OIKOS

Research article

Dissimilarity analysis based on diffusion maps

Jordan A. Gault[1,2,3], **Jan A. Freund**[2], **Helmut Hillebrand**[1,2] **and Thilo Gross**[1,2,3]

[1]HIFMB, Helmholtz Institute for Functional Marine Biodiversity, Oldenburg, Germany
[2]Institute for Chemistry and Biology of the Marine Environment, University of Oldenburg, Oldenburg, Germany
[3]Helmholtz Center for Polar and Marine Research, Alfred-Wegener Institute, Bremerhaven, Germany

Correspondence: Jordan A. Gault (jordan.gault@uni-oldenburg.de)

Oikos
2023: e10249
doi: 10.1111/oik.10249
Subject Editor: Werner Ulrich
Editor-in-Chief: Dries Bonte
Accepted 26 July 2023

Compositional measurements from species assemblages define a high dimensional dataspace in which the data can form complex structures, termed manifolds. Comparing assemblages in this dataspace is difficult because the data is often sparse relative to its dimensionality and the complex structure of the manifold introduces bias and error in measurements of distance. Here, we apply diffusion maps, a manifold learning method, to find and characterize manifolds in high-dimensional compositional data. We show that diffusion maps embed the data in reduced dimensions in which the Euclidean distance between data points approximates the distance between them along the manifold. This is especially useful when species turnover is high, as it provides a way to measure meaningful distances between assemblages even when they harbor disjoint sets of species. We anticipate diffusion maps will therefore be particularly useful for characterizing community change over large spatial and temporal scales.

Keywords: biodiversity change, compositional dissimilarity, diffusion maps, dimensionality reduction

Introduction

Quantifying compositional variation between communities, or beta diversity, is a prerequisite to understanding the processes that govern their assembly and structure. Measures of compositional dissimilarity can be used to quantify beta diversity among a set of communities (Jost et al. 2010). To calculate compositional dissimilarity, the vectors of species abundances (or relative abundances) at two or more sites are considered as points in multivariate space and some measure of pairwise (dis)similarity is calculated between them. These pairwise distances are commonly the starting point for two analyses (Tuomisto and Ruokolainen 2006): First, they can be used as the input for ordination methods, which order the sites along major axes of variation in reduced dimensions. This provides a lower-dimensional representation of the data which can be used to understand the relationship between community composition and environmental and spatial factors (Faith et al. 1987, Tuomisto and Ruokolainen 2006, Legendre and Legendre 2012). Second, pairwise distances can be used as the response variable for regression analysis with the aim of predicting how different two

NORDIC SOCIETY OIKOS
www.oikosjournal.org

communities will be, based on the distance between them along environmental, spatial, or temporal gradients (Tuomisto and Ruokolainen 2006, Ferrier et al. 2007, Lichstein 2007, Anderson et al. 2011, Woolley et al. 2017).

Ordination and regression methods that utilize pairwise distances run into issues of distortion and saturation respectively when comparing communities harboring disjoint sets of species. For instance, consider an environmental gradient along which complete replacement of the species inventory occurs multiple times. Ordinating sites along such a so-called 'long-gradient' introduces distortions, the most severe of which is the horseshoe effect, wherein opposite ends of the gradient curve in towards one another in the ordination plot, erroneously showing that communities become more similar at opposite ends of the gradient (Podani and Miklós 2002, Legendre and Legendre 2012). Additionally, many measures of ecologically meaningful distance are formulated in terms of species overlap and are thus bounded between 0 and 1, where a distance of zero means sites are identical with respect to species identity and abundance and a distance of one means no species are shared. Once complete replacement of the species inventory occurs, such a distance will reach its maximum value of one after which no increase in compositional dissimilarity will be apparent despite continued compositional change along the gradient. Regression of pairwise distances between sites along such a gradient is therefore hampered by the saturation of the distance metric.

More fundamentally, the comparison of communities is difficult because compositional data can form complex structures in multivariate space, and what is needed is a way to visualize and measure distances along such structures. For example, consider a simple case where three species exhibit a uni-modal response to a single environmental gradient (Fig. 1a). Sampling species assemblages along the gradient and plotting the resulting relative abundances yields a three-dimensional curve, termed a 'manifold' (Fig. 1b). Because the environmental gradient constrains the species' abundances, the manifold represents the community compositions that can be realized. A natural measure of distance between communities, therefore, is the distance between them along the manifold (Orlóci 1975, Beals 1984, Barter and Gross 2019). However, calculating the distance directly between two points does not necessarily approximate their distance along the manifold. For instance, in Fig. 1b, the distance calculated directly between the communities at the beginning and end of the environmental gradient (red arrow) indicates that they are very similar. In fact, these two communities can only occur at opposite extremes of the environmental gradient, representing the traversal of the entire manifold (purple arrow).

An additional difficulty is that compositional data is often high-dimensional and sparse: we typically have a low number of samples relative to the dimensionality of the data space and the number of corners in the data space is even larger than the number of dimensions. Consider that an N-dimensional space has 2^N corners. For example, an assemblage with only a single species has two corners: species one can be abundant or rare. An assemblage with two species can exhibit the following four combinations, or corners, of species one and species two: abundant/abundant, abundant/rare, rare/abundant, rare/rare. So, already a 10-species data space has $2^{10} = 1024$ corners which is greater than the number of samples in many ecological datasets. However, because the environment constrains the possible combinations of species' abundances, compositional data often cluster around a manifold that is relatively low-dimensional compared to the data space. This means that practically obtainable ecological datasets may still contain enough information to characterize the manifold. Moreover, quantifying the position of a community on the manifold provides a lower-dimensional description of the

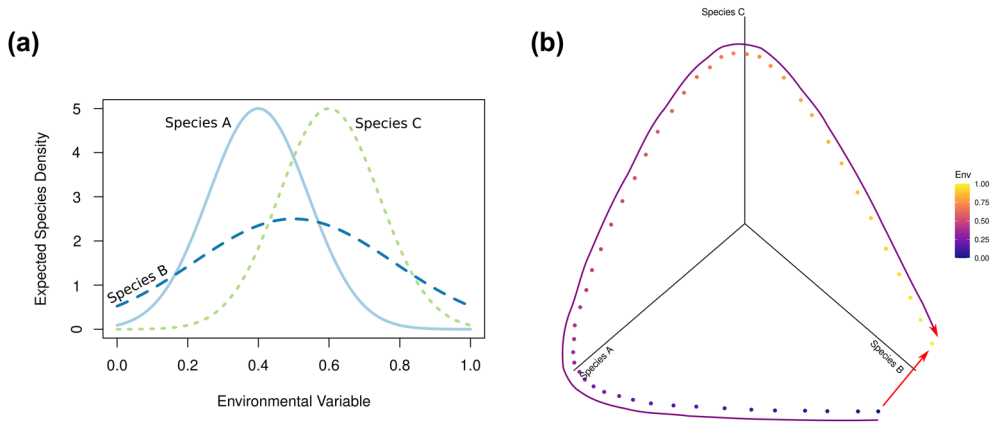

Figure 1. Example of how species responses to an environmental gradient cause compositional data points to cluster around a lower-dimensional manifold in the data space. (a) Response curves for three species to a single environmental gradient. Each curve defines the expected number of individuals sampled at a given value of the environmental gradient. (b) Plot of the relative abundance of the three species at 50 locations along the environmental gradient. Abundances were obtained as Poisson variates with expectation values defined by the response curves. Points are shaded by their location along the environmental gradient. The red arrow denotes the distance calculated directly between the communities at the beginning and end of the environmental gradient. The purple arrow denotes the distance between them along the manifold.

community that captures the main features of its composition. What is needed is a method that can reduce the dimensionality of the data and characterize the manifold on which the data lie.

