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

Abstract

Community ecologists and macroecologists have long sought to evaluate the importance of environmental conditions and other drivers in determining species composition across sites. Different methods have been used to estimate species-environment relationships while accounting for or partitioning the variation attributed to environment and spatial autocorrelation, 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 and space to explain variation in multi-species occurrence and abundance. These methods included distance-based regression (MRM), constrained ordination (RDA and CCA), generalised linear and additive models (GLM, GAM), and tree-based machine learning (regression trees, boosted regression trees, and random forests). Depending on the method, the spatial model consisted of either Moran's Eigenvector Maps (MEM; in constrained ordination and GLM), smooth spatial splines (in GAM), or tree-based non-linear modelling of spatial coordinates (in machine learning). We simulated typical ecological data to assess the methods' performance in (1) fitting environmental and spatial effects, and (2) partitioning the variation explained by the environmental and spatial effects. Differences in the fitting performance among major model types – (G)LM, GAM, machine learning – were reflected in the variation partitioning performance of the different methods. Machine learning methods, namely boosted regression trees, performed overall better. GAM performed similarly well, though likelihood optimisation did not converge for some empirical test data. The remaining methods performed worse under most simulated data variations (depending on the type of species data, sample size and coverage, autocorrelation range, and

response shape). Our results suggest that tree-based machine learning is a robust approach that can be widely used for variation partitioning. Our recommendations apply to single-species niche models, community ecology, and macroecology studies aiming at disentangling the relative contributions of space vs. environment and other drivers of variation in site-by-species matrices.

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. It shapes the ecological niche of species 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, Thompson et al. 2020), 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 at species and community levels (Townsend Peterson et al. 2011; Norberg et al. 2019), (ii) explore the importance of niche selection 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 the influence of environment on species distributions over different spatial scales is the correct estimation of environmental effects while accounting for spatial autocorrelation, as species occurrence and abundance may be

autocorrelated due to spatial dependence caused by environmental autocorrelation and/or other processes such as dispersal limitation. In addition, a popular topic in community ecology has been to disentangle the relative importance of environmentally driven (niche) processes from spatial processes often associated with neutral theory using variation partitioning (Cottenie 2005, Peres-Neto et al. 2006, Soininen 2014, Leibold and Chase 2017).

However, estimating SERs while accounting for spatial autocorrelation and performing variation partitioning have been the subject of different research fields in which different methods are typically used. In biogeography and macroecology, generalised 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), whereas in community ecology multivariate methods, in particular constrained ordination or distance-based methods, have been widely used to infer community assembly processes underlying environmental and spatial effects (Cottenie 2005, Peres-Neto et al. 2006, Tuomisto and Ruokolainen 2006, Soininen 2014). Constrained ordination such as CCA and RDA became popular due to their ease of use, and have been widely used to partition the explained variation according to environmental and spatial effects. However, these methods have been criticised, 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 disregarded 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 thorough comparisons of different models to estimate species distributions have been made (see Norberg et al., 2019, for a comprehensive review and comparison of methods), machine learning has not yet been widely used in community ecology, and as far as we are aware, never for variation partitioning in site-by-species community matrices. Here, we compare ten methods grouped in four families (Table 1) in their ability to recover true environmental and spatial effects and respective fractions of explained variation from simulated spatially explicit site-by-species matrices. We also compare these methods using nine empirical datasets.

Material and methods

Data simulation

In order to compare the different methods, we used 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. Thus, we knew the "true" environmental and spatial fraction of variation in abundance (or occurrence) in each site-by-species matrix, which we then compared with fractions estimated by the different methods in Table 1. We used a simple simulation of multi-species abundances (or occurrences) consisting of a grid of N cells, where each cell was occupied by a different community and was environmentally homogeneous. The occurrence or abundance ($\bar{Y}_{ij}$) of each species i in a given cell j depended on an environmental component (X_{ij}), a spatial component (W_{ij}), or their linear combination (i.e. their additive affect; for the variation partitioning analysis).

Environmental component (X)

We simulated one spatially autocorrelated environmental variable E ($0 \leq E \leq 1$) consisting of a random Gaussian field in which the autocorrelation level was set by the range parameter (A)

of an exponential variogram model, using the R package *gstat* (Pebesma 2004). Species responded to E either linearly or as a Gaussian curve (i.e. a unimodal 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 uniform distribution of β values ranging from 0 to 20). 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 (see examples of linear and Gaussian-shaped responses in Fig. 1, and an example of the spatial pattern of X in Appendix S1, Fig. S1.1).

Spatial component (W)

The spatial component W was defined by simulating an independent random Gaussian field S with a given autocorrelation range (A), with the abundance (or occurrence) W_{ij} directly proportional to S_{ij} ($W_{ij} \propto S_{ij}$). We forced the frequency distribution of W_i to be exactly the same as that of X_i by reordering X_i according to S_i. Therefore, W_i and X_i differed only in their spatial structure (see an example of the spatial pattern of W in Appendix S1, Fig. S1.1).

Multi-species responses ($\bar{Y}_i$)

In order to control how much the environment (X) and space (W) drive the variation of abundance (or occurrence) in a site-by-species community data, 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_i were obtained by first transforming $\bar{Y}_i$ (centred around 0 to make the binomial distribution balanced) into probabilities using an inverse logit function, and then randomly sampling from a binomial distribution. Abundance data Y_i were obtained by adding normally distributed errors (with mean of 0 and standard deviation of 10% of the maximum abundance) to $\bar{Y}_i$, or drawn randomly from a Poisson distribution using the predicted abundances $\bar{Y}_i$ as the mean abundances of the Poisson distribution (to simulate counts data; see an example of the spatial pattern of the resulting Y in Appendix S1, Fig. S1.1). Similarly to standardised coefficients in linear regression, the weights β set the variation in species abundances attributable to the environmental and spatial components. Because X and W have exactly the same frequency distribution and are independent, β is closely related to the partial correlation r between $\beta_X X_i$ or $\beta_W W_i$ and $\bar{Y}_i$, i.e., $\beta_X \approx r_{X\bar{Y}}$ and $\beta_W \approx r_{W\bar{Y}}$. However, the measure of variation that is typically used for variation partitioning is the coefficient of determination R^2, thus we used the r^2 (equivalent to R^2 in a linear regression) as the true R^2. We performed variation partitioning on the r^2 through a sequence of calculations (Borcard et al. 1992) where $[X]$ is the pure fraction of variation attributable to X, $[W]$ is the pure fraction of variation attributable to W, and $[X \cap W]$ is the shared fraction of variation attributable to the collinear effect of X and W:

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

$$r_X^2 = r(Y_i, X_i)^2,$$

$$r_W^2 = r(Y_i, W_i)^2,$$

$$[X] = r_{XW}^2 - r_W^2,$$

$$[W] = r_{XW}^2 - r_X^2,$$

$$[X \cap W] = r_X^2 + r_W^2 - r_{XW}^2.$$

Note, however, that the spatial predictors in the statistical models we consider (either MEMs or spatial coordinates) fit the abundance/occurrence Y determined by the spatial structure of both the spatial variable S and the environmental variable E, but the spatial structure of X is independent of W (other than spurious correlation). As such, the estimated shared fraction according to these methods is systematically higher than the "true" shared fraction. Hence, we used the relative contribution of the pure fraction $[X]$ (or $[W]$) such that $[X]_{rel} = [X]/([X]+[W])$, which can be directly compared to the estimated relative fractions. As expected, the true relative fraction $[X]_{rel}$ is closely related to β_X (Appendix S1 Fig. S1.2).

Statistical methods

We considered four families of statistical methods: distance-based regression, constrained ordination, generalised linear and additive models, and tree-based machine learning (Table 1). These methods can fit 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 variants within methods would result in a huge list, we opted for using the most popular and widely used approaches. Our goal was to compare the performance of different strategies (or "pipelines") for partitioning variation in site-by-species matrices, in that the methods considered vary both in the statistical model and in the specification of the model (i.e. in the type of transformation of the response and in the model independent terms).

