Ecological Informatics 9 (2012) 1–10

Contents lists available at SciVerse ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf

Analyzing floristic inventories with multiple maps

Laurens van der Maaten [a,*], Sebastian Schmidtlein [b], Miguel D. Mahecha [c]

[a] Pattern Recognition and Bioinformatics Laboratory, Delft University of Technology, Mekelweg 4, 2628 CD Delft, The Netherlands
[b] Vegetation Geography, Institute of Geography, University of Bonn, 53115 Bonn, Germany
[c] Biogeochemical Model-Data Integration Group, Max Planck Institute for Biogeochemistry, Hans-Knöll-Str. 10, 07745 Jena, Germany

ARTICLE INFO

Article history:
Received 9 December 2011
Received in revised form 30 January 2012
Accepted 30 January 2012
Available online 8 February 2012

Keywords:
Species occurrence modeling
Data visualization
Multi-dimensional scaling
Non-metric similarities
t-Distributed Stochastic Neighbor Embedding

ABSTRACT

Spatial observations of plant occurrences contain a wealth of information on relations among species and on the relation between species and environmental conditions. Typically, inventory data of this kind are large co-occurrence matrices, and hence, direct ecological interpretations based on expert knowledge are often very difficult. Hitherto, ordination approaches have been used to construct a virtual ordination space (represented as one or multiple scatter plots) in which species that often co-occur are situated close together, whereas species that hardly co-occur are found far apart. In this study, we investigate a recently proposed ordination approach, *multiple maps t-SNE*, that constructs multiple, independent ordination spaces in order to reveal and visualize complementary structure in the data. We compare multiple maps t-SNE to several conventional ordination approaches, exploring a large inventory of vascular plant occurrences (FLORKART). Our results reveal that multiple maps t-SNE is well suited for the analysis of floristic inventories. In particular, multiple maps t-SNE uncovers the major dependencies of species co-occurrences on climate and soil biogeochemical preconditions.

1. Introduction

The construction of large floristic and vegetation databases over the course of the last century opened novel possibilities for regional, national, and continental biogeographical assessments (Bekker et al., 2007; Chytry and Rafajová, 2003; Jalas et al., 1972–1999; Mucina et al., 1993; Myklestad and Birks, 1993). Recent efforts bring together various sources of observation and allow even for global analyses (e.g., Scholes et al., 2008, see also the open access "global biodiversity information facility" http://data.gbif.org). Data collections of this kind allow scientists to address plant ecological questions across ecosystems (Kühn et al., 2004; Schmidtlein, 2004). Extracting the underlying environmental patterns from such data collections requires computationally powerful explorative multivariate tools (Mahecha and Schmidtlein, 2008). Indeed, the application of such tools has a long history in vegetation sciences, and a series of influential textbooks were published more than a decade ago (e.g., Gauch, 1982; Legendre and Legendre, 1998; ter Braak, 1995). Since then, the use of ordination techniques such as principal component analysis (PCA), detrended correspondence analysis (DCA), classical multi-dimensional scaling (CMDS), or non-metric multi-dimensional scaling (NMDS) has become a standard. Nonetheless, novel developments in the field of dimensionality reduction and machine learning in general are largely ignored or only timidly transferred from informatics to other fields (Mjolsness and DeCoste, 2001). As a consequence, scientific progress regarding data exploration tools in ecology and environmental sciences is partly on halt (Mahecha et al., 2007). For instance, an important open issue is how large-scale patterns of plant co-occurrences are related to environmental determinants (Bierman et al., 2010; Kühn et al., 2006; Tautenhahn et al., 2008).

In parallel to a growing availability of plant occurrence observations and ancillary data such as hyperspectral reflectances (Feilhauer and Schmidtlein, 2011), the science of machine learning and pattern recognition have progressed substantially. Today, empirical inference via machine learning provides fundamentally new perspectives. For instance, powerful techniques of non-linear dimensionality reduction such as Isometric Feature Mapping (Isomap; de Silva and Tenenbaum, 2003; Tenenbaum et al., 2000) and Locally Linear Embedding (LLE; Roweis and Saul, 2000) are of high relevance when being transferred to ecological applications (Mahecha and Schmidtlein, 2008; Mahecha et al., 2010). The rationale behind the use of non-linear dimensionality reduction techniques in the analysis of species data is that they may exploit non-linear, higher-order relations between the species that are present in the data. These ordination techniques construct a "species map"[1] in which species that frequently co-occur are depicted close

* Corresponding author.
E-mail address: lvdmaaten@gmail.com (L. van der Maaten).

[1] Throughout the paper, we refer to the ordination space as a "map" or "species map". Note that a species map in this sense is not a geographical map but a two-dimensional or multi-dimensional representation of species co-occurrences: co-occurring species are modeled close together in the species map.

1574-9541/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecoinf.2012.01.005

together; species that hardly co-occur are depicted far apart in the map.

Even advanced non-linear ordination techniques still have issues that are difficult to resolve. One of the common characteristics of both linear and non-linear dimensionality reduction techniques is that they treat the binary vegetation data as points in a high-dimensional metric space. As a result, these techniques assume that the similarities between species (i.e., between the points in this space) are metric (Legendre and Legendre, 1998). The assumption that similarities between species are metric is problematic, in particular, because species co-occurrences are likely to violate the *triangle inequality*. The triangle inequality states that if a point *A* and a point *B* are close together and points *B* and *C* are close together, then also points *A* and *C* have to be close. In general, the triangle inequality does not hold for species co-occurrences though: in vegetation data or floristic inventories, species *A* may co-occur with species *B* and species *B* may co-occur with species *C*, without species *A* ever co-occurring with species *C*. Because low-dimensional ordination plots of species (the "species maps") are designed to satisfy the triangle inequality, it is in most cases unlikely (or in the described case even impossible) that a given co-occurrence structure is accurately modeled in the species map.

In this study, we address the problem described earlier by analyzing vegetation survey data using an ordination technique that tries to faithfully model the non-metric nature of species co-occurrences. Instead of constructing a single two-dimensional or multi-dimensional species map like conventional ordination approaches, the ordination technique we investigate constructs multiple low-dimensional maps that visualize complementary similarity structure (Cook et al., 2007; van der Maaten and Hinton, 2012). Using multiple maps allows for the visualization of similarities that do not obey the triangle inequality, such as species co-occurrences. We apply the ordination technique – called multiple maps t-Distributed Stochastic Neighbor Embedding (or multiple maps t-SNE) – to construct visualizations of the inventory of the German flora (FLORKART). The resulting visualizations reveal the merits of our approach; in particular, (1) they are often better than most conventional methods at revealing the Bray–Curtis distances between the species and (2) they uncover the dependency of species co-occurrences on climate and soil biogeochemical conditions.

2. Materials and methods

2.1. The floristic data

The present study investigates the FLORKART database (see http://netphyd.floraweb.de), which is the outcome of a cumulative mapping project involving literature reviews and thousands of voluntary surveyors in several organizational subunits (Haeupler and Schönfelder, 1989). FLORKART contains vascular plant species counts for Germany, which have proved to be of significant value for biogeographical analyses (e.g., Bierman et al., 2010; Kühn et al., 2004, 2006; Schmidtlein, 2004; Tautenhahn et al., 2008). The floristic records include species data at all taxonomic levels, including subspecies and aggregates. Here we use a revised version of the database, in which records were considered at the species level only (Mahecha and Schmidtlein, 2008). The resulting database contains presence–absence data for 3917 vascular plant species recorded between 1950 and 2000, yet the majority of the data were collected between 1980 and 1990. For our purpose of investigating species co-occurrences, we excluded species that occur in a very small number of locations, as well as species that occur in almost all surveyed locations. Specifically, we only consider species that are found in at least 100 surveyed locations, and in at most 1000 surveyed locations. This leaves a total of $D = 850$ species to be considered in our experiments.