Diffusion maps are a manifold learning method that find complex manifolds in high-dimensional data (Coifman et al. 2005). In doing so, diffusion maps provide a lower-dimensional embedding for the data in which the Euclidean distance between data points approximates the distance between them along the manifold. Crucially, diffusion maps rely only on comparisons between objects that are sufficiently similar, avoiding the issues associated with comparing very dissimilar objects.

Diffusion maps have previously been used to characterize the niche space of bacteria based on genomic data (Fahimipour and Gross 2020) and to infer functional traits of phytoplankton based on species associations estimated from monitoring data (Ryabov et al. 2022). In both cases, the focus was on individual species and their location in a multidimensional trait space defined by their functional characteristics, wherein the primary difficulty lies in comparing species with very different traits. When comparing ecological communities, the primary difficulty is instead comparing assemblages with very different species compositions.

Here, using simulated species assemblages, we show that diffusion maps can be used to embed high-dimensional compositional data in reduced dimensions, effectively capturing the main features of community composition in a few variables. Additionally, we demonstrate that this embedding can be used to calculate unbounded distances between data points, yielding good measures of compositional dissimilarity even when species turnover is high.

Diffusion maps

Diffusion maps use the notion of a diffusion process to explore the structure of multivariate data. But instead of a physical diffusion process, diffusion maps are inspired by the notion of a diffusion process on a network in the data space. Diffusion maps treat data points – in this case vectors of species total or relative abundances – as nodes in a network with weighted links defined by some measure of distance between them. Importantly, nodes are connected to only a small number of nearest neighbors such that links represent local distances. In practice, this is achieved by calculating an appropriately chosen measure of distance between all data points and then discarding distances over some threshold. We are left with a set of shorter, trusted distances which are treated as weighted links between data points, forming a network.

By modeling a random walk of particles on the thesholded network, diffusion maps essentially integrate over local distances to yield a global representation of the manifold. But rather than explicitly simulating a random walk, the structure of the network is explored using harmonic analysis (Coifman et al. 2005, Barter and Gross 2019). Specifically, the eigenvalues and eigenvectors of the network Laplacian are computed. The smallest, non-zero eigenvalues encode the direction of largest variation of the data (they span the manifold) and their associated eigenvectors define a new, lower dimensional coordinate system in which the data points are embedded. Within this new coordinate system, or 'diffusion map', the Euclidean distance between points approximates the 'diffusion distance' between them along the manifold, thus allowing the calculation of distance between data points that are very dissimilar.

The use of local versus global distances is the most important distinguishing factor between diffusion maps and other commonly used dimensionality reduction methods in ecology such as principal coordinates analysis (PCoA) and nonmetric multidimensional scaling (NMDS). To illustrate this difference, we diffusion mapped the data plotted in Fig. 1b. To construct the diffusion map (Fig. 2a), we first calculated the Horn distance (Horn 1966) between all pairs of sites (Supporting information). We then thresholded the distance matrix, keeping the two-nearest neighbors for each site. By integrating over these local distances, the diffusion map essentially spreads the manifold out over a single dimension containing the most variation and successfully recovers the order of the sites along the one-dimensional environmental gradient. Contrast this with the two-dimensional configuration produced by NMDS (Fig. 2b). Because NMDS seeks to preserve the rank order of the entire distance matrix, it is forced to place samples at the end of the gradient close to one another which misorders them along MDS1 (the same misordering results from using a one-dimensional configuration but is easier to see in two dimensions).

Within the diffusion map, the Euclidean distance between sites (i.e. the diffusion distance) now approximates the distance between them along the manifold, quantifying the amount of compositional change that is accumulated across the environmental gradient (Fig. 2c). In contrast, plotting the non-thresholded distances indicates that compositional dissimilarity increases and then decreases along the environmental gradient (Fig. 2d).

Evaluation with simulated and empirical data

We evaluated the ability of diffusion maps to capture the main features of multivariate compositional data using simulated species assemblages and an empirical dataset. First, we simulated species assemblages at 100 sites structured by two environmental gradients. Both gradients ranged from 0 to 1 such that the environmental space is represented by a unit square. The response of a given species to these environmental factors was modeled as a bivariate Gaussian response surface. We artificially generated 1000 such response surfaces whose location, width, and orientation were randomly chosen with respect to the environmental space (Fig. 3a–c, see the Supporting information for full details). Each response surface can be thought of as the expected density of a species

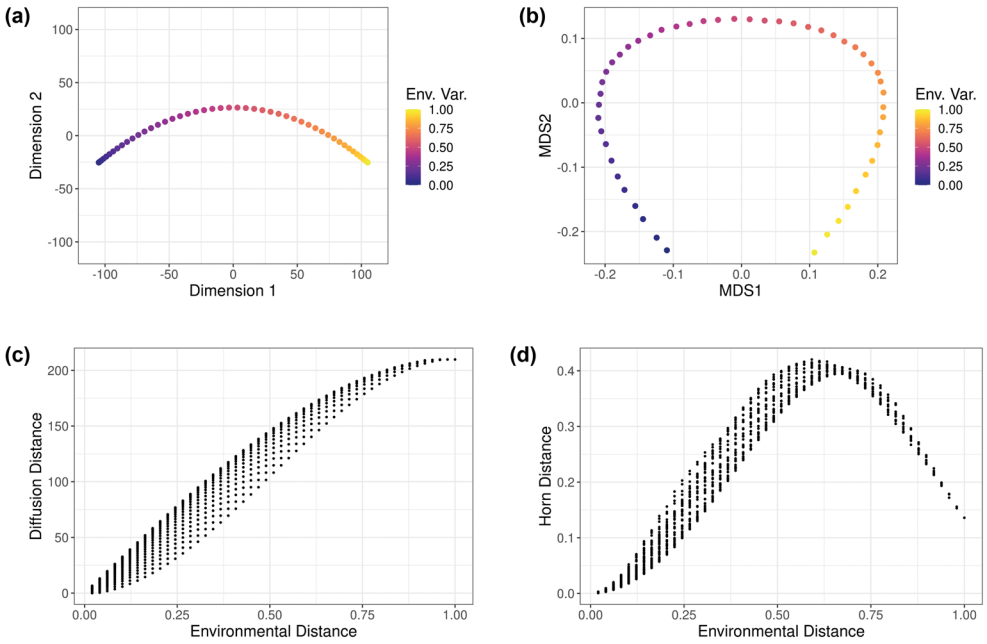

Figure 2. Example of how diffusion maps use local distances to characterize and calculate distances along a manifold. (a) Diffusion map and (b) two-dimensional NMDS configuration of the data plotted in Fig. 1b. By using local distances, the diffusion map 'unrolls' the manifold in a single dimension, correctly ordering sites along the environmental gradient. In contrast, the use of global distances forces NMDS to place sites at opposite ends of the environmental gradient close together, misordering sites along the environmental gradient. (c) Calculating the Euclidean distance between sites within the diffusion map correctly indicates that compositional dissimilarities accumulates along the environmental gradient. (d) Plotting global distances incorrectly indicates that compositional dissimilarity does not accumulate along the environmental gradient.

with respect to a given pair of environmental values. To generate assemblages, we chose 100 locations, or sites, on the environmental space as an evenly spaced 10 × 10 grid on the interval of environmental values from 0.1 to 0.9 (Fig. 3a–c). For a given site, we took the value of each response surface as the expectation value of a Poisson sampling process. Poisson variates were thus generated for each species to yield a vector of sampled abundances for each site on the environmental space. Each vector was transformed to proportional abundance. To account for sampling error, we repeated this sampling process 100 times, yielding 100 different realizations for each site. All analysis were performed and results summarized over these 100 realizations.