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 site-by-species 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 species' abundance or occurrence. Still, since this method is often considered suitable for disentangling environmental and spatial sources of variation (see Tuomisto & Ruokolainen, 2006, and the discussion initiated with Legendre et al. 2005 followed by e.g., Tuomisto and Ruokolainen 2008, Laliberté 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 both the environment and the spatial coordinates (i.e., "geographical" distance). 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), and Redundancy Analysis (RDA; van den Wollenberg 1977) and its variant, the distance-based RDA (dbRDA; Legendre and Anderson 1999). 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). When using RDA, the species abundance data were Hellinger-transformed (as recommended in Legendre & Gallagher 2001 and Legendre and Legendre 2012). In addition, when the response to the environment was Gaussian-shaped, we first log-transformed the abundance data to improve model fitting (Fig. 1b, red line). 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 site-by-species matrix but a matrix composed of principal components resulting from a principal coordinates analysis (PCoA) of the distance matrix. The environmental effect was modelled as linear in CCA and dbRDA, and as either a linear or a second-degree polynomial in RDA, depending on the simulated response. The spatial effects were modelled with distance-based MEM variables using the function "dbmem" (with default settings) of the R package *adespatial* (Dray et al. 2016). 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 and additive models

We estimated (i) generalised linear models (GLM) using the *stats* R package (R Development Core Team 2017) and (ii) generalised additive models (GAM) using the R package *mgcv* (Wood 2017). For the two methods, we fitted a second-degree polynomial for the environmental variables, and the spatial effects were modelled with MEMs in GLM (the same as in constrained ordination) and with thin plate regression splines on the spatial coordinates of the sites in GAM. All models were fitted as single models to each species and the predictions were then stacked into a matrix of predicted species abundances. We used a quasi-Poisson error distribution in GLM for abundance data, as a quasi-Poisson distribution had better estimation convergence. We also tried to use GLM with Poisson and negative binomial error distributions for abundance data, but these models had more 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 (with the exception of GAM if splines are used for every predictor); instead, the relationship is learned entirely from the data. As such, the environmental predictor was simply *E* and the spatial predictors were simply the spatial coordinates (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*). W fitted a tree with the minimum number of observations in any terminal ("leaf") node fixed to 5.

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 sampled 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 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. 2020). 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.

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 two variants of the algorithm – two tree depths (i.e. interaction depths). 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.

Comparison of method performance on simulated data

Our study comprises two exercises. In Exercise 1 we assessed the flexibility of the methods in Table 1 to approximate the simulated linear or Gaussian-shaped responses of species to the environment, and their responses to spatial gradients which are typically more complex and non-linear. In Exercise 2 we partitioned the variation in abundance/occurrence caused by the spatial and environmental component, and compared these to the true spatial and environmental variation fractions.

Exercise 1. Model fitting performance

For the first exercise, the simulated occurrence or abundance of species ($\bar{Y}_{ij}$) was determined by the environment alone ($\bar{Y}_{ij} = X_{ij}$ and $W_{ij}, = 0$) or by the spatial component alone ($\bar{Y}_{ij} = W_{ij}$ and $X_{ij} = 0$), and we only considered univariate single-species data instead of a multi-species site-by-species data. The goal of this exercise was to test if the major types of strategies are flexible enough to fit the different types of environmental responses and spatial structures considered here (Appendix S1, Table S1.1). As such, we examined (G)LM, BRT, and UniRF for the environment, and the same methods plus GAM for space (see model specification in Table 1) – note that we did not use GAM to fit the environmental effects because we modelled it as a quadratic effect, similarly to GLM. To assess the performance of the different methods, we calculated the Pearson correlation between the simulated abundance (or occurrence) values with the values predicted by the method. We used $\bar{Y}$, not Y, as the truth. To assess their predictive ability and propensity to overfitting, we also performed leave-one-out cross validation by predicting successively for the data points left out from model estimation.

Exercise 2. Variation partitioning performance

For the second exercise, we considered multivariate responses of multiple species in site-by-species matrices where $\bar{Y}_{ij}$ was determined by a linear combination of X_{ij} and W_{ij} (equation 3).

We then fitted the models from Table 1 to these data (using environmental and spatial predictors as specified in the Table) and performed variation partitioning based on the models' R^2 (or pseudo R^2, Table 1) to obtain the relative fraction of variation explained by the environment $[\hat{X}]$, which is an estimate of the true simulated fraction $[X]$. Sometimes either $[X]$ or $[W]$ were close to 0, and thus $[\hat{X}]$ or $[\widehat{W}]$ could be slightly negative. In these cases, we set $[\hat{X}]$ or $[\widehat{W}]$ to 0. In addition, because R^2 metrics can be inflated by overfitting, we also calculated cross-validated fractions of variation. We performed leave-one-out cross-validation by predicting successively for the data points left out from the training data and then calculating the respective R^2 using these predictions. Note, however, that we did not perform cross-validation in a few methods, because the package does not retrieve the predictions at the same scale as the response (in CCA and dbRDA) or because it cannot be applied to the specific model (in MRM). For each method, we preferably chose the R^2 or pseudo-R^2 metric that is implemented in the respective R package, but given that no consensus exists on what metric to use in some methods, we also compared alternative metrics. Their influence on the performance of the different methods was negligible (see Results). The R^2 metrics included an eigenvalue-based metric (R2-mv; used for ordination methods as retrieved by the R package *vegan*), 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) (see the metrics used for specific methods in Table 1 and their definitions in Appendix S1).

To assess the sensitivity of the different methods to variations in the simulated data, we varied the type of data (normal, counts, or binary), the size of the grid (N = 25, 100, or 400), the autocorrelation range (A = 5, 25, or 50), and the type of response to the environment, either a linear response or a Gaussian-shaped response with varying niche breadth ($2\sigma^2$ = 0.002, 0.02, or 0.2) (see also Appendix S1, Table S1.2). The reference scenario with fixed parameter values when the target parameter varied was N = 100, R = 25 and $2\sigma^2$ = 0.02. In addition, because undersampling is common in ecology, we also assessed the

performance by fitting the models to a random subsample (50 out of 400 grid cells). For each simulation parameter combination, a range of predictive weights determining the true trend in the variation fractions (β_X and β_W; $\beta_X+\beta_W=1$) were simulated, from 0 to 1 increasing by 0.1, resulting in variation of the relative importance of environment and space. Each combination of parameters was replicated 5 times (resulting in a total of 1485 runs). The performance of each method was assessed by calculating the RMSE between the true $[X]$ and estimated $[\hat{X}]$ fractions of variation (the lower the better) across the whole range of β_X, including replicates.

Comparison of methods on empirical data

To explore the impacts of method choice using empirical data, we compared the all the methods in Table 1 by performing variation partitioning on nine empirical datasets (see Appendix S2 for the respective references). GLM, GAM and MVRT had fitting issues for many datasets and were excluded from the analysis. Each empirical dataset consisted of a species abundance matrix, some environmental variables and the geographical coordinates of the sampled sites.

Results

Exercise 1. Model fitting performance

We found some differences in the fitting performance (measured by the Pearson correlation between the predicted and the simulated response data) of the major types of statistical models (Fig. 2). For the environmental effects, BRT followed by RF performed best. GLM often produced worse fits when using normal and binary data, which was mainly due to the inability of linear models (in particular GLM with normal error distribution) to fit Gaussian-shaped responses, even when the response variable was log-transformed and a quadratic effect was modelled (see Fig. 1 for an example). However, GLM with a Poisson error

distribution using count data performed similarly to BRT and RF (see an example of a fit in Fig. 1). All methods were in general robust to overfitting when modelling the environmental effect, and no obvious differences in performance were found among the different models (Fig. 2). For the spatial effects, the overall performance dropped in all models, meaning that the fits for the spatial effect were worse than those for the environmental effect. GLM with MEMs as predictors were generally better than GAM and BRT. However, the predictive ability as measured by cross-validated predictions was to some extent lower than the fitting ability (i.e. without cross-validation) in all models, but especially in GLM. Overall, and although not perfectly, all models seem appropriate to fit the generated data, though we observed a trade-off in the fitting performance for environmental and spatial effects – GLM with MEMs performing better for approximating the effect of space but tree-based methods were in general better for the effect of the environment.