The collection of the FLORKART database covers Germany at a geographical grid resolution of $6' \times 10'$. For our purposes, cells with less than 50% of their area within Germany or fewer than 100 records were considered insufficiently mapped and therefore excluded. The remaining data contain presence–absence data for $N = 2762$ locations in Germany. Mathematically, the data thus consists of $N = 2762$ binary vectors in a $D = 850$-dimensional space. In total, the data contains a total of 1, 794, 758 species recording, so roughly 10% of the bits in the data is set to 1 (to indicate the presence of a specific species at a specific location). From the binary vectors, we compute a $D \times D$ species co-occurrence matrix, i.e. a matrix with elements $p_{j|i}$ that represent the probability that species j occurs at a specific location, given that species i occurs at that location. The co-occurrences are normalized in such a way that $\sum_j p_{j|i} = 1$. Mathematically, each entry $p_{j|i}$ is given by:

$$p_{j|i} = \frac{N_{ij}/N_i}{\sum_{k \neq i} N_{ik}/N_i}, \tag{1}$$

where N_{ij} represents the number of locations in which species i and j occur, where N_i represents the number of locations in which species i occurs, and where we define $p_{i|i} = 0$ (since we are only interested in co-occurrences of species).

To be able to interpret the visualizations constructed using various ordination techniques, we make use of meta-information on the spatial locations in which the species occur. In particular, we computed the mean (long-term) temperature and precipitation value for each species by averaging temperature and precipitation values over all grid cells in which that species occurs. The temperature and precipitation values are estimated from data at the Climate Research Unit of the University of East Anglia, UK (CRU; New et al., 2002), and were adjusted in order to remove temporal inconsistencies by the Potsdam Institute for Climate Impact Research (PIK; Österle et al., 2003). We also computed minimum and maximum temperatures for each species, but we found these to be strongly correlated with mean temperature, which is why we do not consider minimum and maximum temperatures here.

In addition, we extracted fractions of spatial coverages of the major substrate types from the digital version of the German geological map (BGR, 1993), and computed fractions of substrate types per species (again, by averaging over each grid cell in which a species occurs). Herein, we interpret these fractions of substrate types as proxies for the biogeochemical preconditions for the FLORKART species; we differentiate fractions of mud, raised-bogs, other bogs, sandy soils, loess based soils, lime deficient, lime-stone based, other-bedrock, and unclassified soils.

In order to coarsely characterize the environmental space of each FLORKART species, we work with mean values[2] of each environmental variable estimated across grid cells in which that species effectively occurs. Species with a universal occurrence are, therefore, associated with the mean of the environmental space, while a species that occurs only once is characterized by its specific values on the respective grid cell.

2.2. t-Distributed Stochastic Neighbor Embedding

t-SNE is a recently introduced non-linear ordination technique (van der Maaten and Hinton, 2008). Its input consists of a collection of N conditional probability distributions P_i with entries $p_{j|i}$ that are proportional to the probability that species j occurs on a location at which species i occurs. These probabilities are defined in such a way that $\sum_j p_{j|i} = 1$ and $p_{i|i} = 0$, and they can readily be computed from the raw species data.

[2] We note here that species distributions along environmental gradients are rarely truly Gaussian (Oksanen and Minchin, 2002). Yet, the mean is the most parsimonious measure of centers of distribution if an entire flora is taken into account.

The aim of t-SNE is to model each species i by a point $\mathbf{y}_i$ in a two-dimensional species map in such a way that the similarities $p_{j|i}$ are modeled as well as possible in the species map. The similarity between two points in the map is defined to be proportional to the density under a Student t-distribution with a single degree of freedom, i.e., to the density under a Cauchy distribution:

$$q_{j|i} = \frac{\left(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2\right)^{-1}}{\sum_{k \neq i} \left(1 + \|\mathbf{y}_i - \mathbf{y}_k\|^2\right)^{-1}} \text{ for } \forall i \forall j : i \neq j, \quad (2)$$

where we set $q_{i|i} = 0$. By using a heavy-tailed distribution[3] to measure similarities in the low-dimensional map, t-SNE allows points that are only slightly similar to be visualized much further apart in the map. This is beneficial, because it helps to resolve a problem that is the result of the exponential difference between high and low-dimensional space known as the *crowding problem*. For more details on how the use of a heavy-tailed function to measure similarities in the map addresses the crowding problem, we refer to van der Maaten and Hinton (2008).

In t-SNE, the error between the species co-occurrences $p_{j|i}$ and the similarities between species in the two-dimensional map $q_{j|i}$ is measured by means of a sum of the natural divergences between the conditional distributions P_i and Q_i, i.e., by measuring a sum of Kullback–Leibler divergences:

$$C(Y) = \sum_i KL(P_i || Q_i) = \sum_i \sum_{j \neq i} p_{j|i} log \frac{p_{j|i}}{q_{j|i}}. \quad (3)$$

The asymmetric nature of the Kullback–Leibler divergence leads the cost function to focus on appropriately modeling the large pairwise similarities $p_{j|i}$ between the input objects. In other words, similar input objects really need to be close together in the low-dimensional map in order to minimize the cost function $C(Y)$. In this respect, t-SNE is similar to many other non-linear ordination techniques that focus on preserving small pairwise distances (e.g., Belkin and Niyogi, 2002; Lawrence, 2011; Roweis and Saul, 2000; Sammon, 1969; Weinberger et al., 2007). As the cost function $C(Y)$ is non-convex, the minimization of $C(Y)$ is typically performed using a gradient descent method. Code for t-SNE is available from http://homepage.tudelft.nl/19j49/tsne.

2.3. Multiple maps t-SNE

The probabilistic nature of t-SNE allows for a natural extension to a multiple maps version (multiple maps t-SNE; van der Maaten and Hinton, 2012) that can successfully visualize non-metric species co-occurrences. Multiple maps t-SNE constructs a collection of M two-dimensional maps,[4] all of which contain N points (one for each of the N species). In each map with index m, a point with index i has an "importance weight" $\pi_i^{(m)}$ that measures the importance of point i in map m. Because of the probabilistic nature of multiple maps t-SNE, the importance weights are constrained in such a way that $\forall i \forall m : \pi_i^{(m)} \geq 0$ and $\forall i : \sum_m \pi_i^{(m)} = 1$. These constraints can be enforced by representing the importance weights $\pi_i^{(m)}$ in terms of unconstrained weights $w_i^{(m)}$ as follows:

$$\pi_i^{(m)} = \frac{e^{-w_i^{(m)}}}{\sum_{m'} e^{-w_i^{(m')}}}. \quad (4)$$

Multiple maps t-SNE redefines the conditional probabilities $q_{j|i}$ that represent the similarity between objects i and j in the visualization as the weighted sum of the pairwise similarities between the points corresponding to input objects i and j across all M maps:

$$q_{j|i} = \frac{\sum_m \pi_i^{(m)} \pi_j^{(m)} \left(1 + \|\mathbf{y}_i^{(m)} - \mathbf{y}_j^{(m)}\|^2\right)^{-1}}{\sum_{k \neq i} \sum_{m'} \pi_i^{(m')} \pi_k^{(m')} \left(1 + \|\mathbf{y}_i^{(m')} - \mathbf{y}_k^{(m')}\|^2\right)^{-1}} \text{ for } \forall i \forall j : i \neq j, \quad (5)$$

where $\mathbf{y}_i^{(m)}$ represents the low-dimensional model of object i in map m, and where, again, we set $q_{i|i} = 0$. The cost function that is minimized by multiple maps t-SNE is still given by Eq. (3), but now, it is optimized with respect to the locations of the points $\mathbf{y}_i^{(m)}$ in all species maps and with respect to the weights $w_i^{(m)}$.