The amount of turnover along the gradients was tuned by adjusting the width of the species response surfaces. We considered three scenarios: a high turnover scenario where ≈ 40% of site pairs harbor disjoint sets of species (Fig. 3d), a low turnover scenario where no site pairs harbor disjoint sets of species (Fig. 3g), and a variable turnover scenario where ≈ 25% of site pairs harbor disjoint sets of species (Fig. 3a). In the variable turnover scenario, turnover and species richness varied along environmental variable one, with higher-turnover/lower-richness from 0 to 0.5 and lower-turnover/higher-richness from 0.5 to 1. Average species richness was 158 (SD = 20) for the high turnover scenario, 532 (SD = 104) for the low turnover scenario, and 179 (SD = 60) for the variable turnover scenario.

Diffusion maps of the simulated assemblages were constructed by first calculating the Horn index of overlap (Horn 1966) between all pairs of sites. The similarity matrix was then thresholded by retaining only the ten nearest neighbors for each site. See the Supporting information for full details on construction of the diffusion maps.

To evaluate the ability of diffusion maps to capture the main features of compositional change across the environmental gradients, we asked three questions: first, we asked if the diffusion maps successfully identify the main dimensions of variation in the data; second, we asked if the location of the sites in the diffusion maps correspond with their location along the environmental gradients, correctly recovering the underlying structure of the data; and third, we asked if the diffusion distance successfully captures the relationship between environmental distance and compositional dissimilarity. Additionally, we compared the results obtained using diffusion maps to a more traditional workflow using a bounded measure of dissimilarity and two commonly used dimensionality reduction techniques: principal coordinates analysis (PCoA) and non-metric multidimensional scaling (NMDS). All analyses were carried out in the R programming language (www.r-project.org).

To determine whether diffusion maps successfully identify the main dimensions of variation, we plotted the inverse of the ranked non-zero eigenvalues. The inverse eigenvalues should be larger for meaningful dimensions of variation and relatively small for non-meaningful dimensions of variation. Because the composition of the simulated assemblages is determined entirely by the two environmental gradients, the

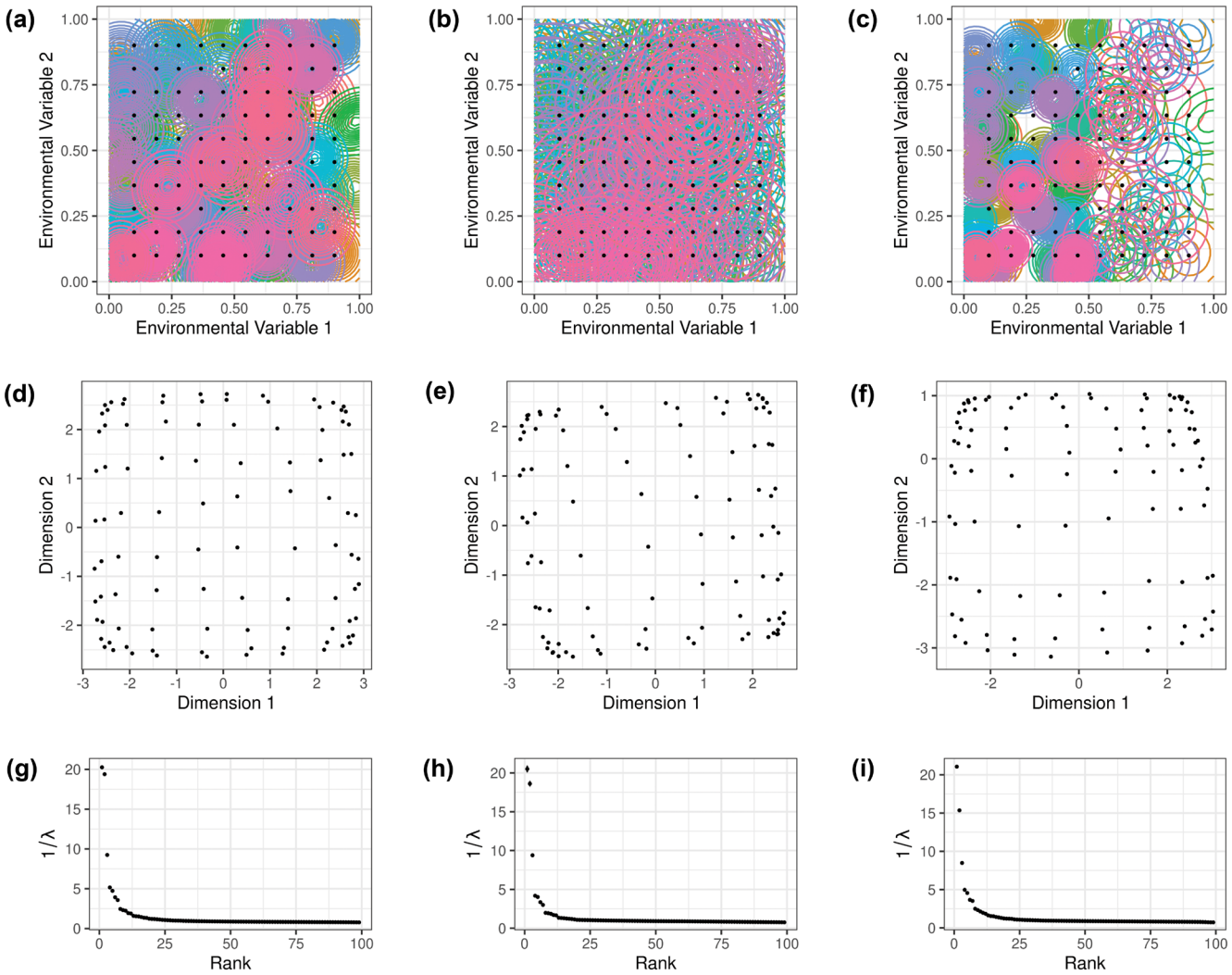

Figure 3. Species response surfaces and sampled sites, selected diffusion maps, and eigenvalue spectrum for all three turnover scenarios. Response surfaces for 100 species in relation to two environmental gradients for the (a) high turnover, (b) low turnover scenarios, and (c) variable turnover scenarios. Each color denotes a separate species and points represent the 100 sampled sites. Only 100 of the 1000 species are shown for visual clarity. Diffusion map of a single realization of the (d) high turnover, (e) low turnover and (f) variable turnover scenarios. Ranked inverse eigenvalue (1/λ) spectrum summarized (mean and standard deviation) across all 100 realizations of the (g) high turnover, (h) low turnover and (i) variable turnover scenarios.

two largest non-zero inverse eigenvalues should identify the major axes of variation in the data. The eigenvectors associated with these two eigenvalues provide a new coordinate system for the data, i.e. the diffusion map.

