Ecology, 101(1), 2020, e02908

Comparison of distance-based and model-based ordinations

David W. Roberts[1]

Department of Ecology, Montana State University, P.O. Box 173460, Bozeman, Montana 59717-3460 USA

Citation: Roberts, D. W. 2020. Comparison of distance-based and model-based ordinations. Ecology 101(1):e02908. 10.1002/ecy.2908

Abstract. Distance-based ordinations have played a critical role in community ecology for more than half a century, but are still under active development. These methods employ a matrix of pairwise distances or dissimilarities between sample units, and map sample units from the high-dimensional distance or dissimilarity space to a low-dimensional representation for analysis. Distance- or dissimilarity-based methods employ continuum or gradient ecological theory and a variety of statistical models to achieve the mapping. Recently, ecologists have developed model-based ordinations based on latent vectors and individual species response models. These methods employ the individualistic perspective of Gleason as the ecological model, and Bayesian or maximum-likelihood methods to estimate the parameters for the low dimensional space represented by the latent vectors. In this research I compared two distance-based methods (NMDS and t-SNE) with two model-based methods (BORAL and REO) on five data sets to determine which methods are superior for (1) extracting meaningful ecological drivers of variability in community composition, and (2) estimating sample unit locations in ordination space that maximize the goodness-of-fit of individual species response models to the estimated sample unit locations. Environmental variables and species were fitted to the ordinations by generalized additive models (GAMs) with Gaussian, negative binomial, or Poisson distribution models as appropriate. Across the five data sets, 22 models of environmental variability and 449 models of species distributions were calculated for each of the ordination methods. To minimize the effects of stochasticity the entire analysis was replicated three times and results averaged across the replicates. Results were evaluated by deviance explained and AIC for environmental variables and species distributions, averaged by ordination method for each data set, and ranked from best to worst. For the four assessments distance-based methods ranked 1 and 2 in three cases, and 1 and 3 in one case, significantly out performing the model-based methods. t-SNE was the top-performing method, out performing NMDS especially on the more complex data sets. In general the gradient-based theoretical basis and data sufficiency of distance-based methods allowed distance-based methods to outperform model-based methods in every assessment.

Key words: Bayesian ordination; Bryce Canyon; hunting spiders; Kullback–Leibler divergence; monotone regression; Mt. Field Massif; random effects ordination; Shoshone National Forest; unconstrained ordination.

Introduction

Ordination has been a central element of ecological community analysis for six decades. During that period the vast majority of ordination approaches have been based on matrices of pairwise differences in community composition between sample units using any of a large number of geometric distances or dissimilarity indices. Discussion of this variability in metrics or indices is beyond the scope of this work, but is extensively covered in textbooks such as Legendre and Legendre (2012) or Kent (2011). Roberts (2017) characterized and distinguished the properties and suitability of geometric distances vs. dissimilarities in ecological analysis. Although many of the best-performing indices are dissimilarities, as opposed to true metric distances, ordinations based on these pairwise matrices, as opposed to the original sample unit × species matrix, have been characterized under the general term "distance-based" ordinations. Common examples include principal components analysis (PCA, based on a correlation or variance/covariance matrix), principal coordinates analysis (PCoA, based on a distance or dissimilarity matrix), correspondence analysis (CA, based on a χ^2 distance matrix), and non-metric multidimensional scaling (NMDS, based on a distance or dissimilarity matrix). In general, the objective of these ordinations is to map the location of sample units from the high-dimensional space implied by the distance matrices to positions along gradients of community composition in a low-dimensional space. It's important to note that although the pairwise distances in the matrices are based on sample unit species composition, no information on individual species presence or abundance is retained in the distance matrix.

Manuscript received 13 February 2019; revised 4 August 2019; accepted 26 August 2019. Corresponding Editor: Helene H. Wagner.
[1] E-mail: droberts@montana.edu

Recently, new forms of ordination have been proposed that explicitly employ species response models to estimate the location of sample units along gradients of community composition referred to as "latent vectors." In this approach, there is no matrix of pairwise dissimilarities, but rather a set of individual species response models where community composition is determined by the amalgamation of the species-specific responses along the gradients. The new methods are generally referred to as "model-based" or "latent vector" ordination approaches. Proponents for model-based or latent vector ordinations (Walker and Jackson 2011, Hui et al. 2015, Hui 2016) argue that the explicit use of individual species response models provides a more direct foundation for ordination than does the use of dissimilarity or distance matrices, and that the formal introduction of statistical models of species response improves identifiability and evaluations of goodness of fit.

In this research I analyzed the performance of two distance-based methods non-metric multidimensional scaling (NMDS; Shepard 1962*a, b*, Kruskal 1964*a, b*) and t-distributed stochastic neighbor embedding (t-SNE; van der Maaten and Hinton 2008, van der Maaten 2014) —and two model-based methods—random effects ordination (REO; Walker and Jackson 2011) and bayesian ordination and regression analysis (BORAL; Hui 2016).

Ecological models and statistical models

Ordination analysis operates at the interface of two models: (1) an ecological model that defines the assumptions that govern the composition of ecological communities and their variability along environmental gradients, and (2) a statistical model that determines the location of sample units in low-dimensional space with a specified error structure. Choosing the appropriate ecological model and ensuring the suitability of a specific statistical model for the assumed ecological model is a primary challenge in the development and application of ordination methods.

Underlying ecological models.—The ecological model underlying distance-based ordination approaches was articulated by Curtis and McIntosh (1951) and Whittaker (1953), and formalized in Whittaker (1967). Given a species pool, any specific environment will have differential favorability for the species in the pool. Over time, driven by species' environmental responses and biotic interactions among species, the community will achieve a composition that is maximally adapted to that environment. As you move away from the focal location along an environmental gradient, environmental favorability and competitive ability will increase for some species and decrease for others, leading to a change in community composition. For smoothly varying environments we expect smoothly varying community compositions that reflect the underlying change in environment. Communities in close proximity along the gradients should be similar, and similarity will decline progressively with increasing distance along the gradient. Importantly, although individual species abundances may show stochastic variability among sample units along gradients, the whole-community perspective smooths out that variability by focusing on total community composition. I will refer to such models as "Whittakerian."

The ecological model underlying the model-based ordinations evaluated in this paper is the individualistic concept of Gleason (1926, 1939), which posits that each species has an individual response to environmental conditions and other species in its environment. Communities are simply the amalgam of individual species responses with the caveat that each species is part of every other species' environment. Thus the composition of communities can be modeled by modeling the abundance of each species separately and then combining fitted species responses to compose the community. I will refer to such models as "Gleasonian."

Underlying statistical models.—The statistical models underlying distance-based methods share a focus on pairwise distance matrices, but differ in other aspects of the models. In all cases, there is a two-step process: (1) choosing an informative pairwise dissimilarity index or distance, and (2) mapping the high-dimensional space implied by the dissimilarity index to the low-dimensional space of the ordination.

