg 1977) and its variant, the distance-based RDA (dbRDA; Legendre and Anderson 1999). Ordination analyses summarise the variation of a multidimensional object, such as a species-by-site matrix, into a smaller number of dimensions (or axes) to detect general patterns that cannot be uncovered by single-species analyses (Zuur et al. 2007). These ordinations can be constrained by independent explanatory variables such as environmental and spatial variables. Constrained ordination looks for linear combinations of predictors that best explain the variation of the species-by-site matrix (Legendre and Legendre 2012). In RDA, the response variables are modelled with ordinary linear regression, whereas the modern implementation of CCA relies on weighted linear regression on the Chi-square-transformed species matrix (Legendre and Legendre 2012). Although RDA can yield similar results to CCA depending on data transformations (Blanchet et al. 2014), we decided to compare the most widely used version – RDA on Hellinger-transformed species abundance and occurrence data. This transformation downweighs the effect of double species absences (double-zeros) and minimizes arch effects (Legendre and Gallagher 2001). For dbRDA we used the Bray-Curtis dissimilarity measure for both abundance and occurrence data, another widely used distance metric (and the default in the *vegan* R package). The only difference to RDA is that the response matrix is not the species-by-site matrix but a matrix composed of principal components resulting from a principal coordinates analysis (PCoA) of the distance matrix. We used the R functions ‘lm’ from the R *stats* package to perform RDA, ‘cca’ to perform CCA, and ‘capscale’ to perform dbRDA, the latter two from package *vegan* (Oksanen et al. 2017, R Development Core Team 2017).

Generalised linear, additive, and mixed models

Generalised linear models allow for more flexible distributions of response data than constrained ordination (but see more flexible ordination methods in Yee 2015), in particular Poisson and binomial distributions typical of abundance (count) and binary occurrence data. Further, the inclusion of quadratic terms for environmental predictors in Poisson and binomial models can accommodate narrow niche unimodal responses that typically contain many absences (i.e. many zeros) (Fig. 1a). This is not possible when fitting Gaussian regression models with identity link function, which may also run into risk of predicting negative abundances (Fig. 1b). We estimated three types of models: (i) simple generalised linear models with a quasi-Poisson error distribution (GLM-qP) and a log link function for abundance data, and a binomial distribution with a logit link for occurrence data (using the *stats* R package; R Development Core Team 2017); (ii) generalised linear mixed models with a Poisson distribution for abundance data, either with a correction for overdispersion (GLMM-overP) or not (GLMM-P), and a binomial distribution for occurrence data, estimating a random intercept and slope for species (using the R package *HMSC*; Ovaskainen et al. 2017); and (iii) generalised additive models (GAM) with a Poisson error distribution for abundance data and a binomial distribution for occurrence data (using the R package *mgcv*; Wood 2017). For all these methods, we fitted a second-degree polynomial for the environmental variables (i.e. their linear and quadratic effects). The spatial effects were modelled with MEMs (see below), except for GAM, in which thin plate regression splines were fitted to the spatial coordinates of the sites. All models were fitted as single models to each species and the predictions were then stacked into a matrix of predicted species abundances, except GLMM, in which the (random) estimates for each species were bounded by a normal distribution, which makes this model a type of joint species distribution model (Ovaskainen et al. 2017). GLM and GAM were fitted by maximum-likelihood estimation, whereas GLMM was fitted using Bayesian estimation. As a side note, we also tried to use GLM with Poisson and negative binomial error distributions for abundance data, but these models often had convergence issues due to the high number of spatial predictors (MEMs; see below) in relation to sample sizes. We excluded these models from the comparison, although we recognise that these can be appropriate if convergence is achieved.

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 species-environment relationship; instead, the relationship is learned entirely from the data (see model specifications and settings in Table 1).

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 compared two variants of MVRT. In one, we fitted a tree with the minimum number of observations in any terminal ("leaf") node fixed to 5. In the other, we used cross-validation to identify the most parsimonious tree. We first fitted a sequence of trees of size ranging from 1 to 20 splits, and identified the tree size with lowest out-of-sample predictive error using 5-fold cross-validation, and we noted its complexity parameter (CP). We then pruned the full tree with 20 splits until it reached the CP value of the best cross-validated tree.

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 the minimum number of observations in any terminal ("leaf") node set to 5, each fitted to data resampled randomly with replacement, using a random subset (1/3) of the predictors. The predictions of these trees are then averaged to make the overall predictions.

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 implementation in R function 'rfsrc' (package *randomForestSRC*; Ishwaran et al. 2019). We compared two variants of MVRF. In one, as in MVRT, we used the minimum number of observations in any terminal ("leaf") node fixed to 5. In the other, we used cross-validation to find the most parsimonious size of the terminal node – we fitted a sequence of random forests with terminal node size going from 2 to 15, and we chose the tree with the lowest out-of-bag CNMSE. In both cases we 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 their total number.

Multivariate 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). Unlike the univariate version, this implementation allows to explain species covariance once the model has been fitted (though we did not do it, and thus we just fitted a series of univariate boosted regression trees to each species). We used the implementation in R function 'mvtb' (package *mvtboost*), and fitted four variants of the algorithm with two tree depths (i.e. interaction depths) and with or without cross-validation to identify the total number of trees. All four variants used learning rate (shrinkage) of 0.01. The tree depths that we examined were 1, which is a single split, also known as "regression stump", and 3, which allows for a three-way interaction between the predictors. The variant without cross-validation used a fixed number of 1000 trees. In the cross-validated variant, we used a 5-fold cross-validation to identify the most parsimonious number of trees, from 1 up to 1000. We chose the BRT model with number of trees giving the lowest multivariate out-of-sample prediction error.