To determine if the diffusion maps successfully recover the underlying structure of the data with respect to the environmental gradients, we compared the matrix of diffusion map coordinates with the matrix of site coordinates in the environmental space (equivalent to the environmental values at each site) via the Procrustes test. The Procrustes test measures the amount of discordance between two matrices after they have been scaled and rotated to maximize their superposition. Specifically, we computed the sum of squares error, m^2, of a symmetric Procrustes test (Peres-Neto and Jackson 2001). Larger values of m^2 correspond to larger discordance between the two matrices. The Procrustes m^2 was calculated using the 'vegan' package (Oksanen et al. 2020).

To determine if the diffusion distance successfully captures the relationship between environmental distance and compositional change, we calculated both linear and ordinal correlations between the pairwise diffusion distances and pairwise environmental distances. We calculated the pairwise environmental distance between all sites as the Euclidean distance between the locations of the sites in the environmental space.

Using the simulated assemblages, we also compared the diffusion map method to more traditional methods of analysis based on a bounded measure of dissimilarity: the Horn dissimilarity. The Horn dissimilarity can be calculated as the one complement of the Horn similarity used to construct the

diffusion maps. First, we compared the ability of diffusion maps to recover the underlying structure of the data with two dimensionality reduction techniques commonly used in ecology: PCoA and NMDS. Starting with the Horn dissimilarity, we used PCoA and NMDS to project the data into two dimensions. PCoA was performed using R package 'ape' (www.r-project.org, Paradis and Schliep 2019) and NMDS was performed using package 'vegan' (Oksanen et al. 2020). We compared the resulting ordinations with the matrix of site coordinates in the environmental space using the Procrustes m^2. Second, we computed the linear and ordinal correlations between the Horn dissimilarity and the environmental distance.

Finally, we applied diffusion maps to a previously well-characterized empirical dataset: fish assemblages along the Doubs river (Verneaux 1973, Verneaux et al. 2003). The Doubs river and its tributaries span an 832 km network near the France–Switzerland border in the Jura mountains. Fish and insect communities have previously been shown to exhibit continuous compositional change along an upstream–downstream gradient in the river system (Verneaux et al. 2003). This dataset therefore presents an ideal test for diffusion maps ability to recover the main dimensions of variation in compositional data.

We used a subset of the data from Verneaux (1973), included in the R package 'ade4' (www.r-project.org, Thioulouse et al. 2018), which contains the abundance classes for 27 species sampled from 29 sites along the main course of the Doubs river. Because the abundance classes represent course-grained information on the relative abundance of each species, we transformed them to relative abundance and measured community overlap via the Horn index. We retained the 10 nearest neighbors for each sample to construct the diffusion map. For comparison, we also used NMDS to find a two-dimensional configuration for the data using the Horn dissimilarity. From the diffusion map, we calculated the diffusion distance and compared to the full matrix of Horn dissimilarities. Finally, to demonstrate the effect of retaining different numbers of nearest neighbors, we constructed diffusion maps retaining 2, 5, 15 and 29 nearest neighbors.

Results

The diffusion maps successfully capture the main dimensions of variation in the simulated data set for all three turnover scenarios (Fig. 3g–i). The inverse eigenvalue spectrum for all three scenarios reveals a substantial gap between the first two non-zero eigenvalues and the remaining eigenvalues. The relative magnitude of each inverse eigenvalue is proportional to the magnitude of variation contained in the corresponding dimension. The large gap separating the first two inverse eigenvalues from the rest indicates that two dimensions contain most of the variation in the data. Additionally, the eigenvalue spectrum for the variable turnover scenario indicates that one dimension contains higher variation, correctly capturing the reduced amount of turnover along the first environmental gradient relative to the second environmental gradient (Fig. 3i). The two leading inverse eigenvalues and their associated eigenvectors correspond to the two environmental gradients and define the new coordinate system into which the data points can be embedded. Figure 3d–f show this embedding, or diffusion map, for a single realization of the high, low, and variable turnover scenarios, respectively.

The diffusion maps successfully recover the underlying structure of the data with respect to the two environmental gradients for all three scenarios. Visual comparison of the diffusion maps (Fig. 3d–f) with the original sampling grid (Fig. 3a–c) show that they successfully recover the positions of the sites in the environmental space, albeit with some distortion. Note that the orientation of a given diffusion map is arbitrary in the sense that any dimension can be flipped by multiplying the corresponding eigenvector by − 1 without changing the interpretation (in fact, when solving for eigenvectors, the value and sign of the first entry is arbitrary).

Table 1 summarizes the Procrustes m^2 for each of the dimensionality reduction methods – diffusion mapping, PCoA, and NMDS – across all 100 realizations for the high, low, and variable turnover scenarios. For the high turnover scenario the diffusion map (mean = 0.0287, SD = 0.0006) outperformed PCoA (mean = 0.2247, SD = 0.0006). For the low turnover scenario the diffusion map (mean = 0.0471, SD = 0.0020) performed similarly to PCoA (mean = 0.0484, SD = 0.0001). NMDS outperformed diffusion maps and PCoA for the high turnover (mean = 0.0003, SD =), low turnover (mean = 0.0009, SD = 0), and variable turnover (mean = 0.0804, SD = 0.0015) scenarios. Visual inspection of selected Procrustes plots highlights additional features (Fig. 4). For the high turnover scenario, PCoA misorders sites (Fig. 4d) as shown by the multiple criss-crossed arrows. Diffusion maps, on the other hand, do not misorder sites (Fig. 4a). In the low turnover scenario, PCoA orders sites more successfully (Fig. 4e) but shows more distortion of the sampled grid compared to the diffusion map (Fig. 4b). In the variable turnover scenario, PCoA misorders sites and shows a large degree of distortion (Fig. 4f). In contrast, the diffusion

Table 1. Summary of the Procrustes sum of squares error m^2 for all 100 realizations of the high, low, and variable turnover scenarios for diffusion maps, PCoA ordinations, and NMDS ordinations

	High turnover		Low turnover		Variable turnover	
	Mean	SD	Mean	SD	Mean	SD
Diffusion map	0.0287	0.0006	0.0471	0.0020	0.0870	0.0010
PCoA	0.2247	0.0006	0.0484	0.0001	0.3437	0.0006
NMDS	0.0003	0.0000	0.0009	0.0000	0.0804	0.0015