Exercise 2. Variation partitioning performance

The partitioning performance, i.e. the difference between the true and estimated fractions of environmental variation (measured by RMSE), of the different methods varied considerably but was overall consistent among the different types of data (Fig. 3 and 4; Appendix S1, Fig. S1.4, for all tested variants of the different methods). GAMs and tree-based machine-learning performed best across a wide range of scenarios for both binary and abundance data. The remaining methods performed worse, especially the ordination methods. Cross-validated variation partitioning followed the same performance trends in general when comparing the different methods.

The good performance of machine learning was generally consistent among its different variants, though BRT consistently outperformed other ML techniques for abundance, and particularly for the binary data (Fig. 3). An interaction depth of 3 compared to 1 in the BRT algorithm was better for binary data, but it did not make much difference for

abundance data (Appendix S1, Fig. S1.4). The total variation explained by the BRT models was higher compared to other methods (Appendix S1, Fig. S1.5-7), but cross-validation indicated that overfitting was not an issue (Fig. 3). For UniRF, the default sample size parameter of the algorithm (equal to the sample size of the data) worked better in general than a lower sample size (20% of the data sample size), except for linear responses to the environment, for which the lower sample size was better. Multivariate random forest (MVRF) and regression trees (MVRT) had somewhat worse performances than BRT and UniRF.

Ordination methods (CCA, RDA, dbRDA) performed consistently worse than other methods for all types of data (Fig. 3). RDA was in general better than CCA and dbRDA, even for non-linear responses to the environment. Consistently with the results from the first exercise, we observed that they generally underestimated the environmental fraction (Fig. 4), due to their worse fit to non-linear responses, and a tendency of MEMs to fit the spatial structure resulting from "unfitted" environmental effects.

GAMs, together with BRT, were in general the most reliable methods for the variation partitioning (Fig. 3 and 4). In contrast, GLMs had mediocre performances for abundance data and performed poorly for binary data. Further, GLMs sometimes failed to converge due to the large number of MEMs relative to sample size. Although GLMs can better fit environmental effects than the linear models underlying RDA, the low performance can again be due to the unbalanced fitting performance of environmental vs. spatial effects, even if the R^2 values are adjusted to correct for the number of predictors.

Finally, MRM (the distance-based regression) had mediocre performance similar to RDA and GLM, depending on the type of data (Fig. 3 and 4). MRM generally overestimated the environmental fraction (Fig. 4), likely because spatial distance is a limited proxy for spatial structure.

The type of R^2 metric had a minor impact on the results compared to the methods, as observed by the clustering of methods in the performance rankings (Appendix S1, Fig. S1.4).

In general, based on our results, it is hard to highlight specific R^2 metrics for the different methods.

Empirical data

When applied to empirical data, the choice of method had a clear influence on the results of the variation partitioning. GAM, GLM, and MVRT failed to fit for some of data sets, likely because of the large number of predictors (including multiple environmental predictors) relative to sample size. The variation partitioning results were remarkably different among the methods. Even closely related methods, such as dbRDA and RDA, and BRT and UniRF, provided different results (Fig. 5). These results are, nevertheless, consistent with the simulation exercises. The lower environmental fractions estimated by RDA compared to dbRDA (Fig. 5) can also be observed in the simulation results (Fig. 4); and the relatively lower environmental fraction estimated by UniRF compared to BRT (Fig. 5) can also be observed in the simulation results, especially when the responses are linear (Fig. 4).

Discussion

As far as our simulations could tell, the best performing methods were tree-based machine learning methods, in particular BRT, and GAM. This highlights the potential of flexible, yet not widely employed, methods for ecologists aiming at partitioning environment and space in site-by-species matrices, one of the most common types of data in ecology.

Tree-based ML is by far the most user friendly approach from our list: it can fit both narrow and wide responses to the environmental variables (i.e. narrow vs. wide niches), and inherently accounts for interactions in their effects, while simultaneously modelling complex spatial responses. Here we stress that modelling spatial effects in tree-based ML is as simple as inserting the spatial coordinates in the function call. Even though ML has also been criticised for a tendency to overfit (e.g. Wenger and Olden 2012, but see Hastie et al. 2009),

this was not an issue in our analyses, as in our results the cross-validated variation fractions performed as well as the regular ones, with only minor differences for both GAMs and ML. In any case, one can always remedy concerns about overfitting by using cross-validation. We also highlight that tree-based methods can also be used for interpreting modelled relationships and test for effects (Lucas 2020), although directly testing *a priori* hypotheses about species-environment relationships is not possible.

When specific hypotheses about species-environment relationships are considered, it may be instrumental to use parametric models rather than learning algorithms. For example, if we want to test for linear or unimodal (e.g. 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 unimodal responses. Our results indicate that GAMs offer a good compromise, as modelled species-environment relationships can be limited to parametric response shapes (or to splines with a limited number of knots) while spatial effects can be fitted using flexible smooth splines on spatial coordinates. However, sample size is a limitation for the use of GAMs with "free" splines due to the large number of model parameters. When sample size is sufficiently large, though, splines in GAM have the advantage that overfitting is inherently penalised (Wood 2017).

Strikingly, the widely used constrained ordinations (CCA, dbRDA, and RDA), as well as GLMs, underperformed in estimating the environmental vs. spatial fractions of variation. We attribute underperformance in ordination methods to the general underestimation of the environmental component, arguably because of model misspecification (as in Fig 1b – note that constrained ordination models are based on LMs). However, we also observed poor performance when the environmental effect was correctly specified (i.e., linear responses in RDA, as well as linear or unimodal responses in GLM), which suggests that MEMs might also overfit despite the use of adjusted R^2 metrics (in line with Gilbert and Bennet 2010).

Finally, the distance-based regression (MRM) also performed poorly, and we recommend to avoid this method in variation partitioning exercises.

While we have pointed to some better, and some worse, methods for disentangling spatially vs. environmentally driven variation, none of the methods was perfect, and thus there is margin for improvement. Machine learning methods are diverse, and more methods can be tested and even developed specifically for the purposes outlined here (e.g. D'Amen et al. 2017, Nieto-Lugilde et al. 2018). Also, joint species distribution models (JSDM) are becoming popular in community ecology (e.g. Ovaskainen et al. 2017), but these methods are based on generalised linear modelling which also brings its share of limitations. 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 avenues for improvement might include the use of other spatial regression techniques such as those based on Moran's Eigenvectors (Dray et al. 2012, Murakami and Grifith 2019), as well as autoregressive models (e.g. Dormman et al. 2007), though we are not aware of the use of the latter to perform variation partitioning.

In conclusion, we showed that the ordination methods CCA and RDA, which are widely used to partition variation explained by environment and space, provide less accurate estimates of variation fractions under many circumstances. As the particular shapes of species responses to the environment are hard to hypothesise in the context of community or multi-species data, we recommend tree-based machine learning as a flexible alternative that can be widely used with both abundance and occurrence data. If *a priori* hypotheses about species-environment relationships are considered, GAM as a semiparametric method is a good choice for variation partitioning. 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 apply to single-species niche models, community ecology, and macroecology studies.

Funding

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.

Author contributions

The first author performed simulations of metacommunities, consolidated and applied the 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.

Data Accessibility

All data are publicly available (see Appendix S2). All the analyses were carried out in R (R Development Core Team, 2020), and all the computer code is available in https://github.com/duarte-viana/iVarPart.

References

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.

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.

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.

D'Amen, M. et al. 2018. Disentangling biotic interactions, environmental filters, and dispersal limitation as drivers of species co-occurrence. - Ecography 41: 1233–1244.

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 30: 609–628.

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.

Harris, D. J. 2015. Generating realistic assemblages with a joint species distribution model. - Methods Ecol. Evol. 6: 465–473.

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.

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.

Lucas, T. C. D. 2020. A translucent box: interpretable machine learning in ecology. - Ecol. Monogr. *in press*: ecm.1422.