Multiple maps t-SNE can successfully visualize non-metric similarities in a two-dimensional visualization. Recall our introductory example with the three species A, B, and C. If we construct a visualization with two maps, multiple maps t-SNE can give species A an importance weight of 1 in the first map, species B an importance weight of 1 in the second map, and species C an importance weight of $\frac{1}{2}$ in both maps, and it can give all three points nearby spatial locations in both maps (see Fig. 1 for an illustration). In this layout of the maps, the pairwise similarity between points A and C is roughly equal to $1 \times \frac{1}{2} = \frac{1}{2}$, and the pairwise similarity between points B and C is also roughly equal to $\frac{1}{2}$. However, the pairwise similarity between points A and B is 0, because the points A and B have no importance weight in each other's maps. This corresponds to what a human analyst would observe from the maps: points A and B are not close together with high weight in any of the maps, hence, they are not similar. In this way, multiple maps t-SNE can construct visualizations that do not satisfy the triangle inequality. This provides multiple maps t-SNE with an important advantage over conventional ordination approaches when it is used to visualize non-metric similarities such as the co-occurrences of species. As a result, different maps constructed by multiple maps t-SNE often model different "aspects" of the data (Cook et al., 2007). For instance, when multiple maps t-SNE is used to construct word maps based on word association data, different maps tend to represent words that correspond to different "topics" (van der Maaten and Hinton, 2012). Our hope is that when multiple

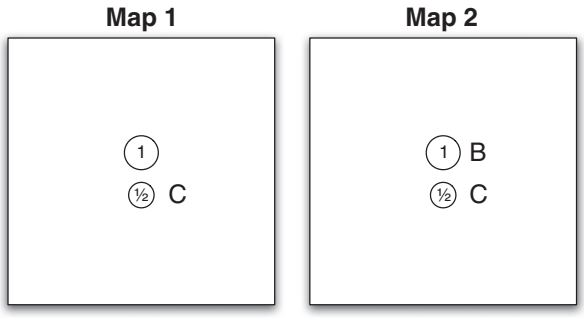

Fig. 1. Illustration of how multiple maps t-SNE can visualize species co-occurrences that do not obey the triangle inequality. Maps 1 and 2 are distinct ordination spaces, where the species A, B, and C are represented. The area of the circles represents the importance weights $\pi^{(m)}$ (in the range $0 \leq \pi^{(m)} \leq 1$) that the species have in each map. Image adopted from van der Maaten and Hinton (2012).

[3] In other studies, distributions with lighter tails (such as the Gaussian distribution) have been explored to model similarities in the map (Cook et al., 2007; Globerson et al., 2007; Hinton and Roweis, 2003). However, these were found to produce inferior results (van der Maaten and Hinton, 2008).

[4] Please note that this is not the same as constructing a $2 \times M$-dimensional representation, and then plotting these as M two-dimensional maps. In such an approach, two species need to be mapped close together in *all* maps in order to represent high species co-occurrence. In multiple maps t-SNE, two species need to be mapped close together (with high weight) in only one of the maps in order to represent high species co-occurrence.

maps t-SNE is used to model species co-occurrence data, it is also capable of separating out different "aspects" of the data.

In multiple maps t-SNE, the minimization of the cost function $C(Y)$ is performed using a gradient descent method. Code that implements multiple maps t-SNE is available from http://homepage.tudelft.nl/19j49/multiplemaps.

3. Results

This section first presents the results of visualization experiments on the FLORKART data set using conventional ordination approaches (in Section 3.1). Subsequently, it presents the results of our experiments with multiple maps t-SNE on the same data (in Section 3.2).

3.1. Visualization with standard ordination techniques

As a baseline result, we first present the results of experiments in which we constructed a single two-dimensional map of the FLORKART database using various conventional ordination approaches. In particular, we experimented with principal component analysis (PCA, Pearson, 1901), classical multi-dimensional scaling (CMDS; Torgerson, 1952), non-metric multi-dimensional scaling (NMDS; Kruskal and Wish, 1978), Isomap (Tenenbaum et al., 2000), and correspondence analysis (CA; Benzécri, 1973). PCA and correspondence analysis are performed on the raw species occurrence data, whereas classical multi-dimensional scaling, non-metric multi-dimensional scaling, and Isomap are performed on one minus the co-occurrence matrix (i.e., on the input distances $\delta_{ij} = 1 - p_{ij}$ with $\delta_{ii} = 0$). As described earlier, t-SNE is applied on the species co-occurrence matrix (i.e., on the co-occurrences p_{ij}).

In Table 1, we present the correlations between the Bray–Curtis dissimilarities (Bray and Curtis, 1957) of the species and the Euclidean distances in the maps constructed by the ordination approaches. The results presented in the table reveal that non-linear ordination techniques (e.g., non-metric multi-dimensional scaling, t-SNE) tend to outperform linear techniques (e.g., PCA, classical scaling, correspondence analysis). According to the results in the table, NMDS and t-SNE are the best-performing ordination techniques for this data set (but note that these performances may vary substantially with the data set under investigation, *cf.* results for tropical vegetation surveys in Mahecha et al., 2007).

In Fig. 2, we visualize the species maps constructed by classical scaling, Isomap (with $k = 4$), non-metric multi-dimensional scaling, and t-SNE. These maps can be interpreted in the logic of conventional ordination methods: each of the points in the low-dimensional map corresponds to a species, and the pairwise distance between two points should reflect the co-occurrence between the two corresponding species (a smaller pairwise distance in the map indicates a higher co-occurrence probability of the corresponding species). Low-dimensional species maps enable the analyst to draw conclusions on the underlying data structure (i.e., in the case of t-SNE, the analyst can draw conclusions on the probability of two species co-occurring). The locations of the points in the visualization appear to be related to certain environmental variables: coloring each species according to mean annual temperatures and mean precipitation sums reveals clear patterns in all ordination spaces, although non-metric multi-dimensional scaling and t-SNE seem to uncover the gradients more accurately (at least in two dimensions). More subtle properties of the environmental drivers are also better revealed by NMDS and t-SNE (see Figs. 3 and 4). Species color codings according to different environmental drivers (in particular, according to other soil biogeochemical preconditions) are presented in Appendix A.

Overall, the visualizations constructed by the single map NMDS and t-SNE reveal the major climatological and geological (and biogeochemical) pre-conditions, despite the fact that the differentiation according to the multiple-maps principle was not applied.

Table 1
Correlation between Bray–Curtis dissimilarities and Euclidean distances in species maps constructed by various ordination approaches (higher is better). For multiple maps t-SNE, the correlation between the Euclidean distances in map 1 and those in map 2 is −0.0539, suggesting that the maps reveal uncorrelated structure in the data.

Technique	Correlation
Principal component analysis	0.1059
Classical multidimensional scaling	0.4171
Non-metric multidimensional scaling	0.7238
Isomap	0.4877
Correspondence analysis	0.5763
t-SNE	0.7368
Multiple maps t-SNE, map 1	0.6793
Multiple maps t-SNE, map 2	0.4290

3.2. Visualization with multiple maps t-SNE

We now turn to experiments in which we used multiple maps t-SNE with $M = 2$ maps to visualize the species co-occurrence data. In order to objectively compare the performance of multiple maps t-SNE with that of the single-map ordination approaches, wepresent the correlation between the Bray-Curtis dissimilarities of the species and the Euclidean distances between the species (in each of the two maps) in Table 1. In the computation of these correlations, the species were weighted according to their importance weights $\pi_i^{(m)}$ in the maps: species with a higher importance weights thus have a larger influence on the observed correlation. We also computed the (weighted) correlation between the Euclidean distances in map1 and those in map 2: this correlation is −0.0539. Hence, the results indicate that the structure that is modeled in each of the two maps correlates with species similarity, but at the same time, that the structure in the two maps is completely unrelated (because there is no correlation between the maps). Indeed, this analysis shows that the two maps constructed by multiple maps t-SNE reveal different aspects of the data.