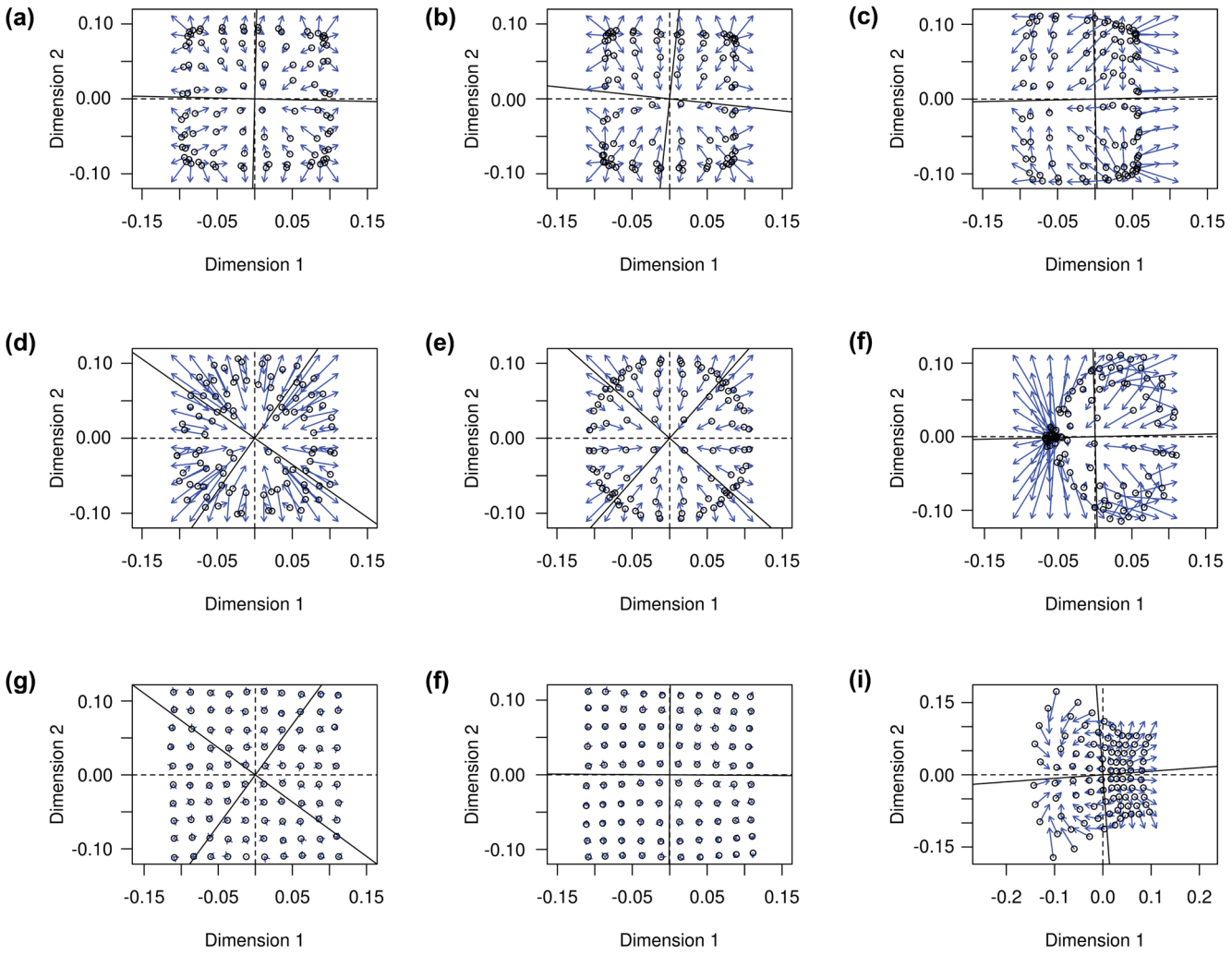

Figure 4. Procrustes errors for diffusion maps, PCoA ordinations, and NMDS ordinations (rows) from a single realization of the high, low, and variable turnover scenarios (columns). Blue arrows point from sites plotted in reduced dimensions to the target matrix of environmental values. Diffusion maps recover the structure qualitatively in the (a) high turnover and (b) low turnover, and (c) variable turnover scenarios. PCoA shows large errors in the (d) high turnover and (f) variable turnover scenarios. PCoA qualitatively recovers the structure in the (e) low turnover scenario but with large distortion of the grid. NMDS recovers the structure in the (g) high turnover, (h) low turnover and (i) variable turnover scenarios.

map correctly recovers the sampled grid (Fig. 4c) In the high, low, and variable turnover scenarios, NMDS successfully recovers the structure of the sampling scheme (Fig. 4g–i).

Diffusion maps successfully recover the relationship between environmental distance and compositional dissimilarity for all three turnover scenarios. For the high turnover scenario, the diffusion distance has higher linear correlation (mean = 0.9512, SD = 0.0009) with environmental distance than the Horn distance (mean = 0.6298, SD = 0.0002) and similar ordinal correlation (mean = 0.9594, SD = 0.0009) as the Horn distance (mean = 0.9631, SD = 0.0008; summarized in Table 2). For the variable turnover scenario, the diffusion distance has higher linear correlation (mean = .8786, SD = 0.0010) with environmental distance

Table 2. Linear and ordinal correlations of the Horn and diffusion distances with environmental distance for the high, low, and variable turnover scenarios

		High turnover		Low turnover		Variable turnover	
		Mean	SD	Mean	SD	Mean	SD
Diffusion distance	Pearson *r*	0.9512	0.0009	0.9230	0.0028	0.8786	0.0010
	Spearman ρ	0.9594	0.0009	0.9351	0.0038	0.8810	0.0011
Horn distance	Pearson *r*	0.6298	0.0002	0.9443	0.0001	0.6379	0.0002
	Spearman ρ	0.9631	0.0008	0.9974	0.0000	0.8718	0.0021

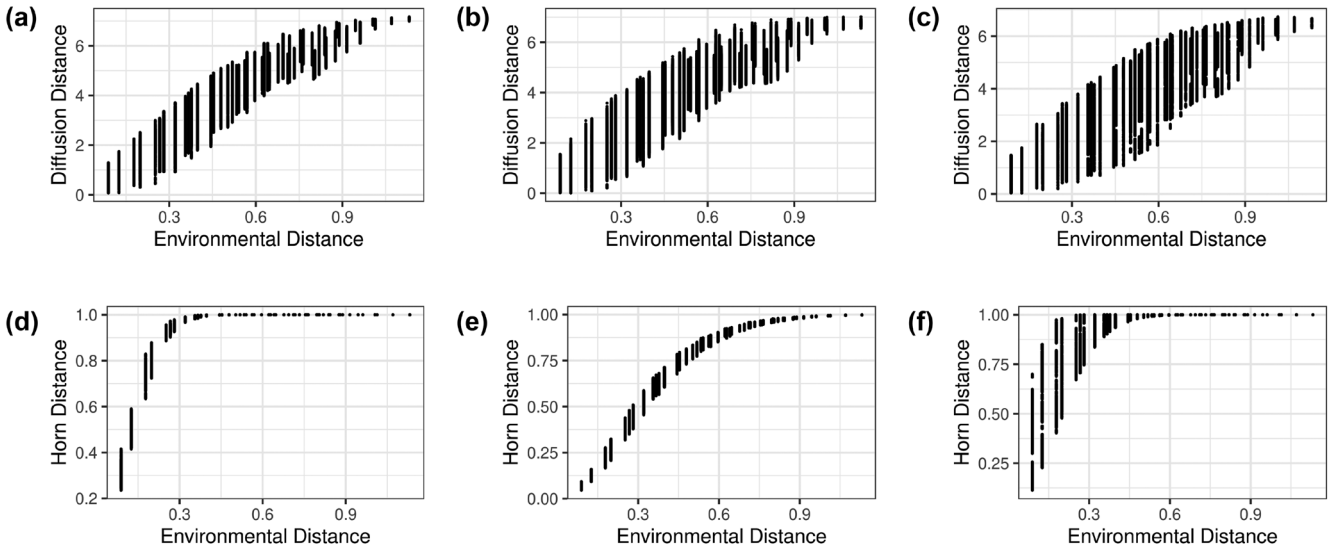

Figure 5. Diffusion and Horn distance in relation to environmental distance for the high, low, and variable turnover scenarios. The diffusion distance increases with environmental distance even after complete replacement of the species inventory for both the (a) high turnover and (c) variable turnover scenarios. The Horn distance does not increase with environmental distance after complete replacement of the species inventory for both the (d) high turnover and (f) low turnover scenarios. The diffusion distance (b) and Horn distance (e) both capture the relationship between compositional dissimilarity and environmental distance for the low turnover scenario.

than the Horn distance (mean = .6379, SD = 0.0002) and similar ordinal correlation (mean = 0.8810, SD = 0.0011) as the Horn distance (mean = 0.8718, SD = 0.0021). For the low turnover scenario, the diffusion distance has similar linear (mean = 0.9230, SD = 0.0028) correlation with environmental distance as the Horn distance (mean = 0.9443, SD = 0.0001) and similar ordinal correlation (mean = 0.9351, SD = 0.0038) as the Horn distance (mean = 0.9974, SD = 0).