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

Murakami, D. and Griffith, D. A. 2019. Eigenvector Spatial Filtering for Large Data Sets: Fixed and Random Effects Approaches. - Geogr. Anal. 51: 23–49.

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.

Pebesma, E. J. 2004. Multivariable geostatistics in S: The gstat package. - Comput. Geosci. 30: 683–691.

Peres-Neto, P. R. et al. 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. - 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.

Thompson, P. L. et al. 2020. A process-based metacommunity framework linking local and regional scale community ecology. - Ecol. Lett. 23: 1314–1329.

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.

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.

Table 1. 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 settings in R**
Constrained ordination	RDA	Redundancy analysis	P	**R2-mv**	**R2-mv**	Linear, polynomial	MEM	- R package: *base, vegan* - Species data: Hellinger-transformed
	dbRDA	Distance-based redundancy 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**	Linear	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, polynomial	MEM	- R package: *base* - Species data: raw - Distribution: quasi-Poisson (abundance), binomial (binary)
	GAM	Generalized additive model	NP	R2-cla, R2-mv, **R2-McF**	**R2-Tju**, R2-Efr, R2-McF	Linear, polynomial (potentially any)	Splines on XY coordinates	- R package: *mgvc* - Species data: raw - 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	XY coordinates	- R package: *mvtboost* - Species data: raw - Distribution: Poisson (abundance), Bernoulli (binary)

							- #trees: 1000 - Learning rate: 0.01 - Interaction depth: 1 or 3 - Internal CV: no
MVRT	Multivariate regression tree	NP	**R2-cla**, R2-mv, R2-McF	**R2-Tju**, R2-Efr, R2-McF	Any	XY coordinates	- R package: *mvpart* - Species data: raw (abundance), numeric 0/1 (binary) - Node size: 5 (no CV) - Interaction depth: 1-20 (CV) - Internal CV: no
UniRF	Univariate random forest	NP	**R2-cla**, R2-mv, R2-McF	**R2-Tju**, R2-Efr, R2-McF	Any	XY coordinates	- R package: *randomForest* - Species data: raw (abundance), categorical 0/1 (binary) - #trees: 500 - Sample size: N or N/5
MVRF	Multivariate random forest	NP	**R2-cla**, R2-mv, R2-McF	**R2-Tju**, R2-Efr, R2-McF	Any	XY coordinates	- R package: *randomForestSRC* - Species data: raw (abundance), categorical 0/1 (binary) - #trees: 500 - Node size: 5

*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

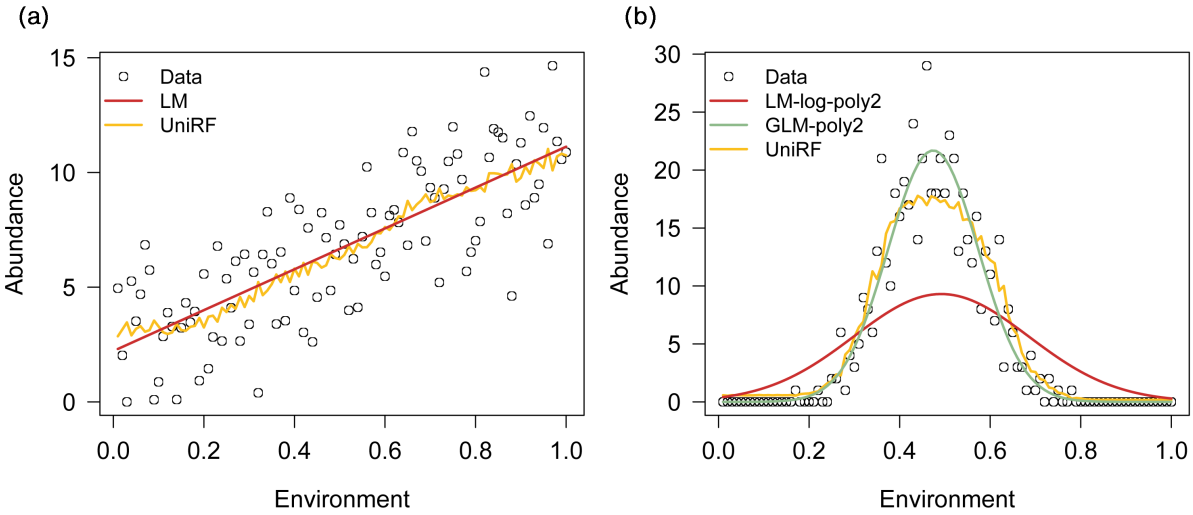


Figure 1. Example of different models fitted to a linear response (a) and a Gaussian-shaped response (b) of one species. For the linear response, a simple linear model (LM) and univariate random forest (UniRF) are fitted. For the Gaussian-shaped response (counts), a LM with log-transformed response data and a quadratic effect (LM-log-poly2), a Poisson generalised linear model (GLM) with a quadratic effect, and UniRF are fitted.

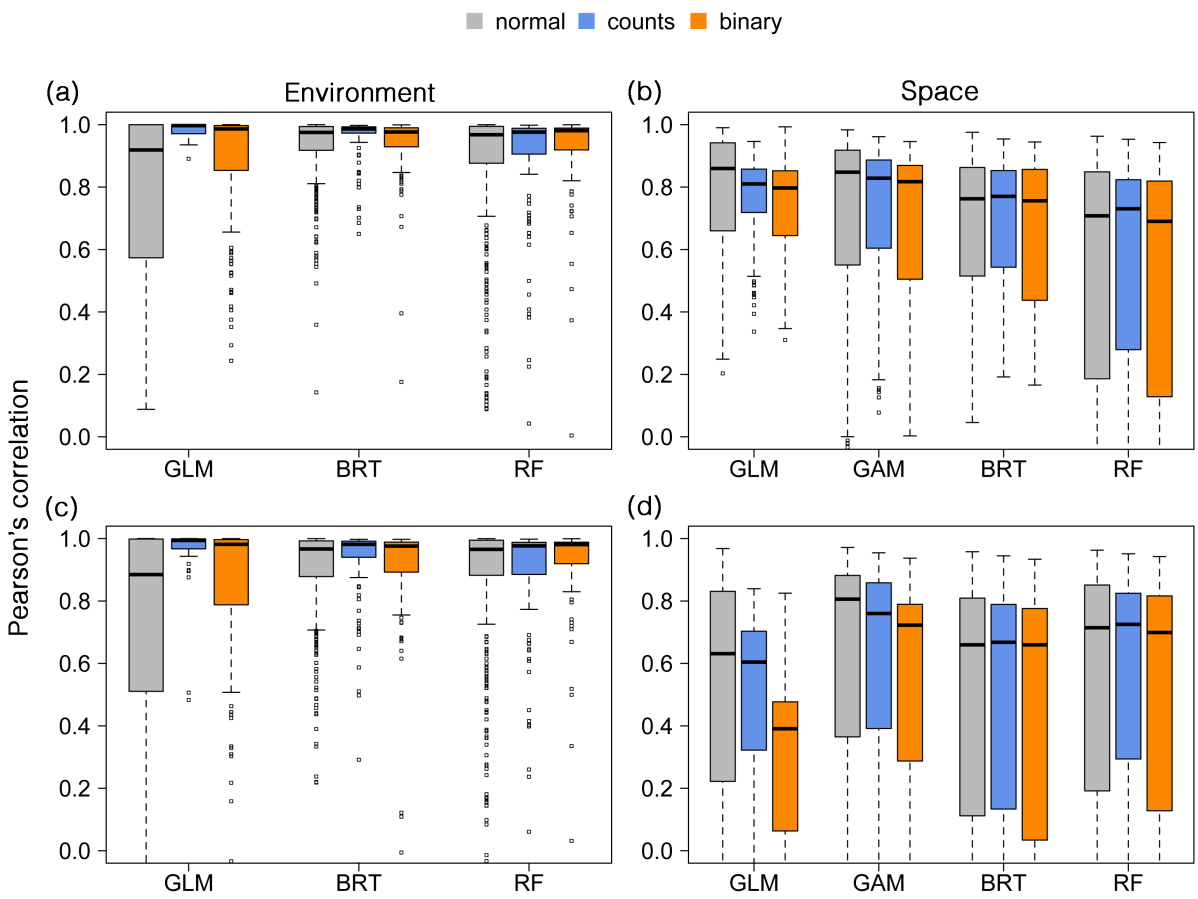


Figure 2. Fitting performance of the major types of statistical models for either environmental effects (left panels) or spatial effects (right panels), fitted to normal, counts, and binary data, across the different generated data (see Appendix S1, Table S1.1), using raw (upper panels) or cross-validated (lower panels) predictions. Note that GAM was fit to the environment similarly to GLM (and not using splines), thus the results are similar to GLM. The performance is given by the Pearson's correlation between the simulated "true" values ($\bar{Y}$) and predictions (Y_{pred}) from the tested methods.