Spatial component (spatial correlation)

Space has been incorporated in statistical models in various ways to deal with autocorrelation. When spatial prediction and interpolation is the goal, geostatistical techniques such as kriging are typically used (Bivand et al. 2011). But often the objective is simply to control for spatial autocorrelation in the residuals, for which autoregressive models and mixed-effects models are commonly used (e.g. Dormann et al. 2007). These models usually assume a simple autocorrelation error structure such as an exponential decay as a function of distance, which might be overall simplistic for complex habitat configurations. Other, more complex, techniques have been used to model space while capturing multiple spatial scales, including high-order polynomials of spatial coordinates (or trend-surface analysis) and eigenvector-based spatial variables (namely Moran's eigenvector maps, MEM; for a more complete overview of spatial methods, see Bivand et al. 2011). The latter has been considered more effective to model multi-scale patterns (Dray et al. 2006, 2012). As such, we used distance-based MEM variables, calculated using the R package *adespatial* (Dray et al. 2016), as spatial predictors in constrained ordination and generalised linear models (but not GAM). For tree-based methods and GAM, that are able to model complex non-linear relationships, it would be redundant to use MEM instead of taking advantage of the non-linear modelling capabilities of the methods. Therefore, for these methods we simply used the spatial coordinates as spatial predictors (see an example of a non-linear spatial surface fitted to abundance data in Appendix S2, Fig. S2.1).

Variation partitioning

We performed variation partitioning to estimate the fractions of variation explained (i.e. R^2) by the exclusive and shared contributions of environment and space. Traditionally, the shared fraction was exclusively attributed to the effect of spatially correlated environment (Borcard et al. 1992, Peres-Neto et al. 2006). However, because spatial autocorrelation in species data is caused by both spatially correlated environment and dispersal limitation, spurious correlations between species distributions and environment can arise (Smith and Lundholm 2010, Clappe et al. 2018). To correct for spurious correlations and avoid inflated fractions of variation explained by the environment, we used the method developed in Clappe et al. (2018). This method partials out spurious correlations by randomising the environmental variables while keeping their spatial structure intact, a method known as Moran Spectral Randomisation (MSR; Wagner and Dray 2015), and then accounting for their contribution in the variation partitioning procedure. As such, we can attribute the exclusive fraction of the environment to spatially uncorrelated environment and the shared fraction to spatially correlated environment. These two fractions were summed up to obtain a global environment fraction [E]. Fraction [S] was the exclusive contribution of space. In some circumstances, when either [E] or [S] are virtually null, their values can be slightly negative. In this case we set negative values to zero. For each method we used an appropriate R^2 or pseudo-R^2 metric to perform the variation partitioning, preferably choosing the metric used in the R package where the method was implemented; however, because different metrics have been used for the same method, we also compared the performance of alternative metrics (see the metrics used in Table 1 and their definitions in Appendix S1). The R^2 metrics used here were categorised as a eigenvalue-based metric (R2-mv; used for constrained ordination methods), a classical MSE-based metric (R2-cla for abundance data and R2-Efr for binary data), a deviance-based metric (R2-McF), and a discrimination metric for binary data (R2-Tju).

Performance of the different methods

To evaluate the performance of each method, we calculated the root mean squared error (RMSE) and the Kendall rank correlation coefficient (τ) between the expected (true) and estimated values of [E] (as a proportion of total variation explained, i.e., [E]/([E]+[S]). The RMSE provides a measure of how far the estimation is from the true trend (the lower the better). The rank correlation is a measure of discrimination, i.e. how well the method discriminates community matrices with low or high [E], relatively to other matrices, but irrespectively of the absolute values of [E]. To calculate a combined index of performance index (hereafter just "performance"), we rescaled the RMSE and rank correlation to range between 0 and 0.5, inverted the rescaled RMSE (now the higher the better), and summed both components. Thus, performance ranged from 0 (worst) to 1 (best). The combined index was consistent with the visual evaluation of method performance.

In addition, because the methods considered here are more or less prone to overfitting (e.g. machine learning methods have been criticised for overfitting; Wenger and Olden 2012), we compared the true (simulated) with the estimated amount of explained variation (i.e. R^2). The true R^2 was estimated by using the true (simulated) abundance as the predicted values and the abundance sampled randomly from a Poisson distribution with means corresponding to true abundance as the observed values (i.e. the data that was used to fit the different models with the different methods; see Appendix S1); and the R^2 for each method was calculated as the sum of all explained variation fractions (i.e. [E] + [S]). We used the same R^2-metric for both the true and estimated values (see Appendix S1 for the calculation of the different R^2 metrics). If overfitting had been observed, we expected the estimated values to be systematically higher than true values, meaning that the model could be fitting random noise produced by the random sampling.

Empirical evaluation

To explore the impacts of method choice on empirical cases, we compared the traditional ordination method RDA and dbRDA as the traditional used methods, with BRT and MVRF as the best performing methods, by performing variation partitioning in nine empirical datasets publicly available. Each empirical dataset consisted of a species abundance matrix, some environmental variables and the geographical coordinates of the sampled sites (see Appendix S3 for dataset details and sources).

Results

The performance of the different methods varied considerably, and was generally better for abundance data compared to binary occurrence data (Fig. 2; Appendix S2, Fig. S2.2 for all tested variants of the different methods). Tree-based machine-learning methods had the best performance across the different scenarios for both binary and abundance data (Fig. 2). These were followed by GAM, which had slightly worse performance. However, the GAM likelihood optimization did not converge for the lowest sample size (N=25) and performed considerably worse for the simulations with narrowest species' responses to the environment (niche breadth = 0.002) (Fig. 3; Appendix S2, Fig. S2.3 for all methods). The remaining methods performed worse (Fig. 2), among which MRM (the distance-based regression) had the worst performance. None of the methods overfitted the data, as the respective total R^2 was virtually always below the true R^2 (Fig. 4).