Plotting the Horn distance against environmental distance for the high and variable turnover scenarios shows that the Horn distance quickly reaches the maximum value of 1, after which no more accumulation of compositional dissimilarity is detected (Fig. 5d, f). In contrast, the diffusion distance shows that compositional dissimilarity increases as environmental distance increases for the high and variable turnover scenarios (Fig. 5a, c). For the low turnover scenario, both the diffusion distance and Horn distance capture the increase in compositional dissimilarity with increasing environmental distance (Fig. 5b, e).

Applying diffusion maps to the Doubs dataset, we successfully recovered an upstream–downstream gradient in compositional change. Figure 6a shows the locations of the samples along the course of the river shaded by distance from the river source. The eigenvalue spectrum indicates that nearly all of the compositional variation can be captured in a single dimension (Fig. 6b). Plotting the diffusion map and shading the points by distance from source reveals that this dimension corresponds closely with position along the river (Fig. 6c). Additionally, the diffusion map indicates potential clustering within the upper and lower halves of the river which may correspond with the salmonid and cyprinid regions previously identified by Verneaux et al. (2003). While an upstream–downstream pattern of compositional change is evident in the two-dimensional NMDS configuration, it does not as clearly indicate how strongly community composition is shaped by a single gradient.

Plotting the diffusion and Horn distances against distance along the river reveals a pattern of increasing compositional dissimilarity with increasing distance between sites (Fig. 6e– f). The Horn distance shows that an increasing number of site-pairs show complete species replacement above an inter-site distance of 150 km (Fig. 6f). However, because the Horn distance reaches a maximum value of one, it is unclear whether compositional dissimilarity reaches a true maximum after 150 km or if it continues to increase. The diffusion distance shows that compositional dissimilarity does in fact asymptote around 150 km, after which the maximum compositional dissimilarity does not continue to increase (Fig. 6e).

Diffusion mapping the Doubs data while retaining increasing numbers of nearest neighbors for each node highlights some considerations that should be taken into account when performing the analysis. If too few nearest neighbors are retained, the network of data points is fragmented into separate components. This is easily detected because eigendecomposition will yield more than one zero eigenvalue if the network has been fragmented. Figure 7a illustrates the consequence of fragmenting the network: several sites form a separate component and are all given the same coordinate of (0, 0). The number of neighbors that fragments the network represents a lower bound on the choice of how many to retain. The diffusion map constructed using five nearest neighbors (Fig. 7b) already closely resembles the diffusion map constructed using 10 nearest neighbors (Fig. 6c).

If too many nearest neighbors are retained, the diffusion map will not be able to 'unroll' the manifold because it will

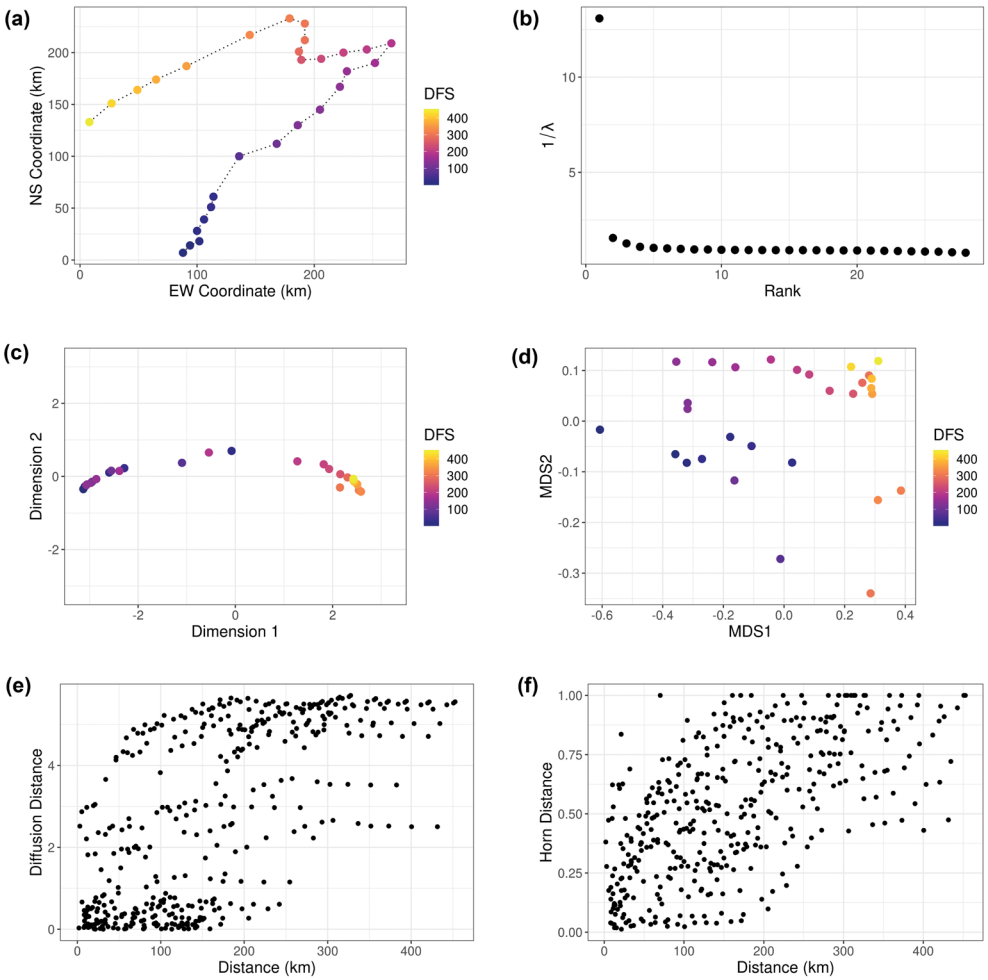

Figure 6. Diffusion map analysis of the Doubs fish assemblages. (a) Geographical location of the sites along the Doubs river. Sites are shaded by distance from source (DFS). (b) Eigenvalue spectrum indicates that a single dimension is sufficient to characterize the data. (c) The diffusion map indicates a strong upstream–downstream gradient in compositional change which is captured in a single dimension. (d) The two-dimensional NMDS configuration of the compositional data does not capture how strongly the upstream–downstream gradient structures the data as effectively as the diffusion map. (e) the diffusion distance captures the compositional turnover that accumulates along the single dimensional manifold and agrees closely with (f) the Horn distance calculated between all site-pairs. Unlike the Horn distance, the diffusion distance captures additional compositional change even after complete species replacement.

be constrained by long-range comparisons. Retaining 15 nearest neighbors, the diffusion map (Fig. 7c) already begins to resemble the two-dimensional NMDS configuration (Fig. 6d). When the full similarity matrix is used to construct the diffusion map (Fig. 7d), it is nearly identical to the NMDS configuration (Fig. 6d).