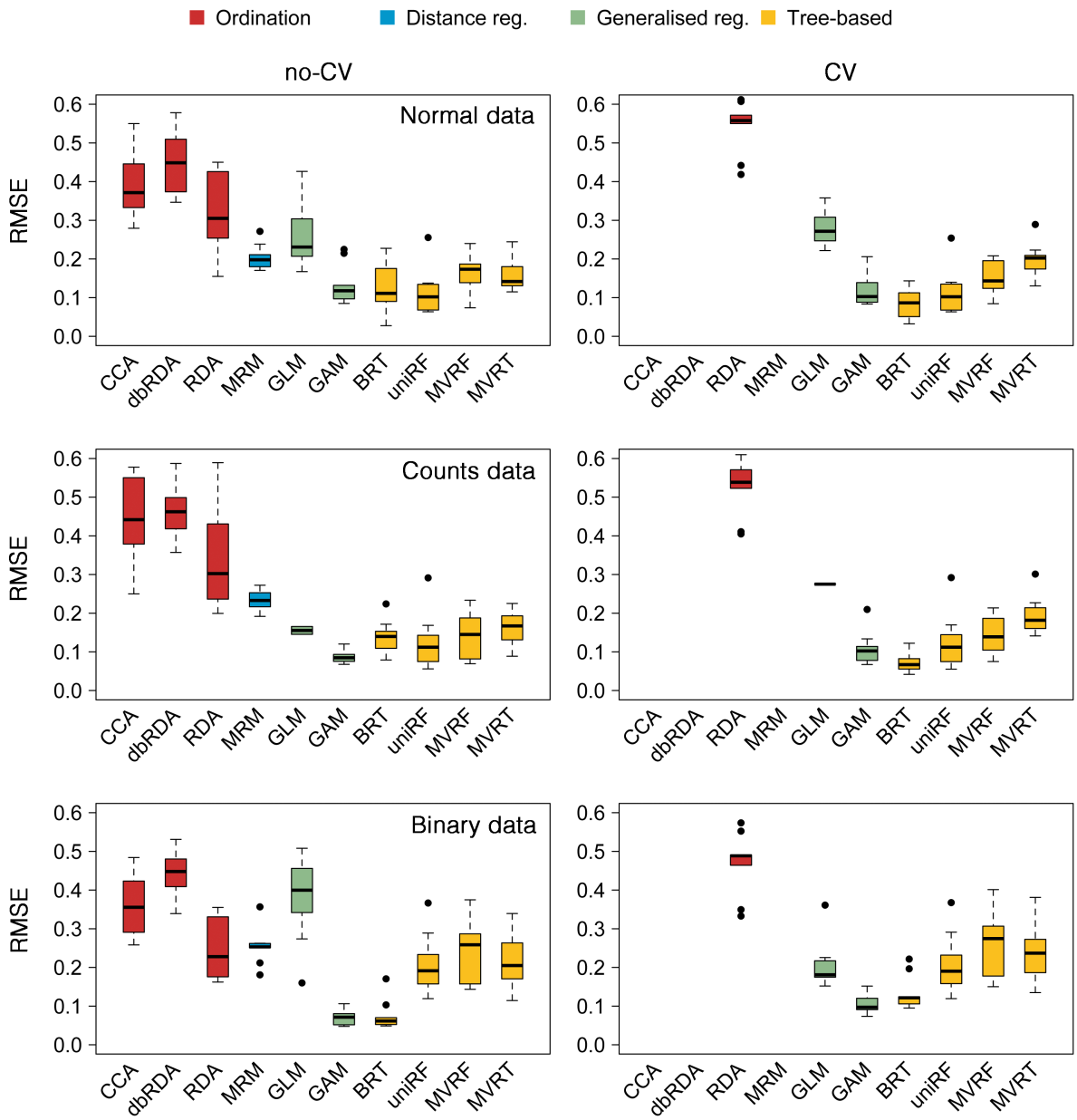


Figure 3. Variation partitioning performance of the different methods for normal, counts, and binary data across the different generated data using raw variation partitioning (left panels) or cross-validated variation partitioning (right panels). The performance is given by the RMSE between the true and estimated fractions of variation. Performance for other variants of each method is in Appendix 1, Fig. S1.4.

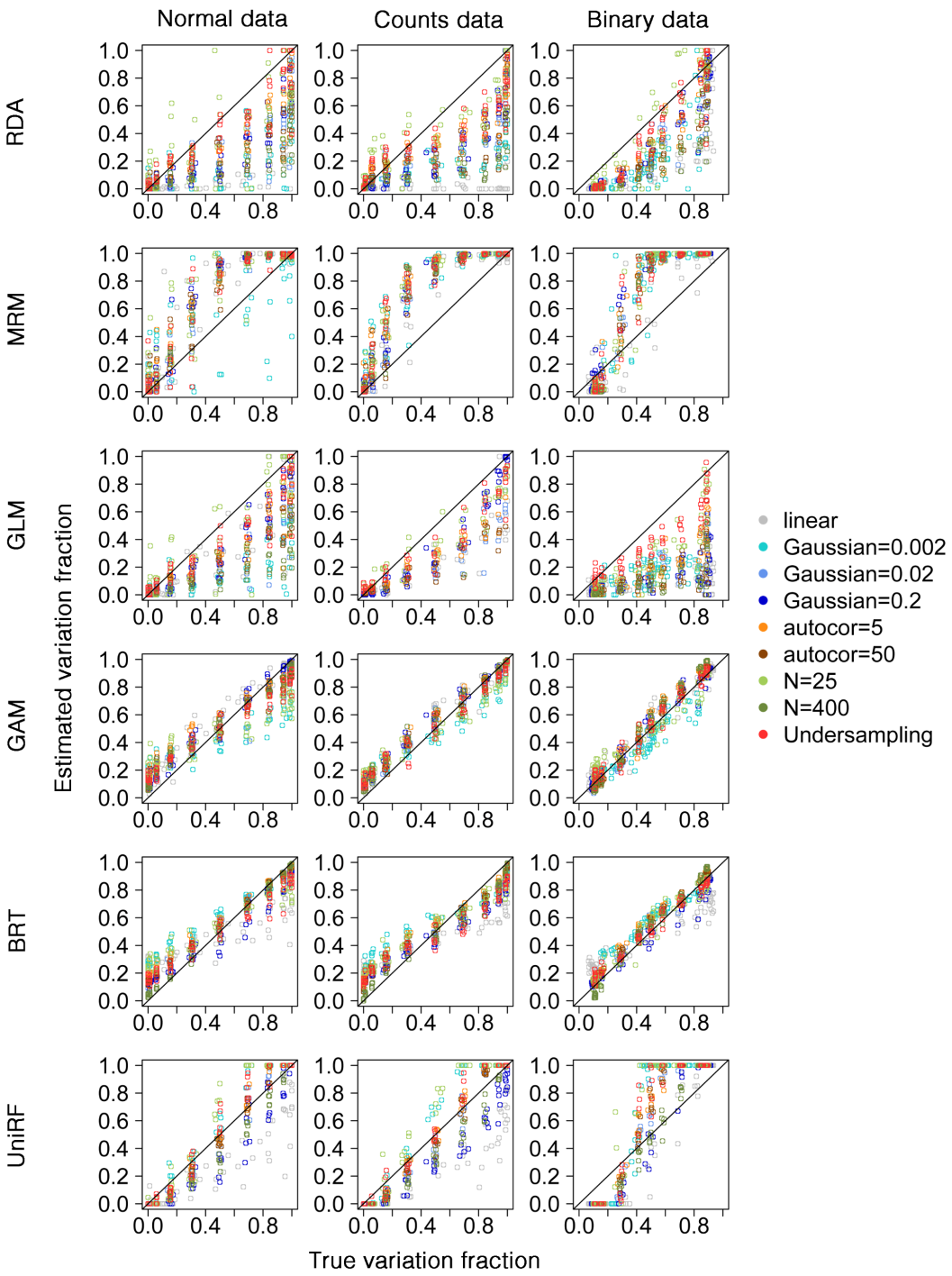


Figure 4. Plots of true vs. estimated fractions of variation ($\hat{R}^2_{X_{rel}}$ vs. $R^2_{X_{rel}}$) for normal, counts, and binary data across different methods (different rows) and generated data (different colours). The black line represents the 1:1 line, where points should fall if the performance of the method had been perfect.