The good performance of machine learning was generally consistent among its different methods, with BRT consistently performing well for both abundance and binary data, together with MVRF for abundance data and UniRF for binary data (Fig. 2). An interaction depth of 3 in the BRT algorithm was better for abundance data, whereas an interaction depth of 1 (a single "stump") was better for binary data (Appendix S2, Fig. S2.2). The total variation explained by the BRT models was higher compared to other tree-based methods and constrained ordination, and similar to GAM (Fig. 3), but we did not observe overfitting (Fig. 4). When cross-validation was used in BRT, MVRF, and MVRT, the performance decreased considerably (Appendix S2, Fig. S2.2) due to underestimation of [S] and consequent overestimation of [E] (Appendix S2, Fig. S2.4).

Ordination methods (CCA, RDA, dbRDA) performed consistently worse than other methods for abundance data (Fig. 2 and 3). However, for binary data, ordination methods performed similarly to GLM, but worse than GAM and the tree-based methods MVRF and MVRT (Fig. 2). Ordination methods were not able to reliably model [E], as observed by the general underperformance when true [E] was high (Fig. 3).

Notwithstanding the good performance of GAM, the parametric regression models (GLM, GLMM) had intermediate performances (Fig. 2). This family of methods had acceptable performances under wider niche breadths and larger sample sizes, but the performance considerably dropped when niche breadth was lowest (sharp abundance peaks at given environments and many absences) and sometimes failed to converge under low sample sizes. GLMM had high rank correlation with true fractions (Fig. 2) but the variation partitioning was biased towards higher spatial fractions [S], especially for binary data (Fig. 3). This is probably due to overfitting caused by MEM variables and the lack of an adjustment procedure for the R^2 of this kind of models. In fact, GLM-qP, whose variation fractions were adjusted for the number of predictors, were less biased in general (lower RMSE), but also had lower discriminatory power (Fig. 2).

In general, the type of R^2 metric had a minor impact on results for abundance data, which can be visualised by the clustering of methods in the general ranking, although deviance-based pseudo-R^2 (R2-McF) tended to provide better performance (Appendix S2, Fig. S2.2). However, for binary data the R^2 metric resulted to be important for some methods. For example, uniRF was the best performing method with R2-Tju, whereas its performance with R2-Efr and R2-McF was amongst the worst. In general, the Tjur pseudo-R2 showed the best results for binary data, except for constrained ordination, for which only the R2-mv was calculated.

When applied to empirical data, the choice of method had a clear influence on the results of the variation partitioning. We fitted dbRDA as an example of a widely used method together with RDA, as a potentially mis-specified model, and BRT and MVRF, as the best performing methods. We also tried GAM and GLM, but these did not converge for some datasets with low sample sizes or a large number of environmental predictors. We thus excluded the latter from further comparisons. The variation partitioning results were quite similar between dbRDA and RDA, and between BRT and MVRF, but were substantially different between ordination and tree-based methods (Fig. 5).

Discussion

Our goal was to find the best method to estimate the relative importance of the environment to explaining variation in species-by-site matrices, over other processes that cause spatial structure in the geographic distribution of species. Clearly, as far as our simulations could tell, the best performing methods were tree-based machine learning methods and GAM – both flexible methods that are entirely or partly non-parametric, respectively. However, GAM failed to converge for low sample sizes and had poorer performance when species responses to the environment were narrowest (a summary of general results and recommendations is provided in Appendix S4).

Tree-based machine learning has several advantages, in that it is able to fit a variety of responses to the environment, including narrow niche responses and potential interactions between different environmental variables, while simultaneously modelling space at various levels of complexity by using only the spatial coordinates of the sampling locations. These methods also dealt better with narrow niche breadth in comparison with the remaining methods. Typically, as spatial extent increases, the size of the environmental gradients also increases beyond the range of values where species occur, causing niche responses to narrow. Tree-based methods are, therefore, an effective tool that can be used across scales. However, machine learning has also been criticised, especially for overfitting (e.g. Wenger and Olden 2012). Our analysis, nevertheless, did not show any significant overfitting, and even if some overfitting had been observed, there are no reasons to believe that the relative contributions of environment and space to explained variation could be biased.

Tree-based methods, however, cannot be used when specific hypotheses about SERs are considered. One may need to test an *a priori* hypothesis and, to that end, it may be instrumental to use parametric models rather than learning algorithms. For example, if we want to test for unimodal, Gaussian-like responses, assuming that more complex responses could be caused by other processes such as biotic interactions, it might be better to limit the scope of the analysis to linear and quadratic responses. Our results indicate that generalised linear models can be used if sample size is sufficient for the models to converge and if the niche responses are not too narrow (see also Appendix S4), which is something that can be guessed in an exploratory data analysis prior to the actual model fitting. In case MEMs are intended to be used and cause convergence issues, GAM with smooth splines for spatial coordinates can be used to avoid loss of degrees of freedom, as overfitting is inherently penalised (Wood 2017). We also note that the environmental component can be modelled with splines when using GAM, though sample size is a limitation and the parametric advantage of GAM is lost.

Even though the ordination methods were based on different data transformations, the results of CCA, dbRDA, and RDA, were largely similar (Appendix S2, Fig. S2.2 and S2.3). The worse performance of these methods is due to the general underestimation of the environmental component, arguably because of model misspecification. We also note that the distance-based regression (MRM) had a poor performance (Appendix S2, Fig. S2.2, S2.3), thus we recommend to avoid this method if the goal is to estimate the relative importance of environment and space to explaining variation in species-by-site matrices.