In Figs. 5, 6, and 7, we present the species maps constructed by multiple maps t-SNE). In these maps, the size of a point corresponds to the importance weight of that point in the map: a larger point corresponds to a higher importance weight in the map. Note that two species have high co-occurrence if there is at least one map in which both corresponding points have high weight *and* in which both corresponding points are close together. As in the plots constructed using traditional single-map ordination approaches, the points are colored according to various environmental and biogeochemical variables.[5] The multiple species maps can also be explored using an online, interactive visualization tool that is available on http://homepage.tudelft.nl/19j49/multiplemaps.

The visualizations constructed by multiple maps t-SNE reveal a variety of fine scale structures that are clearly related to environmental patterns. The visualizations on the lefthand side of Fig. 5 reveal again that mean annual temperature and precipitation sums are related to the floristic gradients. Moreover, the maps on the lefthand side of Figs. 6 and 7 reveal that biogeochemical preconditions have more local effects on the species associations in the species maps, they seem to be more clearly related to the uncovered species gradients compared to the single map t-SNE (see Appendix A). For instance, the antipodal gradients in more lime stone based species and those occurring at sandy soils suggest that the soil pH value (soil acidity) is a clear feature in FLORKART. Our analysis reveals a variety of such features. Some very specific conditions like high mean occurrences of species in moors (bogs) underscore the environmental relevance of the first level of the multiple maps.

[5] Note that the locations of the points in the maps in Figs. 5 to 7 are identical: the only thing that changes is the color code of the species representations.

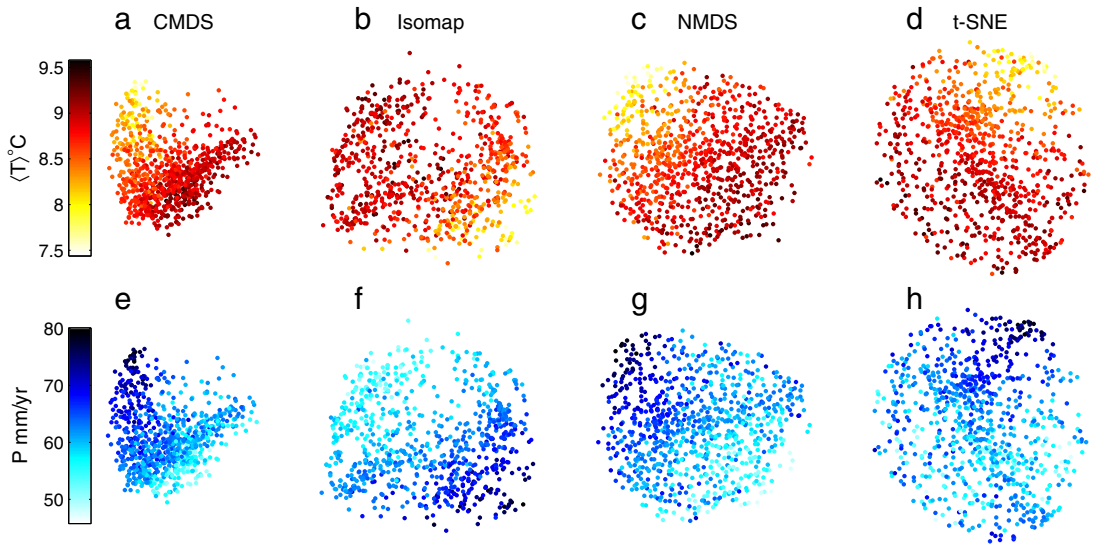

Fig. 2. Two-dimensional embeddings of the FLORKART species occurrence data constructed by CMDS (a, e), Isomap with $k=4$ (b, f), NMDS (c, g), and t-SNE (d, h). Here, each dot represents a species in the two-dimensional ordination space (note that CMDS and Isomap have per construction more dimensions, but only the first two are shown). In panels (a)–(d) each species in the ordination space is color coded by the mean annual temperatures ⟨T⟩°C averaged over the grid cells of occurrences of the respective species. Panels (e)–(h) show the same color coding according to the mean annual precipitation sums P mm/year. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

By contrast, the maps on the righthand side of the figures model structure that is fairly independent from environmental variables. Instead, they appear to model locally varying structures that capture more complex patterns (indeed, it models a different "aspect" of the data). In particular, we find that many of the frequently occurring species are grouped together in the center of the righthand side

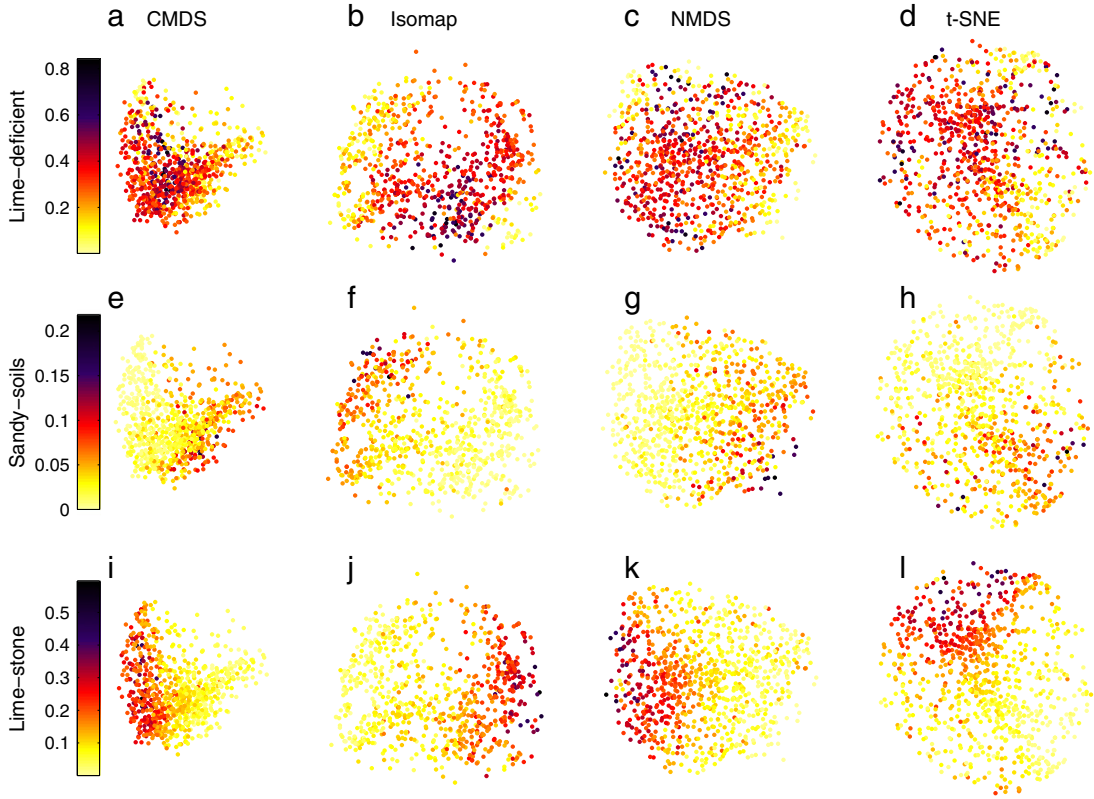

Fig. 3. Two-dimensional embeddings of the FLORKART species occurrence data constructed by CMDS (a, e, i), Isomap with $k=4$ (b, f, j), NMDS (c, g, k), and t-SNE (d, h, l). Here, each dot represents a species in the two-dimensional ordination space (note that CMDS and Isomap have per construction more dimensions, but only the first two are shown). The color coding corresponds to (a)–(d) the mean percentage of "lime deficient" underground in the grid cells where the species was found. The corresponding visualizations (e)–(h) show the mean percentages of "sandy soil" conditions across grid cells; (i)–(l) show the corresponding picture color coded according to the mean percentages of lime stones. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