The statistical challenge in developing distance metrics or dissimilarity indices is ensuring that the pairwise distances calculated on community composition reflect the distances apart of the sample units along the underlying compositional gradients. Beginning with work by Swan (1970), ecologists have employed simulation models to identify indices of pairwise sample unit dissimilarity or distance that reflect locations along the simulated compositional gradients. Gauch and Whittaker (1972) employed an axiomatic approach to species response models and community composition and evaluated the Bray-Curtis index (Bray and Curtis 1957). Faith et al. (1987) evaluated a broad spectrum of geometric distances and dissimilarities in common use by ecologists for their ability to reflect distances in the ecological space of simulated data with known correct locations along the simulated gradients accurately. They explored nine specific simulated models across a range of β diversity, varying data transformations or standardizations, and distances or dissimilarities. In general, they found that species maximum standardized Kulczynski's index, sample-total standardized Manhattan distance, and the Bray-Curtis index maximized the relationship between known ecological distances and distances calculated on the simulated community compositions. Legendre and Gallagher (2001) performed a similar, but simpler, analysis with some of the same distances and dissimilarities employed by Faith et al. (1987) as well as some alternatives. Legendre and Gallagher (2001) determined that Hellinger distance and the Bray-Curtis index

out-performed the alternative distances and dissimilarities considered. Collectively, this body of work (and similar papers not cited for brevity) established the gradient-based model of community composition and the broad applicability of dissimilarities or distances in the analysis of community composition.

Given a distance or dissimilarity matrix, the two distance-based ordinations employed in this research use different statistical models. Nonmetric multidimensional scaling (NMDS) is based on monotone regression (Kruskal 1971), which attempts to order a set of distances as represented in a low-dimensional space to be monotonic with the order of the distances in a full-dimensional matrix. Efficient algorithms for solving monotone regression exist, and have been widely implemented. The algorithms employ iterative adjustments to sample unit locations (and thus pairwise distances) in the low-dimensional space and are generally deterministic from a given initial configuration. However, the algorithm is subject to finding local optima, and numerous initial configurations might be required to find a solution that approximates the global optimum. Non-metric multidimensional scaling is by far the oldest of the algorithms considered here, with early work by Shepard (1962*a, b*) and Kruskal (1964*a, b*) establishing the method half a century ago. Although originally developed for applications in psychology, NMDS has been widely employed in community ecology beginning with Anderson (1971), Dale (1975), Austin (1976), Fasham (1977), Prentice (1977), and Clymo (1980). Minchin (1987) demonstrated the superiority of NMDS for ecological analysis over then-current ordination methods in a benchmark paper, and NMDS is currently a commonly employed unconstrained ordination in ecological research with over 800 scientific papers published on ecological analysis using NMDS (ISI Web-of-Science search, August 4, 2019).

t-distributed stochastic neighbor embedding (t-SNE, van der Maaten and Hinton 2008, van der Maaten 2014) is a relatively new technique in dimension reduction that minimizes the Kullback–Leibler divergence between a probability distribution defined on the sample units in the full-dimensional space and another probability distribution defined on the lower-dimensional space. t-distributed stochastic neighbor embedding employs a Gaussian kernel centered on each sample unit in the high-dimensional space to define the probability that each of the other sample units is a "neighbor" in the high-dimensional space. Because of variability in the density of the high-dimensional space the variance associated with the Gaussian kernel is sample unit-specific, and is varied in a manner to maintain a relatively constant "perplexity," which effectively defines the number of neighbors for each sample unit. In contrast, a Student's t distribution is used as the kernel to determine the probability of neighbors in the low-dimensional space. The relatively longer tails of the t distribution create an emphasis on smaller neighborhoods and de-emphasize larger distances. A stochastic gradient descent algorithm is used to adjust the locations and distances of the points in the low-dimensional space and the iterations stop when a suitable low-dimensional representation is achieved. Like NMDS, t-SNE is subject to finding local optima and may require the use of multiple runs to achieve the best approximation to the global optimum. To my knowledge, t-SNE has only been employed in one ecological analysis to date (van der Maaten et al. 2012).

The two model-based methods explored here operate on similar but distinct statistical models. The REO ordinations assume that species abundances have a quadratic response to environmental variability, but recognize that if a species' optimum lies outside the range of the data then a monotonic response will be observed. The model constrains the locations of sample plots in the ordination to a grid of specified locations in ordination space, and uses a maximum-likelihood approach to estimate species model parameters and sample unit locations simultaneously. Over-fitting is reduced by cross-validation of the models to select the most parsimonious result. The use of a LASSO approach to parameter estimation results in monotonic responses for species where the quadratic term appears small. The model is designed to explicitly simulate modal responses, and complexity in the methodology limits ordinations to two-dimensional solutions.

BORAL is a Bayesian approach that attempts to estimate sample unit locations along latent vectors and fit species response models for each species to the sample unit locations along the latent vectors. BORAL relies on JAGS (Just Another Gibbs Sampler; Plummer 2018) for the Markov Chain–Monte Carlo (MCMC) process. Species response models are monotonic, but can employ either Poisson or negative binomial error distributions. Despite the monotonic assumption, the specification of random effects for sample units results in quasi-quadratic responses along simulated latent vectors (Jamil and ter Braak 2013). Unlike REO, BORAL can simulate three or higher-dimensional ordinations. However, increased dimensionality increases the number of parameters to estimate and high-dimensional attempts may fail because of the inability of the JAGS sampler to estimate parameters.

In principle it is possible to construct latent vector ordinations using joint species distribution models (Ovaskainen and Soininen 2011, Warton et al. 2015, Ovaskainen et al. 2017), but these methods are designed for more complex analyses including environmental covariates, and optionally species trait data and phylogenetic relatedness. The authors of the Hierarchical Models of Species Communities (HMSC) software user's manual (Norberg et al. 2019) note that latent vector ordination is not an appropriate use case for the HMSC approach. Vector Generalized Linear and Additive Models (VGLM and VGAM, Yee 2015) can also produce latent vector ordinations, but as Yee (2015, p. 232) notes such models are limited to low-dimensionality solutions and are highly sensitive to distributional assumptions. Exploratory analysis (not shown) confirmed that these methods were not

appropriate for the data sets considered in this research. Accordingly, neither HMSC nor VGAM were included in this analysis.

Ordination comparisons

Because of the proliferation of ordination methods over the last several decades, many comparisons of the effectiveness of various ordination techniques have been conducted (e.g., Minchin 1987, Podani 1989, Warwick and Clarke 1991, Zhang and Oxley 1994, Giraudel and Lek 2001, Roberts 2009). However, given the recent arrival of t-SNE and model-based and latent vector ordinations relatively few rigorous comparisons have been conducted (although see Hui et al. 2015). Table 1 summarizes the properties of the four methods under consideration.