While we have pointed to some better, and some worse, methods for disentangling SERs, none of the methods assessed here is perfect, and thus there is margin for future improvement. Machine learning methods are diverse, and many more methods can be tested and even developed specifically for the purposes here outlined (e.g. D'Amen et al. 2017, Nieto-Lugilde et al. 2018). Other avenues for improvement might include the refinement of R^2-like indices. For example, one obvious limitation of GLMM (which are widely used in joint species distribution models) was the failure to penalise overfitting caused by the spatial variables (MEMs). Adjustment procedures such as information criteria penalizing complexity or "adjusted" R^2 can be tested (Burnham and Anderson 2002). Cross-validation can also theoretically be used to avoid overfitting, but note that models based on MEMs have not been used with cross-validation (see also Roberts et al., 2017), and non-linear spatial surfaces such as those fitted with GAM and tree-based methods are hardly suited for extrapolation (outside the sampled area). Indeed, we found strong evidence that the worse performance of tree-based methods when using cross-validation is due to the poor predictive ability of the complex spatial effect (resulting in considerable underestimation of the spatial fraction of variation). We, therefore, recommend not to use cross-validation when the objective is to partition explained variation in species-by-sites matrices.

We here proposed and evaluated alternative methods that can be used to estimate the relative contribution of the environment to explain species composition while disentangling spatial dependence. The ecological interpretation of variation partitioning depends on the context of the study, and the integration of other drivers and/or types of information might be needed. For example, to conclude about the relative importance of niche selection and environmental filtering, further information such as species traits might have to be integrated in the analysis (Cadotte and Tucker 2017). The SER is just the first step to quantify the contribution of measured environment to determine species abundance and co-occurrence while accounting for spatial correlation, regardless of other confounding effects, namely biotic interactions. We note that every method used here is valid and reliable whenever the models are correctly specified and assumptions are met. We were only interested in assessing their relative performance for this particular aim of using SERs to perform variation partitioning.

We showed that the popular ordination methods CCA and RDA, which are widely used to partition variation explained by environment and space rather than predicting species occurrence and abundance, fail to provide accurate estimates of variation fractions in many circumstances, and thus should not be used uncritically. We highlight tree-based machine learning as a flexible alternative that can be widely used with both abundance and occurrence data. If a priori hypotheses about SERs are considered, GAM as a semiparametric method is a good choice for performing variation partitioning. We also recommend the use of a deviance-based (McFadden's) pseudo-R^2 for abundance data and Tjur's pseudo-R^2 for occurrence data. By choosing appropriate methods to model different species responses to the environment, from linear responses typically observed at smaller spatial scales to unimodal responses typically observed over larger scales, our recommendations can as well apply to both community ecology and macroecology studies.

Acknowledgements

We thank Jon Chase 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 the Alexander von Humboldt Foundation and sDiv, the Synthesis Centre of iDiv.

Author contributions

The first author performed simulations of metacommunities, consolidated and applied code for the different methods, and led the analysis and writing. All authors conceived and developed the ideas, coded the methods, and contributed to the writing.

References

Bivand, R. S. et al. 2013. Applied Spatial Data Analysis with R: Second Edition. - Springer Science & Business Media.

Blanchet, G. F. et al. 2014. Consensus RDA across dissimilarity coefficients for canonical ordination of community composition data. - Ecol. Monogr. 84: 491–511.

Borcard, D. et al. 1992. Partialling out the spatial component of ecological variation. - Ecology 73: 1045–1055.

Breiman, L. 2001. Random forests. - Mach. Learn. 45: 5–32.

Burnham, K. P. and Anderson, D. R. 2002. Model Selection and Multi-Model Inference: a Practical Information-Theoretic Approach. - Springer-Verlag New-York, Inc.

Cadotte, M. W. and Tucker, C. M. 2017. Should Environmental Filtering be Abandoned? - Trends Ecol. Evol. 32: 429–437.

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

Chesson, P. 2000. Mechanisms of maintenance of species diversity. - Annu. Rev. Ecol. Syst. 31: 343–366.

Chesson, P. L. and Warner, R. R. 1981. Environmental Variability Promotes Coexistence in Lottery Competitive Systems. - Am. Nat. 117: 923–943.

Clappe, S. et al. 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. - Ecol. Lett. 8: 1175–1182.

D'Amen, M. et al. 2017. Spatial predictions at the community level: from current approaches to future frameworks. - Biol. Rev. 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. et al. 2007. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. - Ecography (Cop.). 30: 609–628.

Dray, S. et al. 2006. Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbour matrices (PCNM). - Ecol. Modell. 196: 483–493.

Dray, S. et al. 2012. Community ecology in the age of multivariate multiscale spatial analysis. - Ecol. Monogr. 82: 257–275.

Dray, A. S. et al. 2016. adespatial: Multivariate Multiscale Spatial Analysis.: R package, Version: 0.0-7.

Elith, J. and Graham, C. H. 2009. Do they? How do they? WHY do they differ? on finding reasons for differing performances of species distribution models. - Ecography (Cop.). 32: 66–77.

Ellis, N. et al. 2012. Gradient forests: Calculating importance gradients on physical predictors. - Ecology 93: 156–168.

Friedman, J. H. 2001. Greedy function approximation: A gradient boosting machine. - Ann. Stat. 29: 1189–1232.

Gilbert, B. and Bennett, J. R. 2010. Partitioning variation in ecological communities: do the numbers add up? - J. Appl. Ecol. 47: 1071–1082.

Hastie, T. et al. 2009. The elements of statistical learning: prediction, inference and data mining. - Springer-Verlag, New York in press.

Ishwaran, H. et al. 2019. Package 'randomForestSRC.' in press.