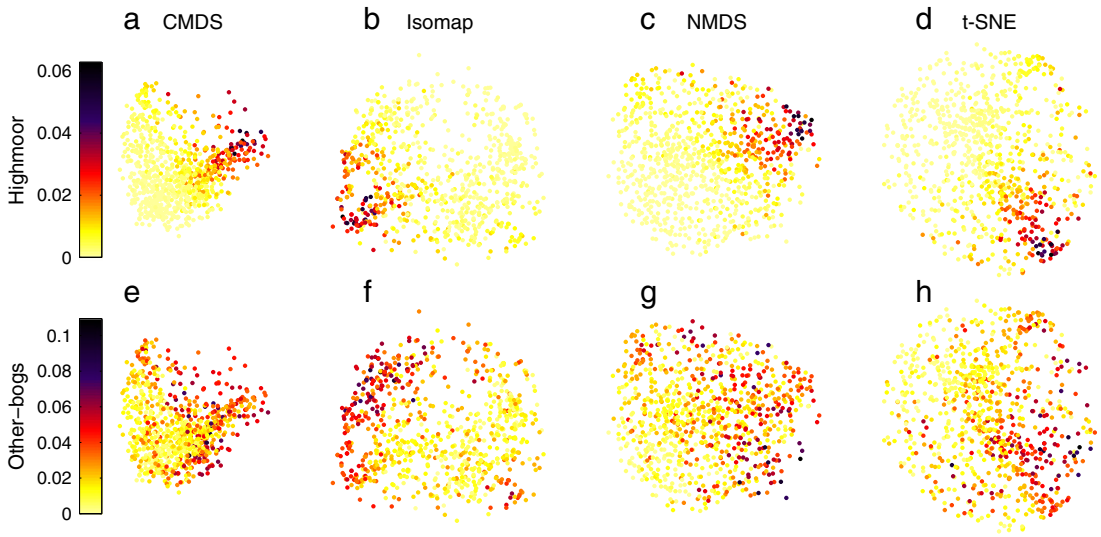

Fig. 4. Two-dimensional embeddings of the FLORKART species occurrence data constructed by CMDS (a, d), Isomap with $k=4$ (b, e) and t-SNE (c, f). Here, each dot represents a species in the two-dimensional ordination space (note that CMDS and Isomap have per construction more dimensions, but only the first two are shown). The color coding corresponds to (a)–(d) the mean percentage of "raised bogs" (here denoted as high-moore) in the grid cells where the species was found. The corresponding visualizations (e)–(h) show the mean percentages of "other bogs" across grid cells. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

map with low weights (i.e. they have a low importance in the right-hand side map). Fig. 8 illustrates that these species are placed here according to their abundance. Widely distributed species simply have a higher probability of co-occurrence than rare species. This rather trivial pattern in the data has only minor relevance in the righthand side maps constructed by multiple maps t-SNE, which allows the lefthand side map to reveal interesting structures in the underlying species co-occurrence data.

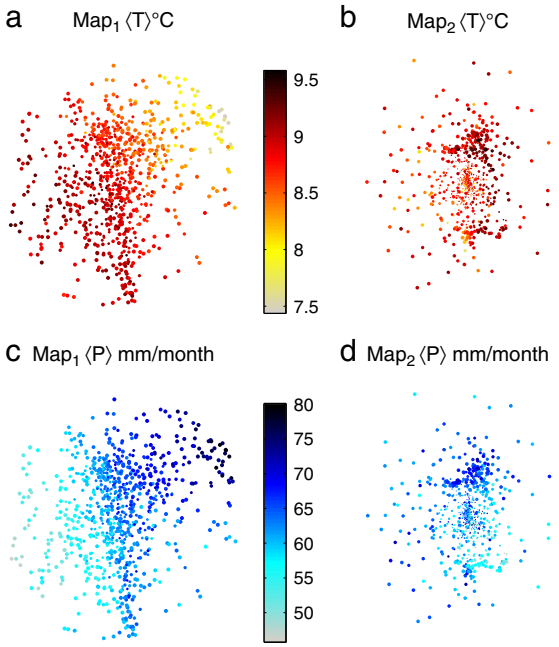

Fig. 5. Two-dimensional embeddings of the FLORKART species occurrence data constructed by multiple maps t-SNE with two maps. Here, each dot represents a species in the two-dimensional ordination space, and the color coding corresponds to, (a, b) the mean annual temperature of the grid cells in which the species was found, and (b, c) the mean of the cumulative annual precipitation sums over the grid cells where the species was found. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

In Fig. 9, we give an overview of the correlation between Euclidean distances in the ordination spaces and the differences in environmental variable. The figure reveals that multiple maps t-SNE is better than traditional ordination approaches at uncovering the major climate gradients. Some of the fine-scale soil conditions, however, seem to be better captured by conventional ordination methods, in particular, by correspondence analysis.

4. Discussion

The results of our experiments have shown that visualizations constructed by multiple maps t-SNE reveal a series of detailed relations of species co-occurrences with environmental determinants. The results suggest that a technique that tries to appropriately model non-metric relations between species also satisfies our ecological curiosity: multiple maps t-SNE reveals[6] how the species gradients within the ordination space are explicable by habitat availability; the latter being defined by specific environmental constraints. An open question is whether the relative positions of the species are in line with an unambiguous geobotanical interpretation.

4.1. Detailed analysis of the results

An in-depth examination of Fig. 10a–d reveals several biogeographical taxon groups or floristic elements (McLaughlin, 1994). For instance, the cutout in Fig. 10a shows plants from the oceanic northwest of Germany, with representatives of the "atlantic" floristic region of Europe (*sensu* (Roisin, 1969). Examples are the aquatic plants *Luronium natans* and *Isolepis fluitans*, or plants from adjacent moorland or heathland (for instance *Erica tetralix*, *Genista anglica*, *Narthecium ossifragum*, and *Myrica gale*). All these species are tied to more or less humid or wet conditions and often to acidic soils. On the opposite site of the ordination (cutout Fig. 10b) we find species of calcareous slopes in Southern Germany like *Sesleria alba*, *Buphthalmum salicifolium*, or *Carduus defloratus*. To the left of this group species of calcareous slopes with a distribution reaching further north and into continental regions prevail (*Anacamptis pyramidalis*, *Aceras*

[6] It should be noted here that species distributions do not solely depend on environmental variables. In particular, the current species distribution also depends on climate fluctuations in the past.