Objectives

The objective of this analysis was to compare distance- or dissimilarity-based ordination methods and model-based ordination methods for their ability rigorously: (1) to extract meaningful ecological drivers of variability in community composition and (2) to estimate sample unit locations in ordination space that maximize the goodness of fit of individual species response models to the estimated sample unit locations.

Table 1. Ordination underlying models.

Ordination	Ecological model	Statistical model
NMDS	Whittakerian	Monotonic regression
t-SNE	Whittakerian	Minimize Kullback–Leibler divergence
BORAL	Gleasonian	Markov Chain–Monte Carlo Bayesian
REO	Gleasonian	Maximum likelihood

NMDS, nonmetric multidimensional scaling; t-SNE, t-distributed stochastic neighbor embedding; BORAL, Bayesian ordination and regression analysis; REO, random effects ordination

The distance- or dissimilarity-based ordinations were represented by the most commonly used distance-based ordination (NMDS) and the more recent t-SNE. The model-based methods were represented by REO (Walker and Jackson 2011) and BORAL (Hui 2016). Unlike NMDS and t-SNE, both REO and BORAL originated in ecological research, but because of their relatively recent invention have been employed in many fewer analyses to date.

Methods

Data

The analyses were conducted on five data sets: one simulated coenocline data set and four well-known data sets commonly used in method comparisons. Table 2 presents a list of parameters describing the size and complexity of the various data sets. Table 3 presents the environmental variables selected for analysis for each data set.

The simulated data (COENOFLEX) come from the R package *coenoflex* (Roberts 2016*a*). Coenoflex simulates individual species responses along simulated gradients. Species' physiological responses are unimodal curves generated by a fuzzy Pi function (Roberts 2008). Species abundances in simulated sample units are calculated to simulate competition with size asymmetry. The particular data set simulated is the example data set from the coenoflex documentation, and comprises 100 sample units along three gradients with 100 species of variable amplitude, abundance, and skew distributed in the three-dimensional simulated space. The Bray-Curtis dissimilarity was calculated on a $\log(x + 1)$ transformation of simulated species abundances.

The second data set (SPIDERS) is the well-known hunting spiders data set (van der Aart and Smeenk-Enserink 1974) as employed by ter Braak (1986) in a demonstration of canonical correspondence analysis (CCA) and by Roberts (2009) in a demonstration of multidimensional fuzzy set ordination (MFSO), as well as by others. The data comprise the abundances of 12

Table 2. Data properties.

	N†	S‡	S_5§	β¶	β_5#	D‖	D_5††	PCO_3‡‡
Coenoflex	100	100	85	7.70	6.66	0.115	0.131	0.291
Spiders	28	12	12	0.85	0.85	0.542	0.542	0.803
Bryce Canyon	160	160	100	11.35	6.79	0.081	0.128	0.432
Shoshone Forest	150	368	108	20.28	6.63	0.047	0.131	0.320
Mt. Field	424	209	144	7.35	4.82	0.119	0.172	0.387

†Number of sample units.
‡Number of species.
§Number of species occurring five times of more.
¶Whittaker β diversity = $(S/r) - 1$, where r is mean sample unit species richness.
#Whittaker β diversity for species occurring five times of more.
‖Matrix density = number of non-zero values/$(N \times S)$.
††Reduced matrix density = number of non-zero values for species that occur $\geq$ five times/$(N \times S_5)$.
‡‡variability explained in a three-dimensional principal coordinates analysis.

Table 3. Environmental variables by data set.

Data set variable	Description	Range
Coenoflex		
I	Simulated gradient	[0,300]
II	Simulated gradient	[0,200]
III	Simulated gradient	[0,100]
Spiders		
Soil dryness	Soil dry mass	[0.956,3.518]
Bare sand	Cover of bare sand	[0.000,4.511]
Fallen leaves	Cover of fallen leaves	[0.000,4.605]
Moss cover	Cover of moss	[0.000,4.331]
Herb cover	Cover of the herb layer	[0.693,4.615]
Reflection	Solar reflection under a clear sky	[0.000,4.382]
Bryce Canyon		
Elevation	Elevation above sea level in meters	[2027,2737]
Slope	Slope steepness in percent	[0,65]
Aspect value	(cos(azimuth-30) + 1)/2	[0,1]
Shoshone forest		
Elevation	Elevation above sea level in meters	[1780,3085]
Slope	Slope steepness in percent	[0,73]
Aspect value	(cos(azimuth-30) + 1)/2	[0,1]
Site Water balance	Precipitation minus potential evapotranspiration in mm	[−531,244]
Spring temperature	Mean spring temperature in °C	[−1.48,5.64]
Frost-free Days	Growing season length between frosts in days	[132,236]
Mt. Field		
Elevation	Elevation above sea level in meters	[910,1380]
Drainage	Soil drainage class	[1,5]
Radiation	Direct solar radiation	[9.10,16.50]
Slope	Slope steepness in degrees	[0.0,38.00]

spider species in 28 sample units with six associated environmental variables. Bray-Curtis dissimilarity was calculated on the raw counts.

The third data set (BRYCE) is the Bryce Canyon vegetation data available in R package *labdsv* (Roberts 2019). The data comprise 169 non-tree vascular plants in 160 sample units located in Bryce Canyon National Park, Utah, United States. The data are cover-class midpoints of an eight-class cover scale [midpoints = 0.1, 0.5, 3.0, 15.0, 37.5, 62.5, 85.0, 97.5]. The data were $\log(x+1)$ transformed for the calculation of the Bray-Curtis dissimilarity. The data have previously been analyzed by Roberts (2008), Blanchet et al. (2008), De Cáceres and Legendre (2008), and others. Three environmental variables that have previously been shown to be informative were selected for analysis.

The fourth data set (SHOSHONE) is the Shoshone National Forest data available in R package *optpart* (Roberts 2016*b*). The data are 150 sample units randomly selected from a larger database of vegetation sample units on the Shoshone National Forest, Wyoming, United States. Percent cover estimates were made for 368 vascular plant species in the sample units based on a 12-class cover scale [midpoints = 0.1, 3.0, 10.0, 20.0, 30.0, 40.0, 50.0, 60.0, 70.0, 80.0, 90.0, 95.0]. Bray-Curtis dissimilarity was calculated on $\log(x+1)$ transformed abundances. The data have been previously employed in ecological methods research by Roberts (2008, 2015). Six environmental variables were selected for analysis.