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

Legendre, P. and Anderson, M. J. 1999. Distance-based redundancy analysis: testing multispecies responses in multifactorial ecological experiments. - Ecol. Monogr. 69: 1–24.

Legendre, P. and Gallagher, E. 2001. Ecologically meaningful transformations for ordination of species data. - Oecologia 129: 271–280.

Legendre, P. and Legendre, L. 2012. Numerical ecology. - Elsevier Science BV.

Legendre, P. et al. 2005. Analyzing beta diversity: Partitioning the spatial variation of community composition data. - Ecol. Monogr. 75: 435–450.

Legendre, P. et al. 2008. Analyzing or explaining beta diversity? Comment. - Ecology 89: 3238–3244.

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

Leibold, M. A. et al. 2004. The metacommunity concept: a framework for multi-scale community ecology. - Ecol. Lett. 7: 601–613.

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

Miller, P. J. et al. 2016. Finding structure in data using multivariate tree boosting. - Psychol. Methods 21: 583–602.

Nieto-Lugilde, D. et al. 2018. Multiresponse algorithms for community-level modelling: Review of theory, applications, and comparison to species distribution models. - Methods Ecol. Evol. 9: 834–848.

Norberg, A. et al. 2019. A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels. - Ecol. Monogr. 89: e01370.

Oksanen, J. et al. 2017. vegan: Community Ecology Package. R package version 2.0-9.

http://CRAN.R-project.org/package=vegan. in press.

Ovaskainen, O. et al. 2017. How to make more out of community data? A conceptual framework and its implementation as models and software. - Ecol. Lett. 20: 561–576.

Peres-Neto, P. R. et al. 2006. Variation partitioning of species data matrices: Estimation and comparison of fractions. - Ecology 87: 2614–2625.

Pollock, L. J. et al. 2014. Understanding co-occurrence by modelling species simultaneously with a Joint Species Distribution Model (JSDM). - Methods Ecol. Evol. 5: 397–406.

R Development Core Team 2017. R: A language and environment for statistical computing. - R Found. Stat. Comput. Vienna, Austria. ISBN 3-900051-07-0, URL http//www.R-project.org. in press.

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

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

Ter Braak, C. J. F. 1986. Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. - Ecology 67: 1167–1179.

Townsend Peterson, A. et al. 2011. Niches and Geographic Distributions. - Princeton University Press.

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

Tuomisto, H. and Ruokolainen, K. 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.

Wagner, H. H. and Dray, S. 2015. Generating spatially constrained null models for irregularly spaced data using Moran spectral randomization methods. - Methods Ecol. Evol. 6:

1169–1178.

Wenger, S. J. and Olden, J. D. 2012. Assessing transferability of ecological models: An underappreciated aspect of statistical validation. - Methods Ecol. Evol. 3: 260–267.

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

Yee, T. W. 2015. Vector Generalized Linear and Additive Models: With an Implementation in R. - Springer.

Zuur, A. F. et al. 2007. Analyzing Ecological Data. - Springer Science & Business Media.

Table 1. Overview of the methods compared in this study. See the description of the different R^2 or pseudo-R^2 metrics in Appendix S2. The bold R^2 metrics are those chosen as the most appropriate metrics and thus reported here (the results for the other R^2 metrics can be found in Appendix S2, Fig. S2.2, and Appendix S4).

Method family	**Method name**	**Method full name**	**Type***	**R^2 metric for count data**	**R^2 metric for binary data**	**Environmental response**	**Spatial modelling** (input: X and Y coordinates)	**Specific model settings**
Constrained ordination	RDA	Redundancy discriminant analysis	P	**R2-mv**	**R2-mv**	Linear	MEM	- R package: *base, vegan* - Species data: Hellinger-transformed
	dbRDA	Distance-based redundancy discriminant analysis	NP	**R2-mv**	**R2-mv**	Any	MEM	- R package: *vegan* - Species data: raw - Dissimilarity index: Bray-Curtis
	CCA	Canonical correspondence analysis	P	**R2-mv**	**R2-mv**	Unimodal	MEM	- R package: *vegan* - Species data: raw
Distance-based regression	MRM	Multiple regression on distance matrices	NP[§]	**R2-cla**	**R2-cla**	Any/Linear[¶]	XY Euclidean distances	- R package: *base* - Dissimilarity index: Bray-Curtis
Generalized linear and additive models	GLM	Generalized linear model	P	R2-cla, R2-mv, **R2-McF**	**R2-Tju**, R2-Efr, R2-McF	Linear, quadratic	MEM	- R package: *base* - Species data: raw - Distribution: quasi-Poisson (abundance), binomial (binary)
	GLMM	Generalized linear mixed model	P	R2-cla, R2-mv, **R2-McF**	**R2-Tju**, R2-Efr, R2-McF	Linear, quadratic	MEM	- R package: *HMSC* - Species data: raw - Distribution: Poisson w/ or w/o overdispersion (abundance), binomial (binary) - Iterations: 20,000
	GAM	Generalized additive model	NP	R2-cla, R2-mv, **R2-McF**	**R2-Tju**, R2-Efr, R2-McF	Linear, quadratic	Splines on XY coordinates	- R package: *mgvc* - Species data: raw