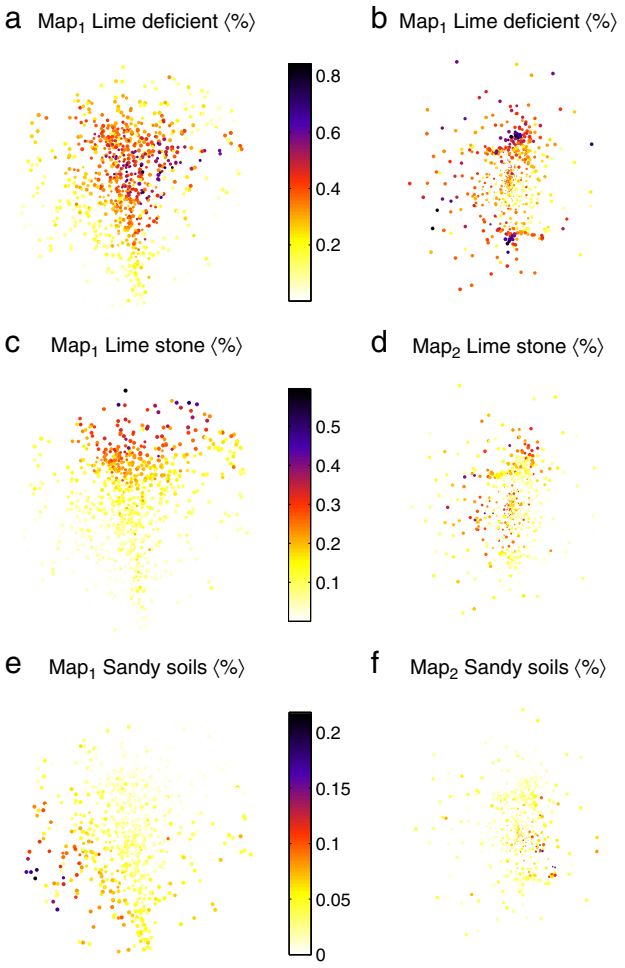

Fig. 6. Two-dimensional embeddings of the FLORKART species occurrence data constructed by multiple maps t-SNE with two maps. Here, each dot represents a species in the two 2D ordination spaces. The color coding corresponds to (a) and (b) the mean percentage of "lime deficient" soils in the grid cells where the species was found. The corresponding visualizations (c) and (d) show the mean percentages of "lime stone" conditions across grid cells; (e) and (f) show the corresponding picture color coded according to the mean percentages of sandy soils. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

anthropophorum, or *Pulsatilla vulgaris*). On the right side of the graph, taxa of nutrient-poor wetlands are aggregated. Habitats span from raised bogs (*Carex limosa*, *Scheuchzeria palustris*, or *Drosera longifolia*) to calcareous fens (*Blysmus compressus*, *Eleocharis quinqueflora*, or *Liparis loeselii*). Plants of raised bogs reaching to the northwest (like *Andromeda polifolia* or *Lycopodiella inundata*) link back to the "atlantic" domain at the bottom of the graph. This is only to mention a few out of many meaningful groupings that are often connected by taxa with mediating properties.

4.2. Comparing ordination approaches

A disadvantage of multiple maps ordination is that it adds an additional free parameter that needs to be set by the user, viz. the number of maps. An approach to automatically set the number of maps is proposed by (van der Maaten and Hinton, 2012): construct ordinations for increasing number of maps, and monitor the so-called *neighborhood preservation ratio*. The neighborhood preservation ratio measures what ratio of the most frequently co-occurring pairs of species is modeled as nearest neighboring points in the ordination space.

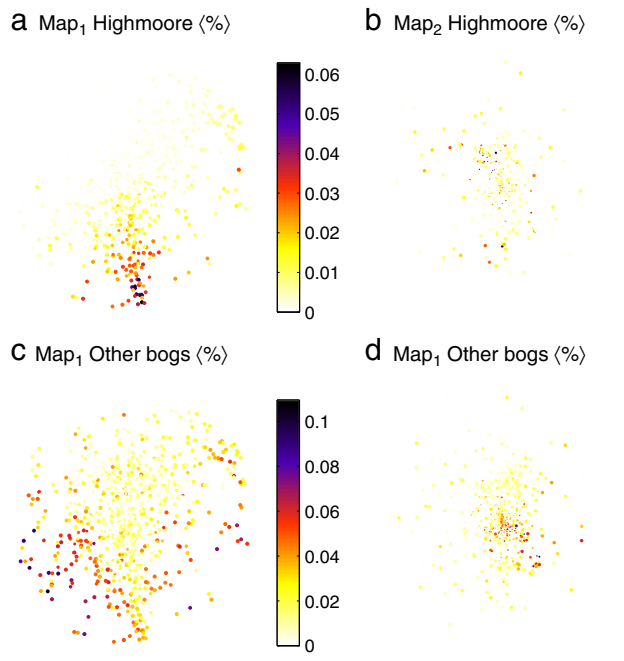

Fig. 7. Two-dimensional embeddings of the FLORKART species occurrence data constructed by multiple maps t-SNE with two maps. Here, each dot represents a species in the two 2D ordination spaces. The color coding corresponds to (a) and (b) the mean percentage of "high-moore" in the grid cells where the species was found. The corresponding visualizations (c) and (d) show the mean percentages of "other bogs". (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The neighborhood preservation ratio typically stops increasing after a certain number of maps are used in the multiple maps ordination. A sensible choice for the number of maps used in the final species ordination is the number of maps at which the increase of the neighborhood preservation ratio stops, i.e., the number of maps at which adding new maps does not significantly change the ordination anymore.

The results presented in Section 3 (in particular, those in Fig. 9) reveal that there is no single ordination approach that outperforms the other approaches in explaining all of the species correlations with environmental variables. As a result, there is not only a strong need for developing new ordination approaches that more appropriately model the data properties, but is probably at least as important that data analysts consider *an ensemble* of approaches. Indeed, our results suggest that each of the approaches in such an ensemble can reveal different, relevant information.

4.3. Observational limitations

Despite the convincing results presented earlier, we have to be aware of possible limitations inherited from data biases. While ideally, the binomial botanical nomenclature leads to an efficient, non-redundant encoding of floristic elements, substantial sampling biases are unavoidable in the construction of national scale faunistic and floristic databases (Petřík et al., 2010). Spatial data collections often suffer from slight taxonomic confusions but also from systematic sampling artifacts. The reason is that such observations depend on expert knowledge. This expertise is, however, itself organized spatially and inconsistent depending on the expert's calibration training. Moreover, the analyzed FLORKART data are influenced by sampling artifacts, mainly caused by the federal organization of the floristic survey (Mahecha and Schmidtlein, 2008). Problems of this kind are well known, but even highly standardized post-processing schemes are not fully capable of removing errors of this kind (Bierman et al.,

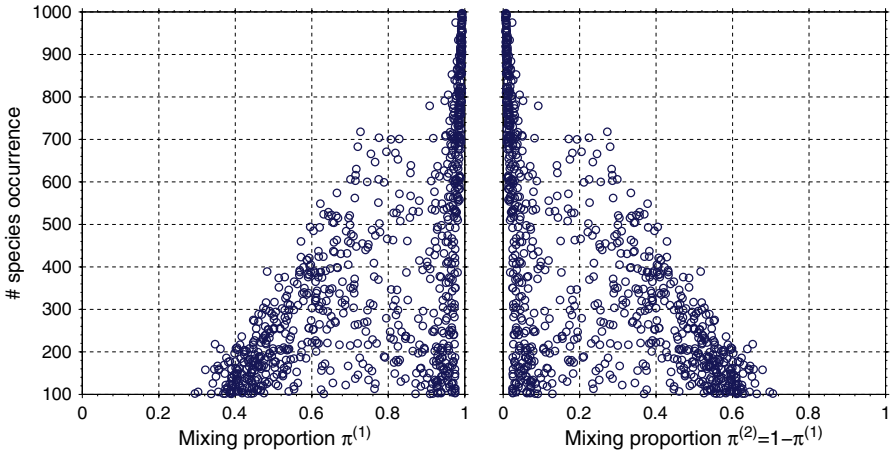

Fig. 8. The incidence (number of occurrences) of each species as a function of its importance weight (mixing proportion π) in both maps. Often occurring species tend to be represented in map 1, whereas rare species have higher weights in the second map.

2010). An analysis of the environmentally not interpretable map 2 could mean that sampling bias related patterns in the floristic co-occurrences might be captured there, but we cannot provide any empirical evidence for that.