The fifth data set (MTFIELD) comes from the Mount Field Massif in south-central Tasmania as described by Minchin (1983, 1989). The data set consists of sample plots of 100 m^2 with species percent cover estimated as the midpoint of six cover classes [midpoints = 0.05, 0.55, 5.5, 20, 50, 85]. There are 424 sample plots with a total of 209 species. The Bray-Curtis dissimilarity was calculated on $\log(x+1)$ mid-point abundances. Four environmental variables were selected for analysis. The data are published at figshare[2] as mtfield_veg.csv and mtfield_site.csv for the vegetation and site data, respectively.

Ordinations

NMDS was conducted using the *bestnmds* function from R package *labdsv*. The function calls the *isoMDS* function of R package *MASS* (Venables and Ripley 2002) with one instance starting from a PCoA as the initial conditions, and an additional 99 random starts. The result with the lowest stress was selected for analysis.

The t-SNE ordinations were fitted with the *besttsne* function in R package *labdsv* which calls the *Rtsne* R package (Krijthe 2015). *Rtsne* is quite flexible, with multiple options; the analyses in this paper employed the same Bray-Curtis dissimilarity matrix used in the NMDS ordinations, with perplexity set at 30 (but reduced to five for the small SPIDERS data), and θ set to 0.0. Like the NMDS analysis, t-SNE was initiated from a PCoA solution as well as 99 random starts; the result with the lowest Kullback—Leibler divergence was selected for analysis.

Latent vector ordination (BORAL) was conducted with the *boral* function from the R package of the same name (Hui 2016, 2018). For the BORAL analyses all species that occurred fewer than five times were deleted, and the data were converted to integers, if necessary, by a ceiling function. The data were modeled with a negative binomial error distribution for species and sample unit random effects. If the model failed to converge because of limitations of JAGS, the model was refit with a simpler Poisson error distribution.

The REOs were fit with the *ltm.ecol* function in the *reo* R package (Walker 2011). For REO, all species that occurred fewer than five times were deleted, and the data were converted to presence–absence. *ltm.ecol* employs a regularization parameter λ to control the complexity of the fits, and was supplied with a vector of candidate

[2] https://doi.org/10.6084/m9.figshare.9912551.v1

values in the set {0.1, 1.1, 2.1, 3.1, 4.1}. The models were fit with cross-validation (cvic=TRUE) and 20 points along each latent vector.

Analyses

For each data set ordinations were constructed as described above. For NMDS, BORAL, and t-SNE both two-dimensional and three-dimensional solutions were sought; REO is currently limited to two-dimensional solutions. An R script to replicate the analyses is included as Metadata S1.

Environmental analyses.—For each ordination type and dimensionality, for each data set, environmental variables were fitted post hoc to the ordination with the *calibrate* function from R package *labdsv,* which employs a GAM (Wood 2000, 2017) using the ordination coordinates as explanatory variables. Models were first fitted with thin-plate spline smoothers; if the models failed because of insufficient degrees of freedom they were re-fit with independent splines. For each data set, for each ordination type and dimensionality, for each variable, the deviance explained [(null deviance–residual deviance)/null deviance] was calculated, rounded to three digits, and ranked from highest to lowest by ordination type. For each ordination type and dimensionality, for each data set, mean rank was calculated across all variables in the data set. Finally, for each ordination type and dimensionality, a global mean rank was calculated across the five data sets and ranked. Means that differed less than 1% were considered ties. This process was then repeated for Akaike's information criterion (AIC) ranked from lowest to highest, with mean rank calculated across variables for each data set, along with a global mean across the five data sets.

Species response analysis.—For each ordination type and dimensionality, for each data set, species response models were fitted by the *predict* function from R package *labdsv* using a GAM for all species that occurred at least five times. Species were modeled using a thinplate spline with a negative binomial error function. If the models failed because of insufficient degrees of freedom they were re-fit with independent splines. For each ordination type and dimensionality, for each data set, models for each species were ranked for deviance explained and AIC and averaged by ordination type and dimensionality to produce mean ranks. Global means were then calculated across the data sets and ranked. Means that differed by less then 1% were considered tied.

All analyses were conducted using R (version 3.6.0, R Development Core Team 2019), employing packages *boral* (Hui 2018), *reo* (Walker 2011), *labdsv* (Roberts 2019), *coenoflex* (Roberts 2016*a*), mgcv (Wood 2019), and *Rtsne* (Krijthe 2015).

Results

In total, three replicates of 35 ordinations were calculated (seven ordination types by five data sets). Across the five data sets, 22 models of environmental variability and 449 models of species distributions were fitted. Fig. 1 presents an example of the environmental

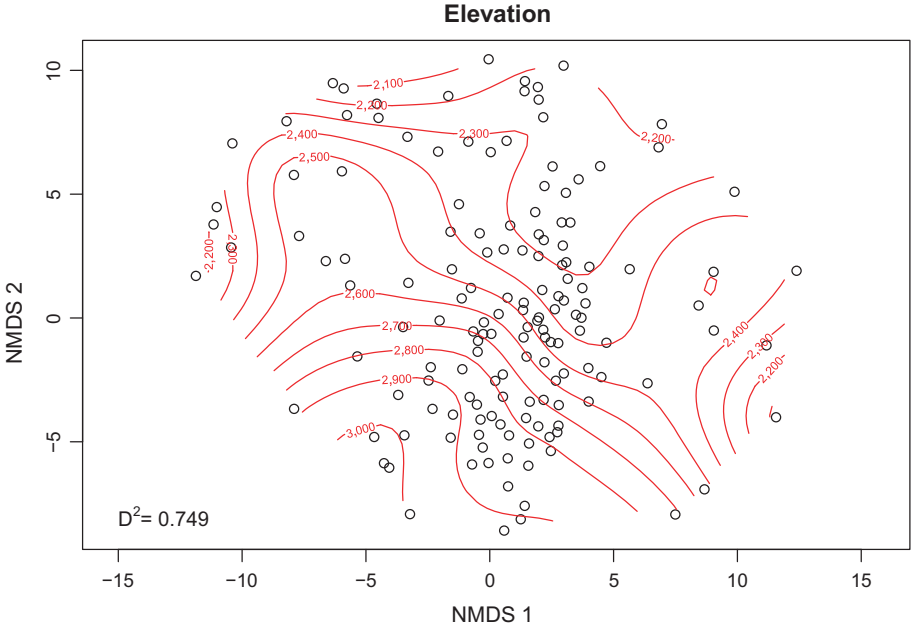

Fig. 1. Example demonstrating the generalized additive model thin-plate spline for elevation in the Shoshone National Forest on an nonmetric multidimensional scaling (NMDS) ordination with a Gaussian error distribution. Black circles are sample units; red lines are isolines of elevation (m). D^2 = deviance explained.