Figure 5

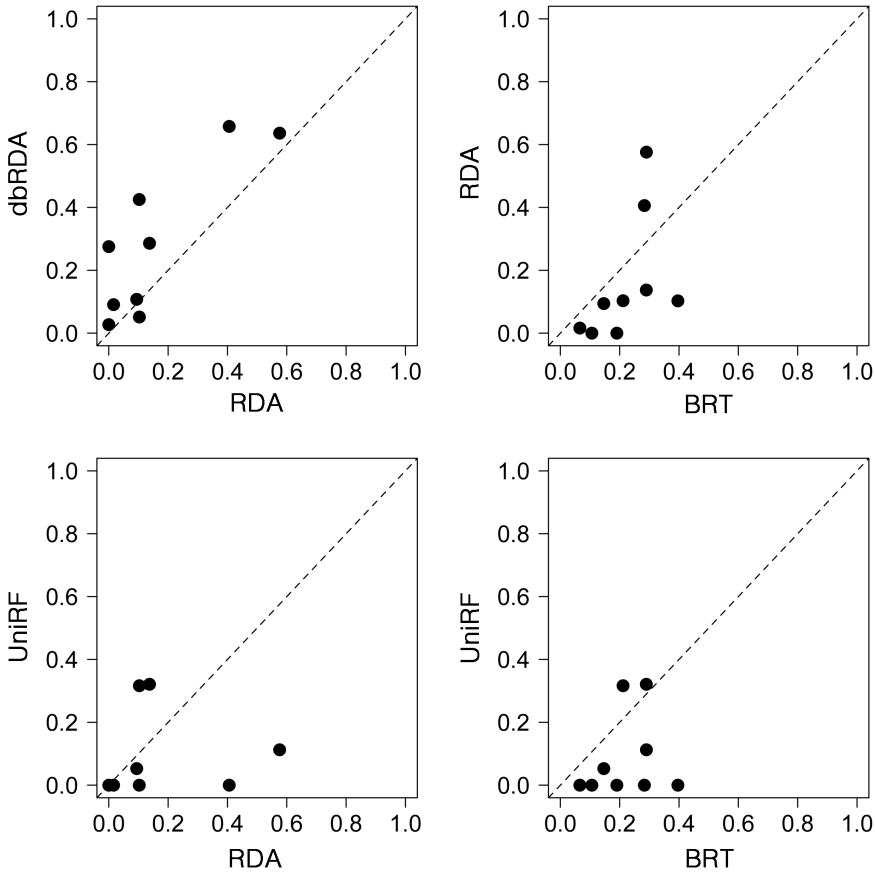
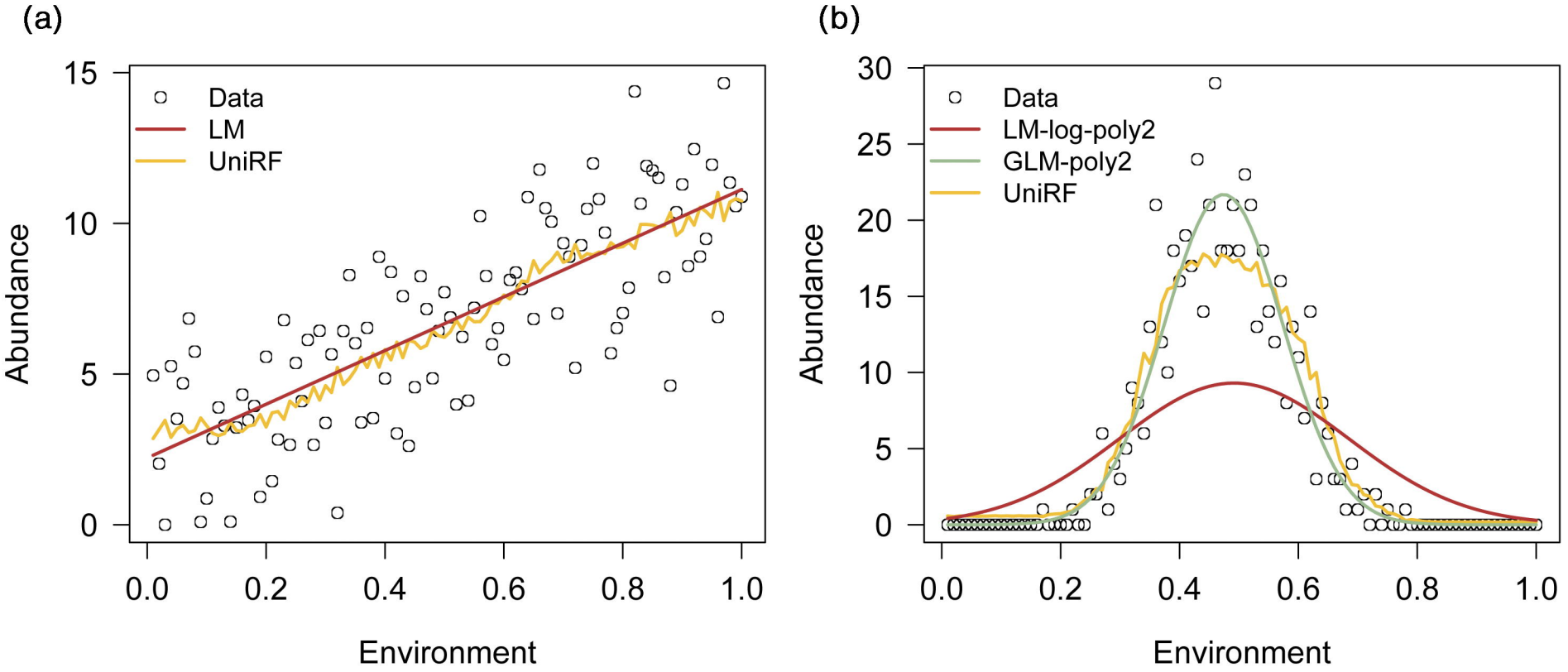