Discussion

Two distinct but closely related approaches can be taken to understand how beta diversity is related to environmental and spatial gradients (Tuomisto and Ruokolainen 2006). First, we might wish to know how community composition changes in response to environmental factors or geographical location (i.e. will a community have a specific composition for a specific set of environmental values or at a specific location). Second, we might wish to know how dissimilar two communities will be based on how environmentally dissimilar or geographically distant they are (i.e. how much compositional change can we expect for a given change in environmental conditions or geographic distance).

When seeking to understand how community composition changes in response to environmental factors or geographical location, it is often desirable to reduce the dimensionality of the data, summarizing the main features of community composition in a reduced number of variables. We show that diffusion maps can find and characterize manifolds in high dimensional compositional data and that the location of samples on the manifold serves as a lower-dimensional characterization of the community composition. In the simulated assemblages, the manifold on which the data lie is a two-dimensional plane defined by the two environmental factors that structure the communities. We show that

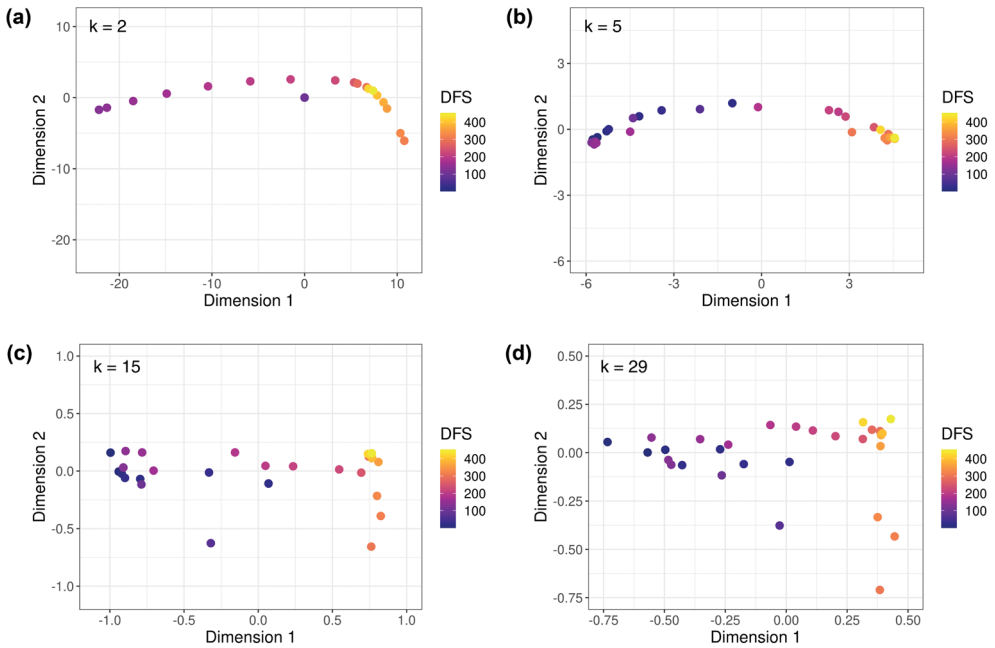

Figure 7. The number of nearest neighbors retained (k) for each node affects the ability of diffusion maps to characterize the manifold. (a) Retaining too few (k = 2) nearest neighbors fragments the network into separate components, evidenced here by the multiple sites placed at the origin. (b) A modest increase in the number of nearest neighbors (k = 5) yields a single network component while excluding long-range comparisons. (c) At k = 15, over half of the original distance matrix is retained and long-range comparisons begin to bias the diffusion map. (d) When the distance matrix is not thresholded at all (k = 29), long-range comparisons hinder the ability of the diffusion map to characterize the manifold and it closely resembles the two-dimensional NMDS configuration in Fig. 6d.

diffusion maps correctly identify the dimensionality of the manifold and recover the location of the communities in the environmental space. Because the coordinates of each community in the diffusion map correspond to their location in the environmental space, these coordinates serve as a compact description of community composition and its relationship to the two environmental gradients.

Principal coordinates analysis (PCoA) and non-metric multidimensional scaling (NMDS) are two commonly used methods for reducing the dimensionality of compositional data. Using simulated assemblages, we compared diffusion maps with the ability of these methods to correctly recover the underlying structure of the data and found that diffusion maps consistently outperform PCoA, especially when turnover is high. Diffusion maps and NMDS both correctly recover the structure of the data, but NMDS consistently performed better as measured by the Procrustes sum of squares error, m^2. This is not surprising given the simple underlying structure of the simulated data and the logic behind NMDS. NMDS seeks a lower-dimensional configuration of the data in which the rank order of pairwise distances is as close as possible to the original dissimilarity matrix. The ability of NMDS to optimally recover the location of the simulated communities in the environmental space indicates that compositional change along the two environmental gradients is accurately represented by the full dissimilarity matrix. Despite this, diffusion maps are still able to recover the underlying structure of the data nearly as well as NMDS using only local distances.

Applying diffusion maps to the Doubs dataset demonstrates the advantage of using local distances to characterize the structure of more complex data. Compared to the NMDS configuration, the diffusion map more clearly indicates that community composition is changing along a single dimension which corresponds well with position along the river. While the pattern of compositional change along the course of the river is somewhat evident along the first dimension of the NMDS configuration, the diffusion map more effectively reveals the structure of the data by considering a thresholded subset of the original distance matrix. As an increasing number of long range comparisons are retained, the diffusion map increasingly resembles the NMDS configuration, demonstrating how integrating over local distances allows the diffusion map to essentially 'unroll' the manifold along the dimensions of greatest variation.

Most importantly, by embedding the manifold on which the data lie in a lower number of dimensions, diffusion maps provide a way to calculate meaningful distances between data points along the manifold. Using simulated assemblages, we show that the diffusion distance recovers the relationship between environmental distance and community dissimilarity for all three turnover scenarios. This is particularly useful when species turnover is high because diffusion maps can be used to make meaningful comparisons between communities

along environmental or spatial gradients even when complete replacement of the species inventory occurs multiple times. However, we also show that diffusion maps are effective when turnover is low. Compositional data can form complex manifolds even in the absence of complete species turnover, in which case raw dissimilarities may not reflect the distance between sites on the manifold (for example Fig. 1, 2). By integrating over local distances, diffusion maps essentially quantify the amount of compositional change that accumulates across the manifold. The diffusion distance can thus be used to quantify the relationship between compositional dissimilarity and environmental/geographical distance over a range of scales.

The only tune-able parameter in diffusion maps is the number of nearest neighbors retained in the similarity matrix. The choice of number of nearest neighbors represents a balancing act between choosing too few and fragmenting the network, and choosing too many thereby hindering the ability of diffusion maps to effectively characterize the manifold. A value of 10 nearest neighbors has performed well across a number of applications and can be recommended as a good rule of thumb (Barter and Gross 2019, Fahimipour and Gross 2020, Ryabov et al. 2022). Retaining 10 nearest neighbors worked well even for the relatively small Doubs dataset, and larger data sets should be less sensitive to the choice of number of nearest neighbors (Ryabov et al. 2022). However, the robustness of the resulting diffusion map to the choice of number of nearest neighbors should be checked for any analysis.