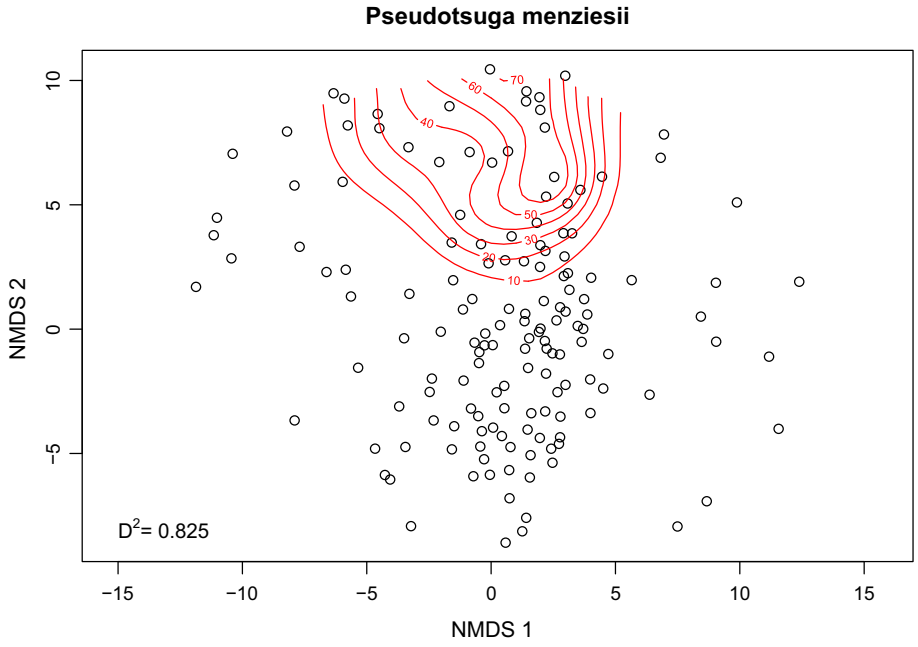

FIG. 2. Example demonstrating the generalized additive model thin-plate spline for *Pseudotsuga menziesii* on the Shoshone National Forest with a negative binomial error distribution. Black circles are sample units; red lines are isolines of percent cover for *Pseudotsuga menziesii*. D^2 = deviance explained.

analyses with a GAM model of elevation fitted to an NMDS ordination of the Shoshone National Forest. Fig. 2 shows an example of an individual species distribution model, showing the percent cover of *Pseudotsuga menziesii*, a common mid-to-low-elevation tree species, on the same NMDS ordination. Overall results are presented in Tables 4–7, for environmental deviance explained, environmental model AIC, species response deviance explained, and species model AIC, respectively.

Environmental model deviance explained

Table 4 gives the average rank across all variables for each ordination for each data set, as well as global mean ranks and final overall rank by ordination. Across all five data sets (3-D) NMDS explained the most deviance of environmental variables, followed by 3-D t-SNE. Interestingly, 3-D NMDS did best on the smaller or simpler data sets (COENOFLEX and SPIDERS), whereas 3-D t-SNE did better on the larger or more complex data sets (SHOSH and MTFIELD).

Surprisingly, the 3-D ordinations did not universally out-perform the 2-D ordinations. Specifically, 2-D t-SNE ranked third across all data sets and out-performed 3-D BORAL. Nonetheless, 3-D solutions generally did best. With respect to the stated objectives, distance-based methods ranked 1, 2, 3, and 6, and model-based methods ranked 4, 5, and 7.

TABLE 4. Environmental deviance explained.

	BORAL	REO	NMDS	t-SNE	BORAL	NMDS	t-SNE
Dimensions	2	2	2	2	3	3	3
Coenoflex	6.00	4.22	3.67	3.44	4.00	1.56	5.11
Spiders	4.67	4.72	4.17	4.22	4.22	2.67	3.33
Bryce Canyon	7.00	4.22	5.33	2.78	4.33	2.22	2.11
Shoshone Forest	4.33	6.28	5.83	3.11	3.61	3.78	1.06
Mt. Field	6.08	5.25	6.00	4.17	3.00	2.25	1.25
Mean	5.62	4.94	5.00	3.54	3.83	2.49	2.57
Rank	7	5	6	3	4	1	2

Notes: BORAL, Bayesian ordination and regression analysis; REO, random effects ordination; NMDS, nonmetric multidimensional scaling; t-SNE, t-distributed stochastic neighbor embedding.

Values in the body of the table are mean ranks by method, dimensionality, and data set. Values in the lower tier are column means and ranks.

Environmental model AIC

For environmental variable AICs 3-D t-SNE performed best, followed by 3-D NMDS; other ranks did not change from deviance explained (Table 5). This suggests that the slightly better results obtained by 3-D NMDS in deviance explained were slightly over-fit compared to t-SNE. Distance-based approaches again ranked 1, 2, 3, and 6; model-based methods ranked 4, 5, and 7.

Species distribution models deviance explained

Table 6 gives the deviance explained for the species distribution models. In general, 3-D t-SNE proved superior, followed by 3-D NMDS and then 3-D BORAL. In contrast to the environmental models, 3-D solutions were notably better than any 2-D model. Among the 2-D models, t-SNE tied for best with REO, followed by NMDS and then BORAL. Interestingly, t-SNE again performed best on the larger, more complicated data sets. With respect to the specific objectives, distance-based ordinations ranked 1, 2, 4.5, and 6, and model-based methods ranked 3, 4.5, and 7.

Species distribution models AIC

Table 7 gives the rank results for species distribution AIC. In this case, 3-D t-SNE again ranked first, but 3-D NMDS and 3-D BORAL reversed ranks. This suggests that the 3-D NMDS results were somewhat over-fit, requiring excessive degrees of freedom to achieve best results. Interestingly, althrough the results of the 3-D methods were better than any 2-D method, the differences were less pronounced than for species deviance explained. With respect to the specific objectives, distance-based methods ranked 1, 3, 4, and 6.5, and the model-based methods ranked 2, 5, and 6.5.

Discussion

In general, dissimilarity-based methods outperformed the model-based methods for both environmental interpretation and species distribution models on the tests performed in this research. For the 3-D ordinations either NMDS or t-SNE was ranked first in every analysis, and the dissimilarity-based methods ranked first and second three out of four times. For BORAL 3-D the best result was species AIC where it ranked second to t-SNE.

Table 5. Environmental Akaike information criterion.