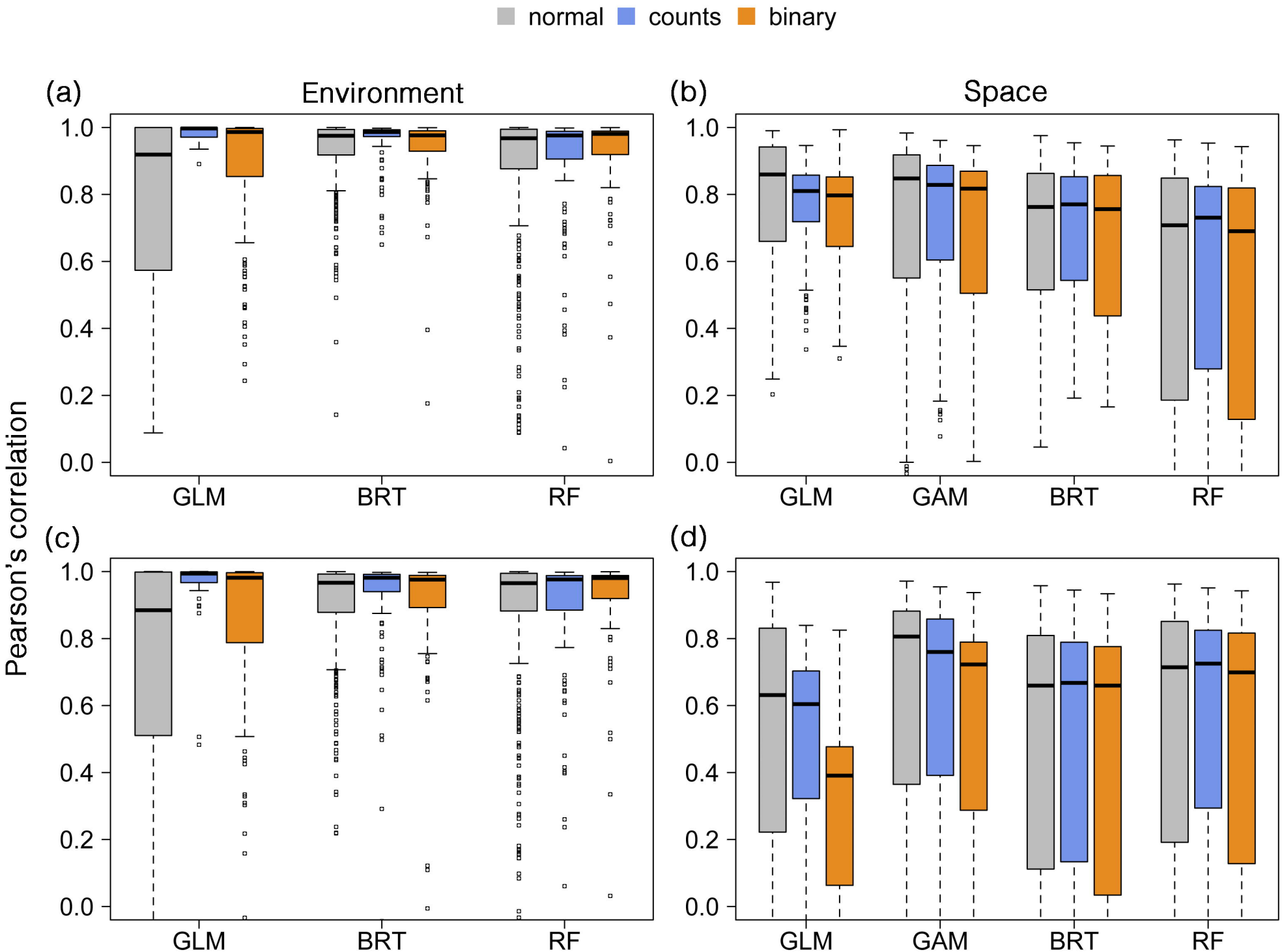
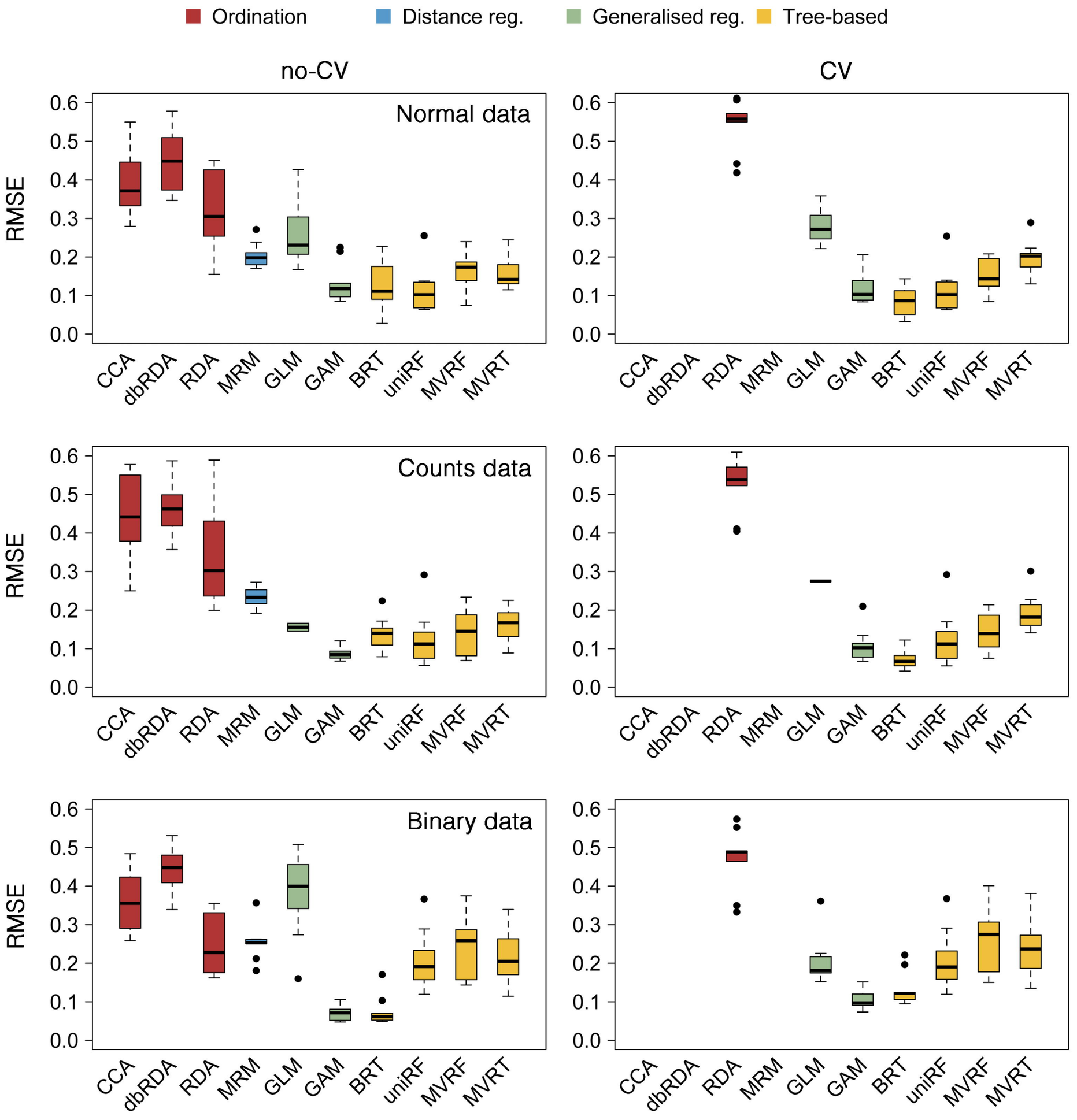
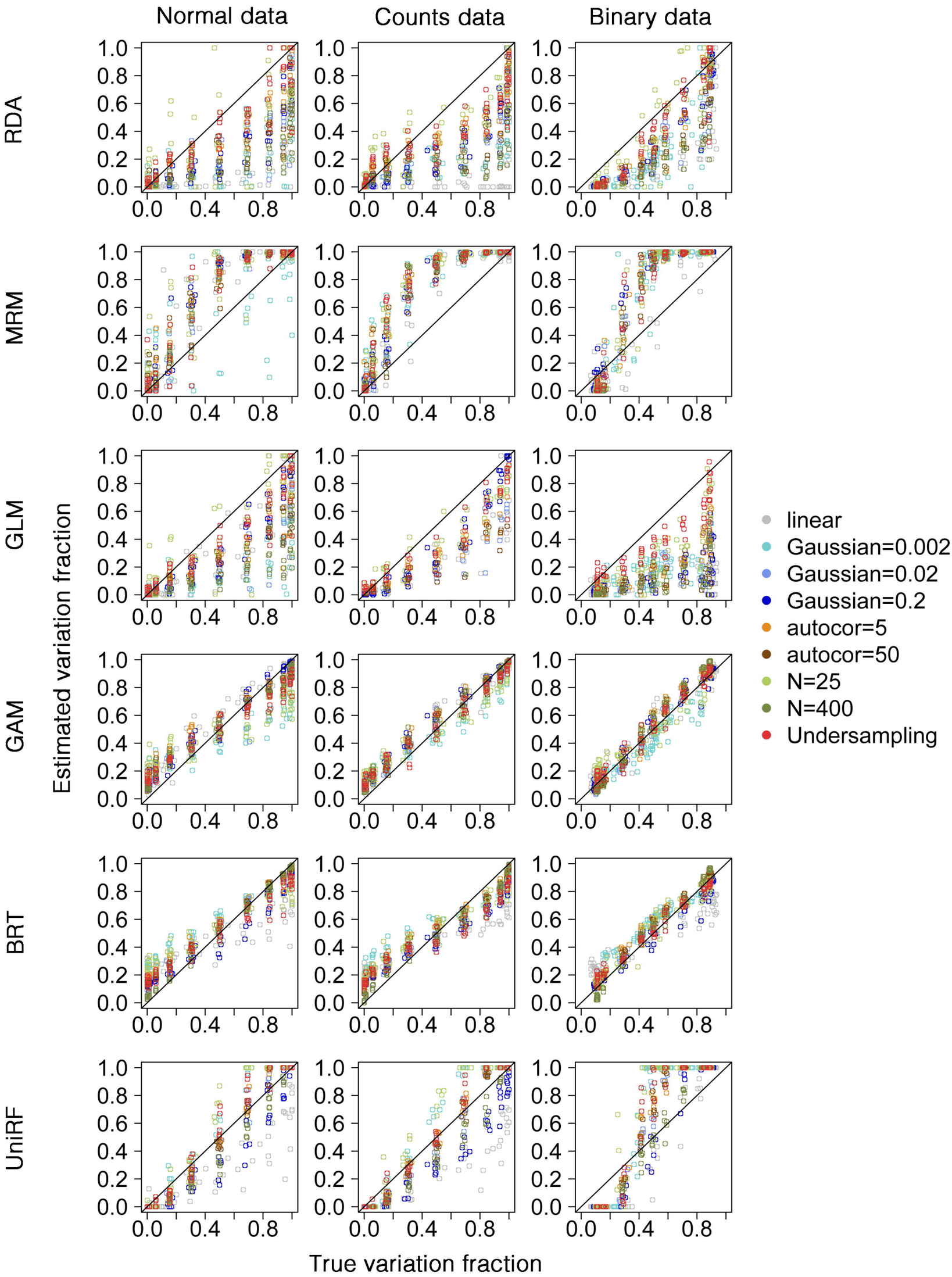