It is important to note that unlike PCoA and NMDS, diffusion maps are not a method for ordination or visualization per se. PCoA and NMDS are typically used to represent high dimensional data in a visually interpretable number of dimensions. Diffusion maps, on the other hand, find lower-dimensional manifolds in high dimensional data. These manifolds may themselves be high dimensional and therefore not easily visualized (Moon et al. 2019). For example, diffusion maps may find a 10-dimensional manifold in 1000-dimensional data. In this case, further dimensionality reduction may be needed to visually ordinate the data. Alternatively, methods such as PHATE (Moon et al. 2019) could be used instead for visualization of the manifold. However, in cases where diffusion maps find a relatively low number of important dimensions, the embedding can be plotted directly to yield interpretable visualizations.

Quantifying community changes at regional levels is difficult because many sites harbor disjoint sets of species (Ferrier et al. 2007). We envision that diffusion maps will be particularly useful when comparing communities across large spatial scales that encompass very different environmental conditions and species assemblages. In this case, the diffusion distance could be used as a response in models that predict the amount of compositional change across large spatial or environmental gradients. Additionally, the diffusion map embedding could be used to visually map community types across these gradients or to predict community types based on environmental and spatial variables. Moreover, there is growing recognition of the need for more comprehensive surveys of ecological communities at local, regional and global scales. Therefore, the size and complexity of ecological data sets will continue to grow with a concordant need for methods that can be used to analyze them. Diffusion maps show promise as an effective way to understand the structure of complex ecological data sets at a range of scales.

Funding – HH acknowledges funding by the German Federal Ministry for Education and Research and the Belmont Forum for the MARISCO project (Award 03F0836A). HH, JAF and JAG acknowledge funding from the Helmholtz-Incubator pilot project Uncertainty Quantification (Award ZT-I-0029).

Author contributions

Jordan A. Gault: Conceptualization (equal); Data curation (lead); Formal analysis (lead); Investigation (lead); Methodology (equal); Software (lead); Validation (lead); Visualization (lead); Writing – original draft (lead); Writing – review and editing (lead). **Jan A. Freund**: Conceptualization (equal); Funding acquisition (equal); Project administration (equal); Supervision (equal); Writing – review and editing (equal). **Helmut Hillebrand**: Conceptualization (supporting); Funding acquisition (equal); Project administration (equal); Supervision (equal); Writing – review and editing (supporting). **Thilo Gross**: Conceptualization (equal); Methodology (equal); Supervision (equal); Writing – review and editing (equal).

Data availability statement

Data are available from the Zenodo Digital Repository: http://10.5281/zenodo.8252502 (Gault et al. 2023).

Supporting information

The Supporting information associated with this article is available with the online version.

References

Anderson, M. J., Crist, T. O., Chase, J. M., Vellend, M., Inouye, B. D., Freestone, A. L., Sanders, N. J., Cornell, H. V., Comita, L. S., Davies, K. F., Harrison, S. P., Kraft, N. J. B., Stegen, J. C., Swenson, N. G. 2011. Navigating the multiple meanings of β diversity: a roadmap for the practicing ecologist. – Ecol. Lett. 14: 19–28.

Barter, E. and Gross, T. 2019. Manifold cities: social variables of urban areas in the UK. – Proc. R. Soc. A 475: 20180615.

Beals, E. W. 1984. Bray-Curtis ordination: an effective strategy for analysis of multivariate ecological data. – Adv. Ecol. Res., pp. 1–55.

Coifman, R. R., Lafon, S., Lee, A. B., Maggioni, M., Nadler, B., Warner, F. and Zucker, S. W. 2005. Geometric diffusions as a tool for harmonic analysis and structure definition of data: diffusion maps. – Proc. Natl Acad. Sci. USA. 102: 7426–7431.

Fahimipour, A. K. and Gross, T. 2020. Mapping the bacterial metabolic niche space. – Nat. Commun. 11: 4887.
Faith, D. P., Minchin, P. R. and Belbin, L. 1987. Compositional dissimilarity as a robust measure of ecological distance. – Vegetatio 69: 57–68.
Ferrier, S., Manion, G., Elith, J. and Richardson, K. 2007. Using generalized dissimilarity modelling to analyse and predict patterns of beta diversity in regional biodiversity assessment. – Divers. Distrib. 13: 252–264.
Gault, J. A., Freund, J. A., Hillebrand, H. and Gross, T. 2023. Data from: Dissimilarity analysis based on diffusion maps. – Zenodo Digital Repository. http://10.5281/zenodo.8252502.
Horn, H. S. 1966. Measurement of "overlap" in comparative ecological studies. – Am. Nat. 100: 419–424.
Jost, L., Chao, A. and Chazdon, R. L. 2010. Compositional similarity and β (beta) diversity. – In: Magurran, A. E. and McGill, B. J. (eds), Biological diversity: frontiers in measurement and assessment. Oxford Univ. Press, pp. 66–84.
Legendre, P. and Legendre, L. 2012. Numerical ecology. – Elsevier.
Lichstein, J. W. 2007. Multiple regression on distance matrices: a multivariate spatial analysis tool. – Plant Ecol. 188: 117–131.
Moon, K. R., van Dijk, D., Wang, Z., Gigante, S., Burkhardt, D. B., Chen, W. S., Yim, K., Elzen, A. V. D., Hirn, M. J., Coifman, R. R., Ivanova, N. B., Wolf, G. and Krishnaswamy, S. 2019. Visualizing structure and transitions in high-dimensional biological data. – Nat. Biotechnol. 37: 1482–1492.
Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., Szoecs, E. and Wagner, H. 2020. Vegan: community ecology package. – R package ver. 2.5-7, https://CRAN.Rproject.org/package=vegan.
Orlóci, L. 1975. Multivariate analysis in vegetation research. – Springer.
Paradis, E. and Schliep, K. 2019. Ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. – Bioinformatics 35: 526–528.
Peres-Neto, P. R. and Jackson, D. A. 2001. How well do multivariate data sets match? the advantages of a procrustean superimposition approach over the Mantel test. – Oecologia 129: 169–178.
Podani, J. and Miklós, I. 2002. Resemblance coefficients and the horseshoe effect in principal coordinates analysis. – Ecology 83: 3331–3343.
Ryabov, A., Blasius, B., Hillebrand, H., Olenina, I., and Gross, T. 2022. Estimation of functional diversity and species traits from ecological monitoring data. – Proc. Natl Acad. Sci. USA 119: e2118156119.
Thioulouse, J., Dray, S., Dufour, A.-B., Siberchicot, A., Jombart, T. and Pavoine, S. 2018. Multivariate analysis of ecological data with ade4. – Springer.
Tuomisto, H. and Ruokolainen, K. 2006. Analyzing or explaining beta diversity? understanding the targets of different methods of analysis. – Ecology 87: 2697–2708.
Verneaux, J. 1973. Cours d'eau de Franche-Comt´e (Massif du Jura). recherches ´ecologiques sur le r´eseau hydrographique du Doubs. – Essai de biotypologie. – Ann. Sci. l'Universit´e de Franche-Comt´e 3: 1–260.
Verneaux, J., Schmitt, A., Verneaux, V. and Prouteau, C. 2003. Benthic insects andfish of the Doubs river system: typological traits and the development of a species continuum in a theoretically extrapolated watercourse. – Hydrobiologia 490: 63–74.
Woolley, S. N. C., Foster, S. D., O'Hara, T. D., Wintle, B. A. and Dunstan, P. K. 2017. Characterising uncertainty in generalised dissimilarity models. – Methods Ecol. Evol. 8: 985–995.