	BORAL	REO	NMDS	t-SNE	BORAL	NMDS	t-SNE
Dimensions	2	2	2	2	3	3	3
Coenoflex	6.33	4.44	4.00	4.44	3.44	1.00	4.33
Spiders	4.06	4.94	5.00	4.06	4.11	2.72	3.11
Bryce Canyon	6.33	3.00	3.78	3.56	6.33	2.22	2.78
Shoshone Forest	3.44	5.89	5.83	2.89	4.56	4.22	1.17
Mt. Field	5.75	5.00	6.00	3.42	3.42	3.42	1.00
Mean	5.18	4.66	4.92	3.67	4.37	2.72	2.48
Rank	7	5	6	3	4	2	1

Notes: BORAL, Bayesian ordination and regression analysis; REO, random effects ordination; NMDS, nonmetric multidimensional scaling; t-SNE, t-distributed stochastic neighbor embedding.

Values in the body of the table are mean ranks by method, dimensionality, and data set. Values in the lower tier are column means and ranks.

Table 6. Species deviance explained.

	BORAL	REO	NMDS	t-SNE	BORAL	NMDS	t-SNE
Dimensions	2	2	2	2	3	3	3
Coenoflex	5.40	3.66	5.36	4.06	2.54	3.09	3.88
Spiders	4.92	5.00	4.53	4.39	3.33	2.75	3.08
Bryce Canyon	6.00	4.59	5.60	4.53	2.99	2.40	1.88
Shoshone Forest	5.99	4.67	5.51	4.32	3.10	2.65	1.77
Mt. Field	5.82	4.16	5.99	4.56	2.77	2.96	1.74
Mean	5.63	4.42	5.40	4.37	2.95	2.77	2.47
Rank	7	4.5	6	4.5	3	2	1

Notes: BORAL, Bayesian ordination and regression analysis; REO, random effects ordination; NMDS, nonmetric multidimensional scaling; t-SNE, t-distributed stochastic neighbor embedding.

Values in the body of the table are mean ranks by method, dimensionality, and data set. Values in the lower tier are column means and ranks.

TABLE 7. Species Akaike information criterion.

	BORAL	REO	NMDS	t-SNE	BORAL	NMDS	t-SNE
Dimensions	2	2	2	2	3	3	3
Coenoflex	5.50	3.95	5.34	4.54	2.56	2.89	3.22
Spiders	3.97	5.50	4.11	4.47	2.44	3.64	3.86
Bryce Canyon	4.84	3.70	4.54	3.25	4.20	4.06	3.41
Shoshone Forest	4.77	3.64	4.51	3.54	4.31	4.29	2.96
Mt. Field	4.88	3.41	5.29	3.89	3.64	4.17	2.71
Mean	4.79	4.04	4.76	3.94	3.43	3.81	3.23
Rank	6.5	5	6.5	4	2	3	1

Notes: BORAL, Bayesian ordination and regression analysis; REO, random effects ordination; NMDS, nonmetric multidimensional scaling; t-SNE, t-distributed stochastic neighbor embedding.

Values in the body of the table are mean ranks by method, dimensionality, and data set. Values in the lower tier are column means and ranks.

Comparing just the 2-D ordinations t-SNE again ranked highest, and in fact 2-D t-SNE had a better mean rank across all four analyses than 3-D BORAL.

Ecological model effects

For the environmental analyses, it is perhaps to be expected that the dissimilarity-based methods would do well as a consequence of their underlying gradient-based ecological model. It is important to note, however, that these are all unconstrained ordinations and no environmental data were employed in determining the coordinates of the sample units in the ordinations. Rather, environmental variables were fit post hoc to the existing ordination. Aside from the simulated COENOFLEX data set, it is impossible to know that the chosen environmental variables actually determine or constrain the composition of the communities analyzed. Nonetheless, it seems unlikely that the results are spurious or artifactual. I interpret the variability in results as indicative of model sensitivity.

It might be expected, given the Gleasonian ecological model underlying the model-based ordinations, that the model-based ordinations would prove superior for individual species distribution models. Surprisingly, they did not, and the dissimilarity-based methods achieved better results in this case as well. Three-dimensional t-SNE ranked first for both deviance explained and AIC. Three-dimensional NMDS and 3-D BORAL both ranked second and third, with NMDS explaining more deviance and BORAL achieving lower AIC scores.

Statistical model complexity

The fact that model-based methods did not perform as well as t-SNE or NMDS (especially with respect to environmental interpretation) reflects the extraordinary complexity of fitting these models. The models need to estimate multiple parameters for both species and sample units. Using BORAL as an example, for a negative binomial model fit, the number of parameters to fit is given by

$$p = S \times (D + 2) + N \times (D + 1)$$

where p = the number of parameters to estimate, S = the number of species, D = the number of dimensions in the ordination, and N = the number of sample units.

Specifically, for each of S species there needs to be an intercept, a β coefficient for each dimension, and an estimate of the negative binomial scaling parameter θ. Each of the N sample units needs a coordinate for each dimension and an estimate of random effects. For a Poisson fit, the requirements are reduced slightly to

$$p = S \times (D + 1) + N \times (D + 1) = (S + N) \times (D + 1).$$

In contrast, for the distance-based methods the number of parameters to estimate is simply $N \times D$.

Ideally, for each parameter to be estimated we would like some minimal number of observations. Unfortunately, ecological community data sets are often sparse, with many species exhibiting numerous zeros. When fitting species response models some number of zeros will be required to estimate a species' distribution if its limits lie within the range of the data. However, excessive zeros resulting from extended gradients where a species does not occur add little additional information to the model. This topic received considerable attention with respect to fitting species response models as β functions (Austin et al. 1994, Austin and Nicholls 1997, Oksanen 1997).

In Table 8 I present the total number of non zero occurrences for each data set, along with the number of parameters to be estimated by model-based methods (assuming a negative binomial error distribution and random effects for sample units) and distance-based methods. For the model-based methods if we assume that a number of absences (or zeros) equal to the number of presences is ideal for fitting the models, we can compare the number of data points to the number of parameters to estimate. As can be seen in Table 8, except for the relatively dense Mt. Field data set, none of the other data sets offer as many as five data points per parameter, especially for the 3-D ordinations. Arguably, since we are fitting all of the parameters simultaneously they

Table 8. Data points per parameter to estimate for model-based methods and distance-based methods.

	Bayesian ordination and regression analysis					Distance based			
	Parameters			(2*NZ)/P§		Parameters		(N−1)/(2D)¶	
	2-D†	3-D‡	Nonzero data	2-D	3-D	2-D	3-D	2-D	3-D
Coenoflex	640	825	1,110	3.47	2.69	200	300	24.75	16.50
Spiders	132	172	182	2.76	2.12	56	84	6.75	4.50
Bryce Canyon	907	1,176	2,053	4.53	3.49	320	480	39.75	26.50
Shoshone Forest	882	1,140	2,123	4.81	3.72	300	450	37.25	24.83
Mt. Field	1,848	2,416	10,488	11.35	8.68	848	1,272	105.75	70.50

†Two dimensional.
‡Three dimensional.
§Two times the number of non-zero data points divided by the number of parameters to estimate.
¶Number of sample unit pairs divided by number of parameters to estimate, $\big((N^2 - N)/2\big)/(N \times D) = (N - 1)/2D$.

support each other more than indicated. Nonetheless, there is minimal information available to support estimating complex models from these ecological data sets. In contrast, for the distance-based methods the number of data points is the number of pairwise dissimilarities ($(N^2 - N)/2$), and the number of parameters to estimate is only $N \times D$. The ratio of data points to parameters reduces to $\big((N^2 - N)/2\big)/(N \times D) = (N - 1)/(2 \times D)$.