						(potentially any)		- Distribution: Poisson (abundance), binomial (binary)
Machine learning tree-based methods	BRT	Multivariate boosted regression trees	NP	R2-cla, R2-mv, **R2-McF**	**R2-Tju**, R2-Efr, R2-McF	Any	Splines on XY coordinates	- R package: *mvtboost* - Species data: raw - Distribution: Poisson (abundance), Bernoulli (binary) - #trees: 1000 - Learning rate: 0.01 - Interaction depth: 1 or 3 - CV: no CV or 5-fold
	MVRT	Multivariate regression tree	NP	**R2-cla**, R2-mv, R2-McF	**R2-Tju**, R2-Efr, R2-McF	Any	Splines on XY coordinates	- R package: *mvpart* - Species data: raw (abundance), numeric 0/1 (binary) - Node size: 5 (no CV) - Interaction depth: 1-20 (CV) - CV: no CV or 5-fold
	UniRF	Univariate random forest	NP	**R2-cla**, R2-mv, R2-McF	**R2-Tju**, R2-Efr, R2-McF	Any	Splines on XY coordinates	- R package: *randomForest* - Species data: raw (abundance), categorical 0/1 (binary) - #trees: 500 - Node size: 5
	MVRF	Multivariate random forest	NP	**R2-cla**, R2-mv, R2-McF	**R2-Tju**, R2-Efr, R2-McF	Any	Splines on XY coordinates	- R package: *randomForestSRC* - Species data: raw (abundance), categorical 0/1 (binary) - #trees: 500 - Node size: 5 (no CV), 2-15 (CV) - CV: no CV or 1 per node size

*P=parametric, NP=non-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

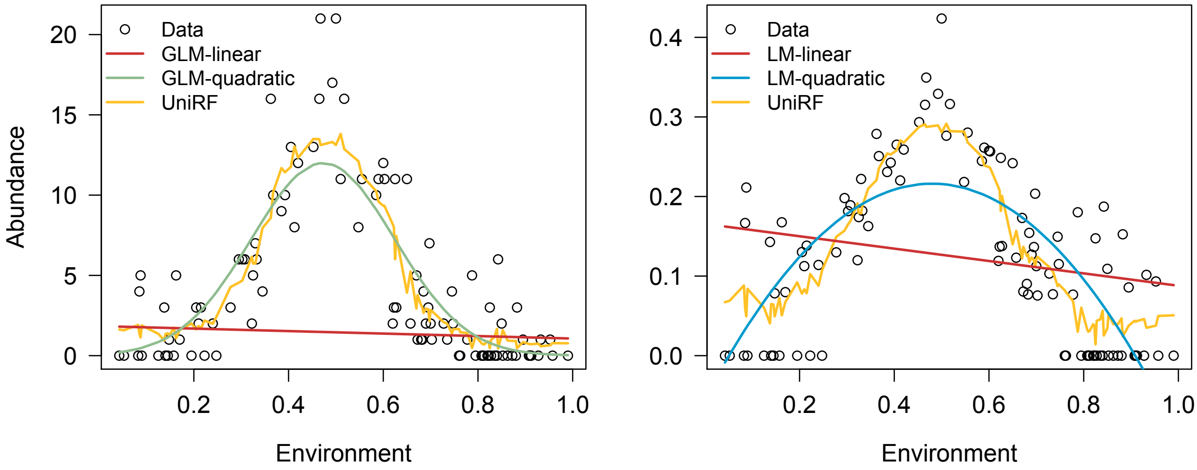

Figure 1. Example of different models fitted to a Gaussian niche of one species. (a) Generalised linear model (GLM) fits with only linear or both linear and quadratic terms, as well as a univariate random forest fit, to Poisson distributed abundances. (b) Linear model (LM) fits with only linear or both linear and quadratic terms, as well as a univariate random forest fit, to Hellinger-transformed abundances.