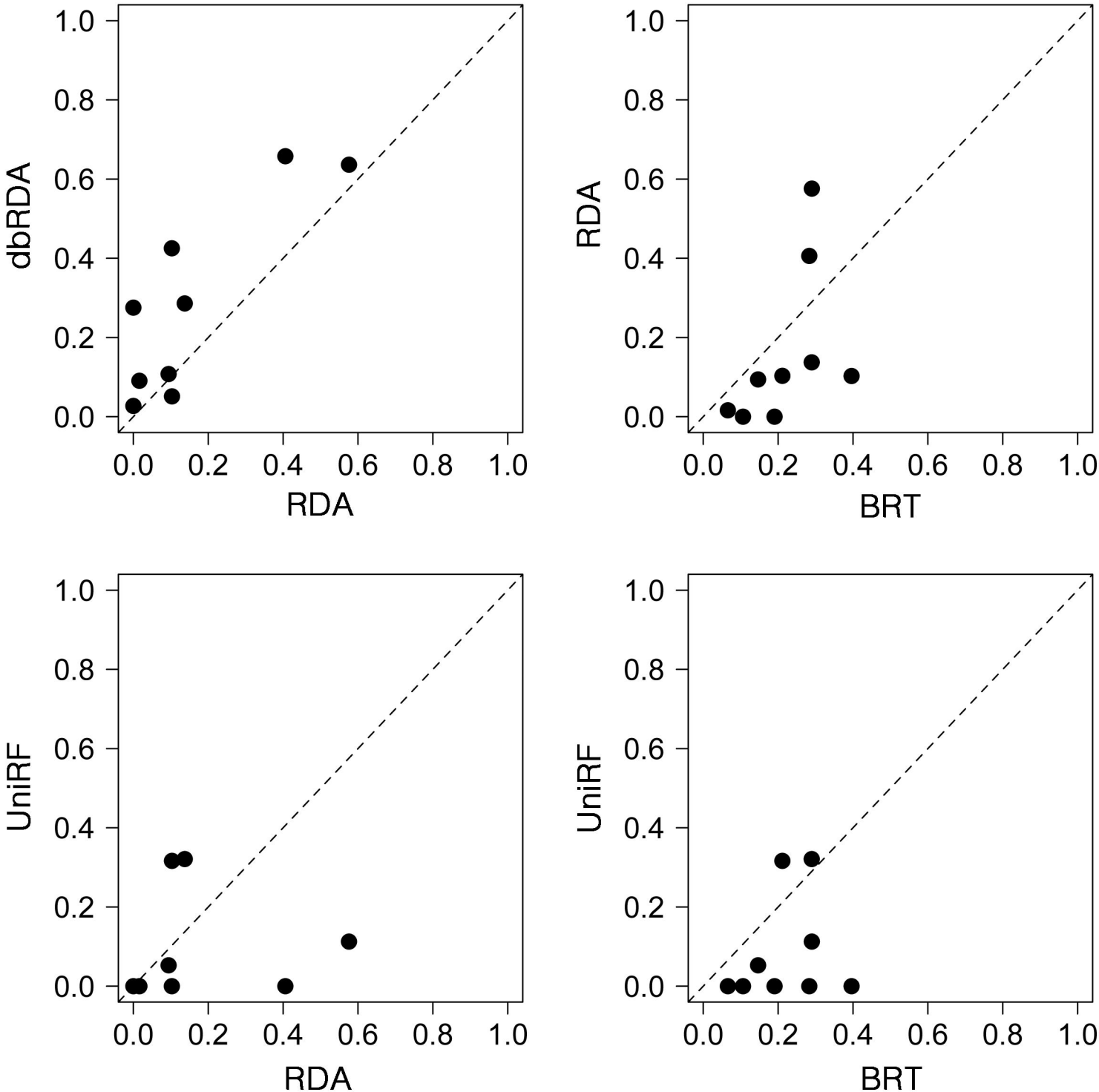
Figure 5. Comparison of the estimated variation fraction among canonical methods (RDA and dbRDA) and tree-based methods (BRT and UniRF) for each empirical dataset. The dashed line represents the 1:1 line. See Appendix S2 for references.


(a)
Data
LM
UniRF
Abundance
Environment
0
5
10
15
0.0
0.2
0.4
0.6
0.8
1.0
(b)
Data
LM-log-poly2
GLM-poly2
UniRF
Abundance
Environment
0
5
10
15
20
25
30
0.0
0.2
0.4
0.6
0.8
1.0


normal
counts
binary
(a)
Environment
(b)
Space
(c)
(d)
Pearson's correlation
1.0
0.8
0.6
0.4
0.2
0.0
GLM
GAM
BRT
RF


Ordination
Distance reg.
Generalised reg.
Tree-based
no-CV
CV
Normal data
Counts data
Binary data
RMSE
0.0
0.1
0.2
0.3
0.4
0.5
0.6
CCA
dbRDA
RDA
MRM
GLM
GAM
BRT
uniRF
MVRF
MVRT


Normal data
Counts data
Binary data
RDA
MRM
GLM
GAM
BRT
UniRF
Estimated variation fraction
True variation fraction
0.0
0.2
0.4
0.6
0.8
1.0
linear
Gaussian=0.002
Gaussian=0.02
Gaussian=0.2
autocor=5
autocor=50
N=25
N=400
Undersampling


dbRDA
RDA
RDA
BRT
UniRF
RDA
UniRF
BRT
0.0
0.2
0.4
0.6
0.8
1.0