It might be assumed that as the density of a data set increases the relative performance of the model-based methods would improve. For 2-D BORAL the best results were obtained on the SPIDER data, which is by far the highest density of the data sets. In fact, the 2-D and 3-D SPIDER models are the only cases where the JAGS Gibbs sampler in BORAL achieved results with a negative binomial distribution; the other four data sets had to be fit with a Poisson error distribution. For 3-D BORAL, the best results were obtained for the COENOFLEX data, closely followed by the MTFIELD data. The COENOFLEX data are not especially dense, but closely align with the assumptions of BORAL, and 3-D BORAL ranked second among the three 3-D methods. The MTFIELD data have the highest ratio of data points to parameters of the five data sets analyzed here, and 3-D BORAL tied for second among the three 3-D methods on the MTFIELD data. Surprisingly, REO did best on the BRYCE data set, which is among the least dense. It does not appear that higher data density is sufficient to overcome the shortcomings of the statistical models of the model-based methods, and it is likely the case that having data that more closely follow the assumptions is more important.

It is conceivable that future improvements to the sampler employed in the MCMC analysis would provide better results for the Bayesian approach, but the results of these comparisons suggest that the methods are data limited, and that the statistical model is not a good fit to ecological data. Even though the random effects for sample units results in a quasi-quadratic response for individual species, directly fitting low-order polynomial models would be preferable, but is simply impossible because of data limitations.

Implications

Given the results presented here it is clear that distance-based non constrained ordinations can be expected to be superior to currently available model-based methods. Despite the appeal of the Gleasonian ecological model, and the seemingly direct approach of modeling individual species, the data are generally insufficient to meet the requirements of the models. It is a common experience among ecologists that species distribution models can only be fit to the more common or dominant species in a community because of lack of statistical power for many less common species. Reducing the data sets employed in these analyses to species occurring at least five times reduced the individual data sets by as much as 70%, and yet many species still produced models no better than the null model as tested by AIC. This is a dilemma inherent to many ecological community data sets.

New Methods

Both *boral* and *reo* are fairly recent developments, and have not previously been rigorously tested on more complex data sets. The results presented here suggest that these methods need to be more thoroughly tested. NMDS is one of the oldest ordination methods still in common use. In recent years the interests of ordination developers moved more to constrained ordination (e.g., CCA, MFSO) or model-based ordination. However, the increased interest in methods of dimension reduction among data scientists has led to a renewal of interest in distance-based unconstrained ordination. Minimization of Kullback–Leibler divergence in t-SNE is statistically more sophisticated than the monotone regression in NMDS, and shows promise for improving the performance of distance-based methods. Based on the results presented here, t-SNE appears quite promising and worthy of more extended testing by the ecological community.

Acknowledgments

I would like to thank Peter Minchin for permission to use the Mount Field data, and Kent Houston for the use of

the Shoshone Forest data. I would like to thank two anonymous reviewers for suggestions that significantly improved the manuscript.

Literature Cited

Anderson, A. J. B. 1971. Ordination methods in ecology. Journal of Ecology 59:713–726.

Austin, M. P. 1976. Performance of four ordination techniques assuming three different non-linear species response models. Vegetatio 33:4349.

Austin, M. P., and A. O. Nicholls. 1997. To fix or not to fix the species limits, that is the ecological question: Response to Jari Oksanen. Journal of Vegetation Science 8:743–748.

Austin, M. P., A. O. Nicholls, and M. D. Doherty. 1994. Determining species response functions to an environmental gradient by means of a Beta-function. Journal of Vegetation Science 5:215–228.

Blanchet, F. G., P. Legendre, and S. Borcard. 2008. Forward selection of explanatory variables. Ecology 89:2623–2632.

Bray, J. R., and J. T. Curtis. 1957. An ordination of the upland forest communities of southern Wisconsin. Ecological Monographs 27:326–349.

Clymo, R. S. 1980. Preliminary survey of the peat-bog Hummell Knowe moss using various numerical methods. Vegetatio 42:129–148.

Curtis, J. T., and R. P. McIntosh. 1951. An upland forest continuum in the prairie–forest border region of Wisconsin. Ecology 32:476–496.

Dale, M. B. 1975. On objectives of methods of ordination. Vegetatio 30:15–32.

De Cáceres, M., and P. Legendre. 2008. Beals smoothing revisited. Oecologia 156:657–669.

Faith, D. P., P. R. Minchin, and L. Belbin. 1987. Compositional dissimilarity as a robust measure of ecological distance. Vegetatio 69:57–68.

Fasham, M. J. R. 1977. A comparison of nonmetric multidimensional scaling, principal components and reciprocal averaging for the ordination of simulated coenoclines and coenoplanes. Ecology 58:551–561.

Gauch, H. G., and R. H. Whittaker. 1972. Coenocline simulation. Ecology 53:446–451.

Giraudel, J. L., and S. Lek. 2001. A comparison of self-organizing map algorithm and some conventional statistical methods for ecological community ordination. Ecological Modelling 146:329–339.

Gleason, H. A. 1926. The individualistic concept of the plant association. Bulletin of the Torrey Botanical Club 53:7–26.

Gleason, H. A. 1939. The individualistic concept of the plant association. American Midland Naturalist 21:91–110.

Hui, F. C., S. Taskinen, S. Pledger, S. D. Foster, and D. I. Warton. 2015. Model-based approaches to unconstrained ordination. Methods in Ecology and Evolution 6:399–411.

Hui, F. K. C. 2016. Bayesian ordination and regression analysis of multivariate abundance data in R. Methods in Ecology and Evolution 7:744–750.

Hui, F. K. C. 2018. boral: Bayesian ordination and regression analysis. R package version 1.7. https://CRAN.R-project.org/package=boral

Jamil, T., and C. J. F. ter Braak. 2013. Generalized linear mixed models can detect unimodal species-environment relationships. Peer J 1:e95.

Kent, M. 2011. Vegetation description and data analysis: A practical approach. Wiley-Blackwell, Hoboken, New Jersey, USA.

Krijthe, J. H. 2015. Rtsne: t-Distributed stochastic neighbor embedding using a Barnes-Hut implementation. https://github.com/jkrijthe/Rtsne

Kruskal, J. B. 1964*a*. Multidimensional-scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika 29:1–27.