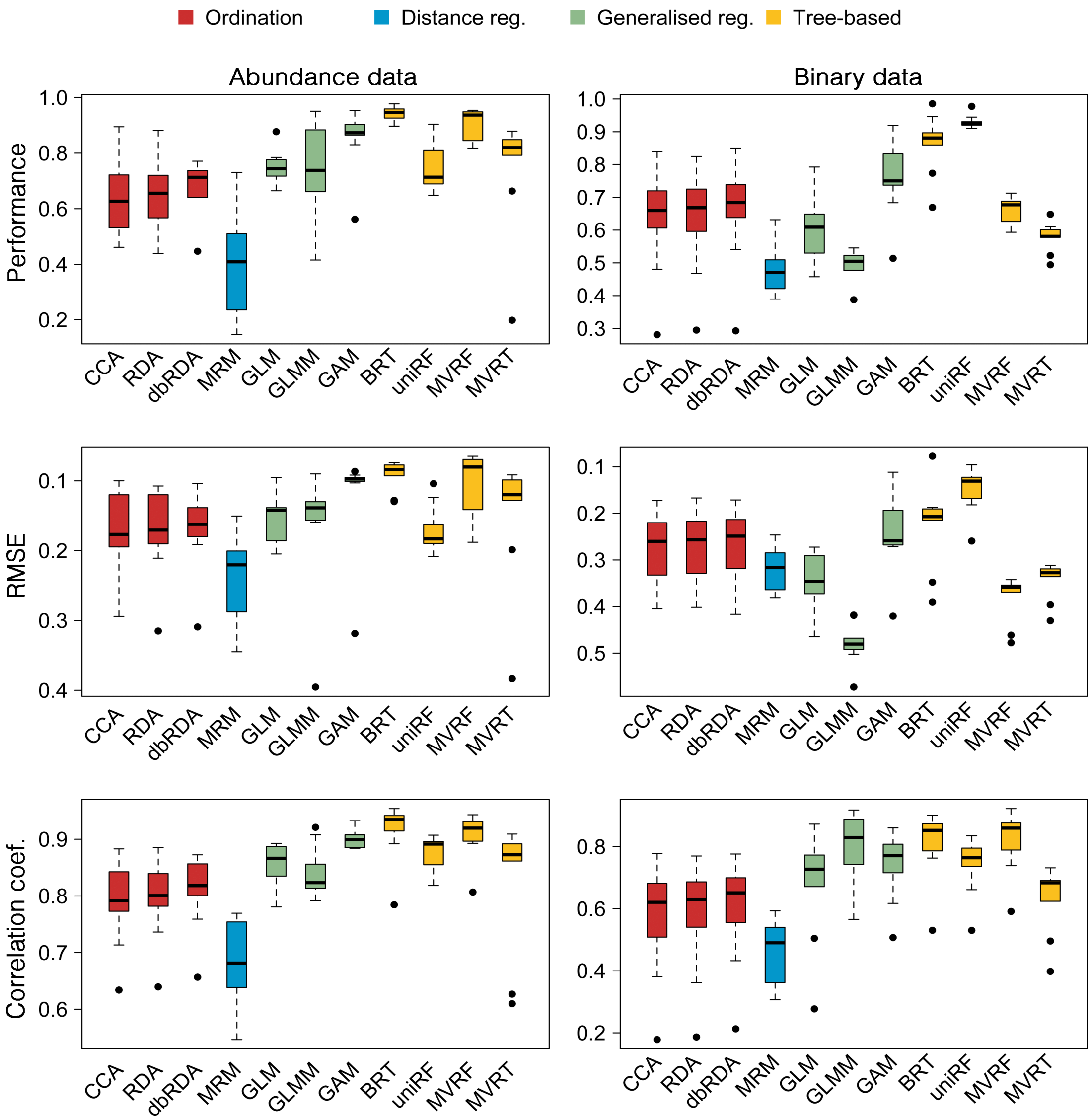

Figure 2. Performance of the different methods for both abundance (left panels) and binary occurrence (right panels) data across the different scenarios: global performance (the higher the better; upper panel), RMSE (the lower the better, but note that the vertical axis is inverted to facilitate comparison; middle panel), and rank correlation coefficient (the higher the better; lower panel). Check the performance for other variants of each method in Appendix 2, Fig. S2.2.

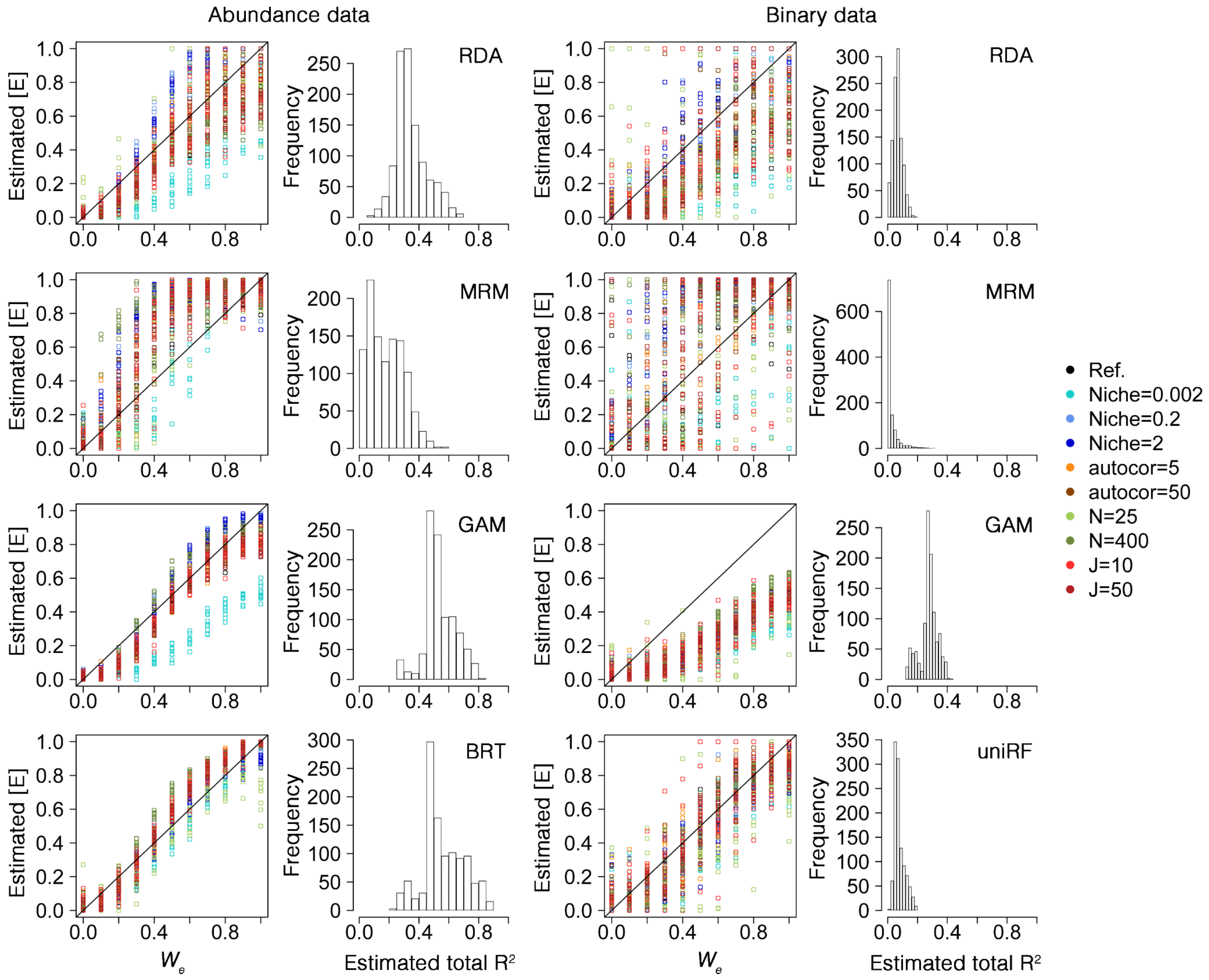

Figure 3. Plots of true vs. estimated fractions of variation [E] (as a proportion of total variation explained; left panels) and the distribution of total variation explained (i.e. total R^2; right panels) for abundance and binary data. The true variation fractions are given by the weights attributed to the environment (W_e) in the simulation model. The black line represents the 1:1 line, where points should fall if the performance of the method had been perfect. The reference scenario (black dots) is defined by N = 100, J = 30, R = 25 and σ = 0.02. The best method of each family of methods is represented (see Appendix S2, Fig. S2.3 to see all methods).