4.4. Future implications

There are several reasons letting us expect a revitalization of the field of botany and vegetation sciences where techniques like multiple maps t-SNE will be valuable tools. First, currently several synthesis efforts seek to synthesize data of regional projects into global data collections (Kueffer et al., 2011), which implies massive amounts of data awaiting further exploration in the near future. Second, alternative species sampling methods are currently being explored and applied. For instance, rapidly evolving DNA-barcoding techniques are capable to provide large-scale inventories with unprecedented levels of accuracy as shown by Lahaye et al. (2008). This study recently reported a first extensive species inventory sampling of entire "hotspots of biodiversity" (geographical areas with extremely high α-diversity). These tools are also not fully error-free, but Lahaye et al. (2008) reported a classification error of less than 10%. A different perspective is provided by "functional" monitoring data. For instance, the recently presented global TRY database of "plant traits" (Kattge et al., 2011) provides an unprecedented collection of plant properties which are expected to help understanding ecosystem functioning in relation to environmental factors. This project is intended to further extend our knowledge of patterns of plant co-occurrences and interactions, where methods as presented in our study may become crucial tools (Kattge et al., 2011). The present conceptual paper shows that multiple maps t-SNE is a promising tool for performing such explorations; it may reveal ecological interactions among species from the non-metric relations between those species.

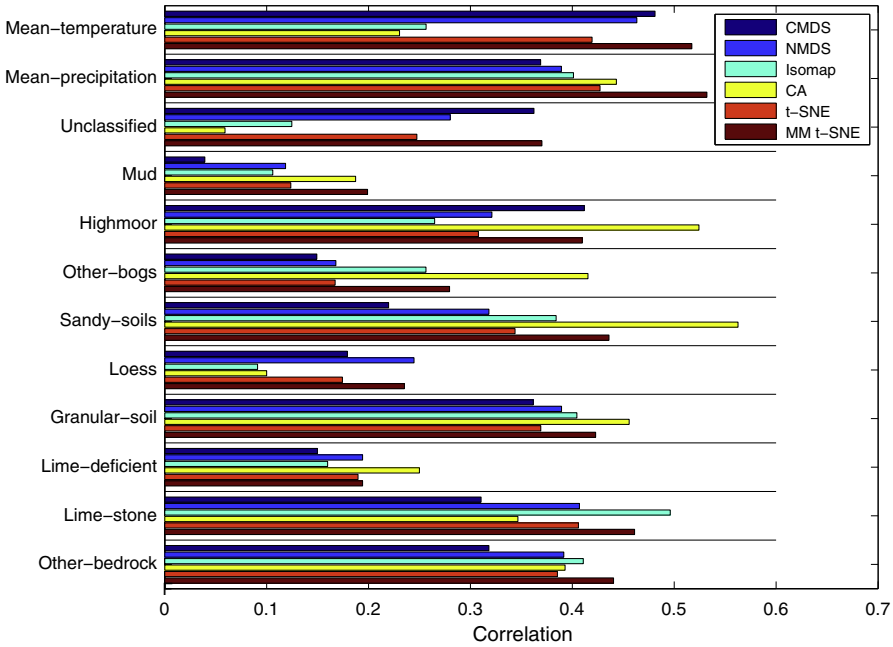

Fig. 9. Correlation between distances in the species maps and variations in environmental variables (for various ordination techniques). The maps constructed by multiple maps t-SNE correlate better with major climatological variables than traditional single-map ordination techniques. Correspondence analysis is better at capturing the relations of species distribution with some soil types (in particular, with high moor and sandy soils). Multiple maps t-SNE consistently provides better models than single-map t-SNE.

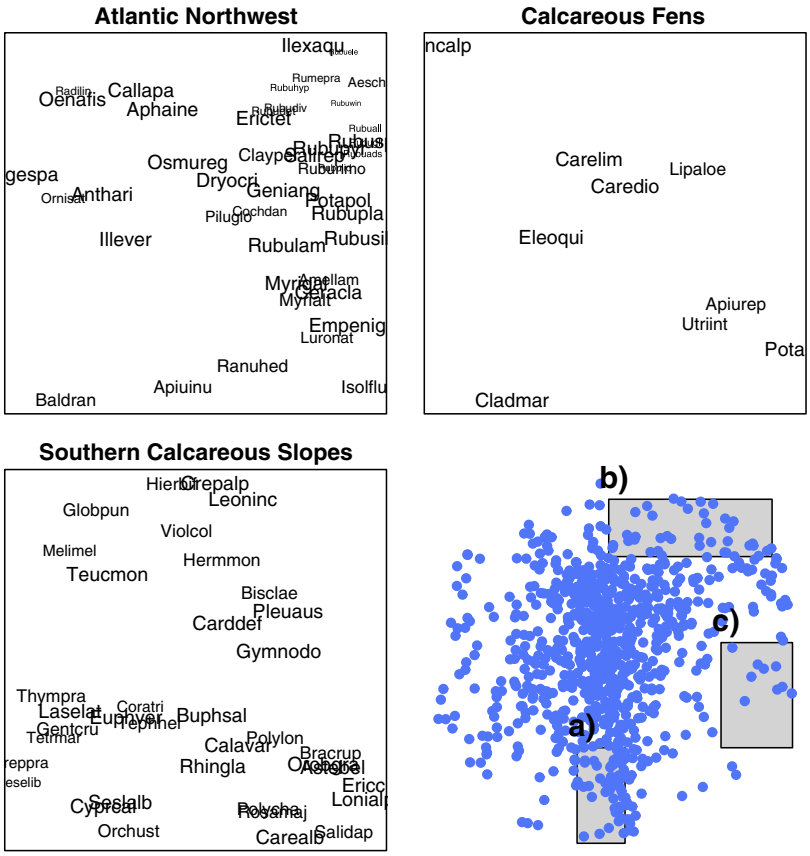

Fig. 10. Scatter plots of the importance weight of each of the species in both maps, as a function of the incidence of the species.

5. Conclusion

The multiple maps approach is very well suited for analyzing binary species co-occurrence data, such as floristic inventories or vegetation monitoring data, because it omits the use of distance measures that pretend metric species-to-species relationships. Working with multiple maps may improve our capability to separate effects that operate on different scales, or that are due to different causes. A disadvantage of working with multiple maps may be that it is not always clear how many maps should be used.

Our experiments revealed that multiple maps t-SNE is one of the best techniques for revealing species co-occurrences and relating large-scale environmental drivers. Therefore, the method may substantially contribute to biogeographical data mining. It certainly is a serious candidate for being included in ensemble approaches of macroecology, where multiple methods reveal different aspects of a data set.

Acknowledgments

This study emerged during a data mining workshop funded by the Max Planck Society (for details, see Reichstein et al., 2009). Laurens van der Maaten acknowledges support by the Netherlands Organization for Scientific Research (NWO; grant no. 680.50.0908), and by the EU-FP7 NoE on Social Signal Processing (SSPNet).

The authors are grateful to the thousands of volunteers mapping the flora of Germany and to the German Federal Agency for Nature Conservation (Bundesamt für Naturschutz) for providing the FLORKART data, in particular Rudolf May. The authors thank Geoffrey Hinton for helpful discussions.

Appendix A. Supplementary data

Supplementary data to this article can be found online at doi:10.1016/j.ecoinf.2012.01.005.