Kruskal, J. B. 1964*b*. Nonmetric multidimensional-scaling—a numerical-method. Psychometrika 29:115–129.

Kruskal, J. B. 1971. Monotone regression—continuity and differentiability properties. Psychometrika 36:57–62.

Legendre, P., and E. D. Gallagher. 2001. Ecologically meaningful transformations for ordination of species data. Oecologia 129:271–180.

Legendre, P., and L. Legendre. 2012. Numerical ecology. Third English edition. Elsevier, Amsterdam, The Netherlands.

Minchin, P. R. 1983. A comparative evaluation of techniques for ecological ordination using simulated vegetation data and an integrated ordination-classification analysis of the alpine and subalpine plant communities of the Mt. Field Plateau, Tasmania. Dissertation, University of Tasmania, Hobart, Tasmania, Australia.

Minchin, P. R. 1987. An evaluation of the relative robustness of techniques for ecological ordination. Vegetatio 69:89–107.

Minchin, P. 1989. Montane vegetation of the Mt. Field Massif, Tasmania—a test of some hypotheses about properties of community patterns. Vegetatio 83:97–110.

Norberg, A., G. Tikhonov, F. G. Blanchet, N. Abrego, and O. Ovaskainen. 2019. User manual for the software packages HMSC-Matlab and HMSC-R 2.0. https://www.helsinki.fi/sites/default/files/atoms/files/hmsc_manual.pdf, Accessed 7/6/2019

Oksanen, J. 1997. Why the beta-function cannot be used to estimate skewness of species responses. Journal of Vegetation Science 8:147–152.

Ovaskainen, O., and J. Soininen. 2011. Making more out of sparse data: hierarchical modeling of species communities. Ecology 92:289–295.

Ovaskainen, O., G. Tikhonov, A. Norberg, F. G. Blanchet, L. Duan, D. Dunson, T. Roslin, and N. Abrego. 2017. How to make more out of community data? A conceptual framework and its implementation as models and software. Ecology Letters 20:561–576.

Plummer, M. 2018. rjags: Bayesian graphical models using MCMC. R package version 4. 8. https://CRAN.R-project.org/package=rjags

Podani, J. 1989. Comparison of ordinations and classifications of vegetation data. Vegetatio 83:111–128.

Prentice, I. C. 1977. Non-metric ordination methods in ecology. Journal of Ecology 65:85–99.

R Core Team. 2019. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Roberts, D. W. 2008. Statistical analysis of multidimensional fuzzy set ordinations. Ecology 89:1246–1260.

Roberts, D. W. 2009. Comparison of multidimensional fuzzy set ordination with CCA and DB-RDA. Ecology 90:2622–2634.

Roberts, D. W. 2015. Vegetation classification by two new iterative reallocation optimization algorithms. Plant Ecology 216:741–758.

Roberts, D. W. 2016*a*. coenoflex: Gradient-based coenospace vegetation simulator. R package version 2.2-0. https://CRAN.R-project.org/package=coenoflex

Roberts, D. W. 2016*b*. optpart: Optimal partitioning of similarity relations. R package version 2.3-0. https://CRAN.R-project.org/package=optpart

Roberts, D. W. 2017. Distance, dissimilarity, and mean–variance ratios in ordination. Methods in Ecology and Evolution 8:1398–1407.

Roberts, D. W. 2019. labdsv: Ordination and multivariate analysis for ecology. R package version 2.0-1. https://CRAN.R-project.org/package=labdsv

Shepard, R. N. 1962*a*. The analysis of proximities—multidimensional-scaling with an unknown distance function. Psychometrika 27:125–140.
Shepard, R. N. 1962*b*. The analysis of proximities—multidimensional-scaling with an unknown distance function 2. Psychometrika 27:219–246.
Swan, J. M. A. 1970. An examination of some ordination problems by use of simulated vegetational data. Ecology 51:89–102.
ter Braak, C. J. F. 1986. Canonical correspondence-analysis—a new eigenvector technique for multivariate direct gradient analysis. Ecology 67:1167–1179.
van der Aart, P. J. M., and N. Smeenk-Enserink. 1974. Correlations between distributions of hunting spiders (Lycosidae, Ctenidae) and environmental characteristics in a dune area. Netherlands Journal of Zoology 25:1–45.
van der Maaten, L. 2014. Accelerating t-SNE using tree-based algorithms. Journal of Machine Learning Research 15:3221–3245.
van der Maaten, L., and G. E. Hinton. 2008. Visualizing data using t-SNE. Journal of Machine Learning Research 9:2431–2456.
van der Maaten, L., S. Schmidtlein, and M. D. Mahecha. 2012. Analyzing floristic inventories with multiple maps. Ecological Informatics 9:1–10.
Venables, W. N., and B. D. Ripley. 2002. Modern applied statistics with S. Fourth edition. Springer, New York, New York, USA.
Walker, S. C., and D. A. Jackson. 2011. Random-effects ordination: describing and predicting multivariate correlations and co-occurrences. Ecological Monographs 81:635–663.
Walker, W. 2011. reo: Random effects ordination. http://R-Forge.R-project.org
Warton, D. I., F. G. Blanchet, R. B. O'Hara, O. Ovaskainen, S. Taskinen, S. C. Walker, and F. K. C. Hui. 2015. So many variables: joint modeling in community ecology. Trends in Ecology & Evolution 30:766–779.
Warwick, R. M., and K. R. Clarke. 1991. A Comparison of some methods for analyzing changes in benthic community structure. Journal of the Marine Biological Association of the United Kingdom 71:225–244.
Whittaker, R. H. 1953. A consideration of climax theory—the climax as a population and pattern. Ecological Monographs 23:41–78.
Whittaker, R. H. 1967. Gradient analysis of vegetation. Biological Reviews 42:207–264.
Wood, S. N. 2000. Modelling and smoothing parameter estimation with multiple quadratic penalties. Journal of the Royal Statistical Society. Series B 62:413–428.
Wood, S. N. 2017. Generalized additive models: An introduction with R. Second edition. Chapman and Hall/CRC Press, Boca Raton, Florida, USA.
Wood, S. N. 2019. mgcv: Mixed GAM computation vehicle with automatic smoothness estimation. R package version 1.8-27. https://CRAN.R-project.ord/package=mgcv
Yee, T. W. 2015. Vector generalized and additive models: With an implementation in R. Springer, New York, New York, USA.
Zhang, J. T., and E. R. B. Oxley. 1994. A comparison of 3 methods of multivariate-analysis of upland grasslands in North Wales. Journal of Vegetation Science 5:71–76.

Supporting Information

Additional supporting information may be found in the online version of this article at http://onlinelibrary.wiley.com/doi/10.1002/ecy.2908/suppinfo