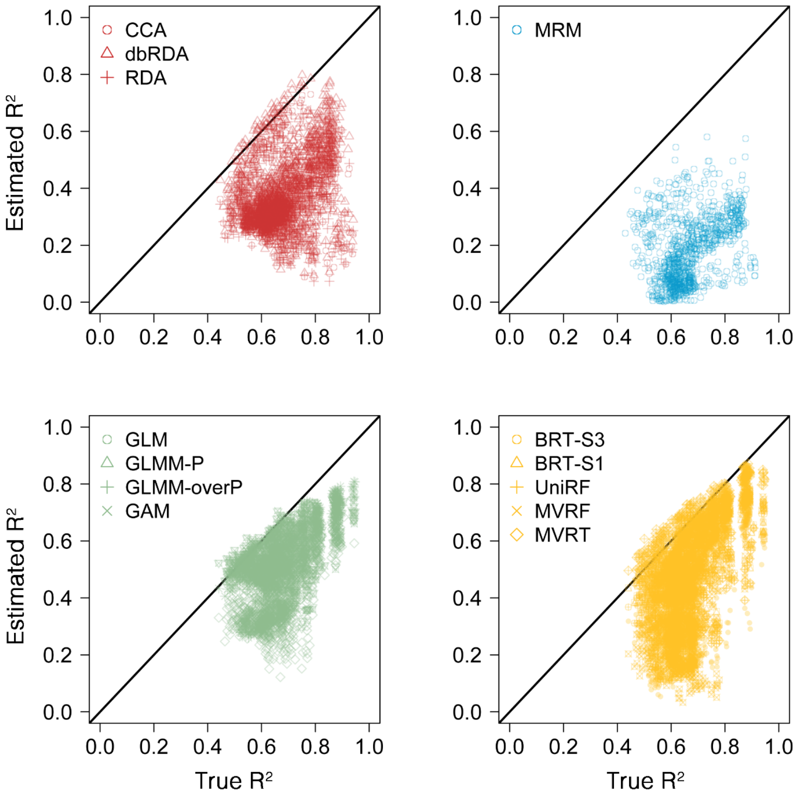

Figure 4. Plots of true vs. estimated total R2. Points above or below the 1:1 line indicate over or underfitting, respectively.

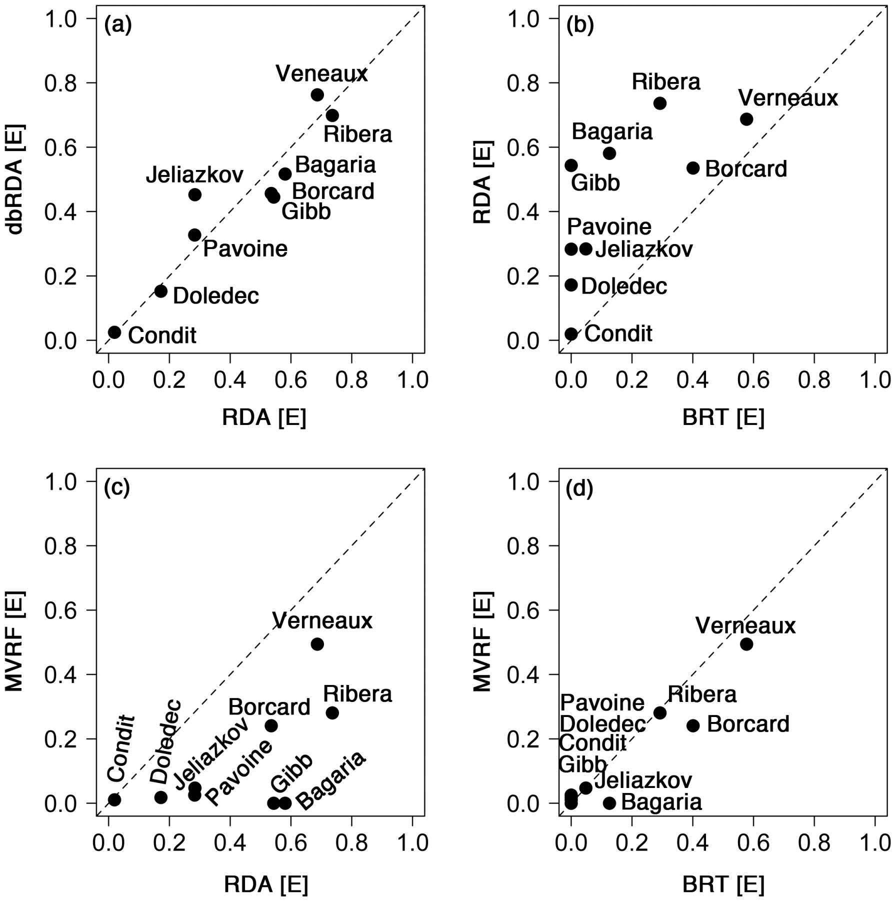

Figure 5. Comparison of the estimated fraction [E] among constrained ordination methods (RDA and dbRDA) and tree-based methods (BRT and MVRF) for each empirical dataset. The dashed line represents the 1:1 line. See Appendix S3 for references.