References

Bekker, R.M., van der Maarel, E., Bruelheide, H., Woods, K., 2007. Long-term datasets: from descriptive to predictive data using ecoinformatics. Journal of Vegetation Science 18, 458–462.
Belkin, M., Niyogi, P., 2002. Laplacian Eigenmaps and spectral techniques for embedding and clustering. Advances in Neural Information Processing Systems, vol. 14, pp. 585–591.
Benzécri, J., 1973. L'analyse des données, 2 vol. Dunod, Paris.
BGR, 1993. Geologische Karte der Bundesrepublik Deutschland 1:1000000. Karte mit Erläuterungen, Textlegende und Leitprofilen. Tech. rep. Bundesanstalt für Geowissenschaften und Rohstoffe, Hannover.
Bierman, S., Butler, A., Marion, G., Kühn, I., 2010. Bayesian image restoration models for combining expert knowledge on recording activity with species distribution data. Ecography 33, 451–460.
Bray, J., Curtis, J., 1957. An ordination of upland forest communities of southern Wisconsin. Ecological Monographs 27, 325–349.
Chytry, M., Rafajová, M., 2003. Czech national phytosociological database: basic statistics of the available vegetation-plot data. Preslia 75, 1–15.
Cook, J., Sutskever, I., Mnih, A., Hinton, G., 2007. Visualizing similarity data with a mixture of maps. JMLR Workshop and Conference Proceedings, vol. 2, pp. 67–74.
de Silva, V., Tenenbaum, J., 2003. Global versus local methods in nonlinear dimensionality reduction. Advances in Neural Information Processing Systems, vol. 15. The MIT Press, Cambridge, MA, pp. 721–728.
Feilhauer, H., Schmidtlein, S., 2011. On variable relations between vegetation patterns and canopy reflectance. Ecological Informatics 6, 83–92.
Gauch, H., 1982. Multivariate Analysis in Community Ecology. Cambridge University Press, Cambridge, UK.
Globerson, A., Chechik, G., Pereira, F., Tishby, N., 2007. Euclidean embedding of co-occurrence data. Journal of Machine Learning Research 8, 2265–2295 (Oct).
Haeupler, H., Schönfelder, P., 1989. Atlas der Farn-und Blütenpflanzen der Bundesrepublik Deutschland. Ulmer, Stuttgart, Germany.

Hinton, G., Roweis, S., 2003. Stochastic neighbor embedding. Advances in Neural Information Processing Systems, vol. 15, pp. 833–840.
Jalas, J., Suominen, J., Lampinen, R., 1972–1999. Atlas Florae Europaeae. Akateeminen Kirjakauppa, Helsinki, Finland.
Kattge, J., Díaz, S., Lavorel, S., Prentice, I.C., Leadley, P., Bönisch, G., Garnier, E., Westoby, M., Reich, P.B., Wright, I.J., Cornelissen, J.H.C., Violle, C., Harrison, S.P., van Bodegom, P.M., Reichstein, M., Enquist, B.J., Soudzilovskaia, N.A., Ackerly, D.D., Anand, M., Atkin, O., Bahn, M., Baker, T.R., Baldocchi, D., Bekker, R., Blanco, C.C., Blonder, B., Bond, W.J., Bradstock, R., Bunker, D.E., Casanoves, F., Cavender-Bares, J., Chambers, J.O., Chapin, F.S., Chave, J., Coomes, D., Cornwell, W.K., Craine, J.M., Dobrin, B.H., Duarte, L., Durka, W., Elser, J., Esser, G., Estiarte, M., Fagan, W.F., Fang, J., Fernández-Méndez, F., Fidelis, A., Finegan, B., Flores, O., Ford, H., Frank, D., Freschet, G.T., Fyllas, N.M., Gallagher, R., Green, W., Gutierrez, A.G., Hickler, T., Higgins, S., Hodgson, J.G., Jalili, A., Jansen, S., Kerkhoff, A.J., Kirkup, D., Kitajima, K., Kleyer, M., Klotz, S., Knops, J.M.H., Kramer, K., Kühn, I., Kurokawa, H., Laughlin, D., Lee, T.D., Leishman, M., Lens, F., Lenz, T., Lewis, S.L., Lloyd, J., Llusià, J., Louault, F., Ma, S., Mahecha, M.D., Manning, P., Massad, T., Medlyn, B., Messier, J., Moles, A.T., Müller, S., Nadrowski, K., Naeem, S., Niinemets, Ü., Nöllert, S., Nüske, A., Ogaya, R., Oleksyn, J., Onipchenko, V.G., Onoda, Y., Ordoñez, J., Overbeck, G., Ozinga, W.A., Patiño, S., Paula, S., Pausas, J.G., Peñuelas, J., Phillips, O.L., Pillar, V., Poorter, H., Poorter, L., Poschlod, P., Prinzig, A., Proulx, R., Rammig, A., Reinsch, S., Reu, B., Sack, L., Salgado-Negret, B., Sardans, J., Shiodera, S., Shipley, B., Siefert, A., Sosinski, E., Soussana, J.-F., Swaine, E., Swenson, N., Thompson, K., Thornton, P., Waldram, M., van Weiher, E., White, M., White, S., Wright, J., Yguel, B., Zaehle, S., Zanne, A., Wirth, C., 2011. TRY—a global database of plant traits. Global Change Biology 17, 2905–2935.
Kruskal, J., Wish, M., 1978. Multidimensional scaling. Sage University Paper Series on Quantitative Application in the Social Sciences, vol. 07–011. Sage Publications.
Kueffer, C., Niinemets, Ü., Drenovsky, R.E., Kattge, J., Milberg, P., Poorter, H., Reich, P.B., Werner, C., Westoby, M., Wright, I.J., 2011. Fame, glory and neglect in meta-analyses. Trends in Ecology & Evolution 26, 493–494.
Kühn, I., Durka, W., Klotz, S., 2004. The flora of German cities is naturally species rich. Evolutionary Ecology Research 6, 749–764.
Kühn, I., Bierman, S., Durka, W., Klotz, S., 2006. Relating geographical variation in pollination types to environmental and spatial factors using novel statistical methods. New Phytologist 172, 127–139.
Lahaye, R., van der Bank, M., Bogarin, D., Warner, J., Pupulin, F., Gigot, G., Maurin, O., Duthoit, S., Barraclough, T.G., Savolainen, V., 2008. DNA barcoding the floras of biodiversity hotspots. Proceedings of the National Academy of Sciences 105, 2923–2928.
Lawrence, N., 2011. Spectral dimensionality reduction via maximum entropy. JMLR Workshop and Conference Proceedings, vol. 15, pp. 51–59.
Legendre, L., Legendre, P., 1998. Numerical Ecology. Elsevier, New York, USA.
Mahecha, M., Schmidtlein, S., 2008. Revealing biogeographical patterns by nonlinear ordinations and derived anisotropic spatial filters. Global Ecology and Biogeography 17, 284–296.
Mahecha, M., Marténez, A., Lischeida, G., Beck, E., 2007. Nonlinear dimensionality reduction: alternative ordination approaches for extracting and visualizing biodiversity patterns in tropical montane forest vegetation data. Ecological Informatics 2, 138–149.
Mahecha, M., Fürst, L., Gobron, N., Lange, H., 2010. Identifying multiple spatiotemporal patterns: a refined view on terrestrial photosynthetic activity. Pattern Recognition Letters 31, 2309–2317 (to appear).
McLaughlin, S., 1994. Floristic plant geography: the classification of floristic areas and floristic elements. Progress in Physical Geography 18, 185–208.
Mjolsness, E., DeCoste, D., 2001. Machine learning for science: state of the art and future prospects. Science 293, 2051–2055.
Mucina, L., Rodwell, J., Schaminée, J., Dierschke, H., 1993. European vegetation survey: current state of some national programmes. Journal of Vegetation Science 4, 429–438.
Myklestad, A., Birks, H.J.B., 1993. A numerical analysis of the distribution patterns of salix l species in Europe. Journal of Biogeography 20, 1–32.
New, M., Lister, D., Hulme, M., Makin, I., 2002. A high resolution data set of surface climate over global land areas. Climate Research 21, 1–25.
Oksanen, J., Minchin, P., 2002. Continuum theory revisited: what shape are species responses along ecological gradients? Ecological Modelling 157, 119–129.
Österle, H., Gerstengarbe, F.W., Werner, P.C., 2003. Homogenisierung und Aktualisierung des Klimadatensatzes der Climate Research Unit of East Anglia, Norwich (in German). Terra Nostra 6, 326–329.
Pearson, K., 1901. On lines and planes of closest fit to systems of points in space. Philiosophical Magazine 2, 559–572.
Petřík a, P., Pergla, J., Wild, J., 2010. Recording effort biases the species richness cited in plant distribution atlases. Perspectives in Plant Ecology, Evolution and Systematics 12, 57–65.
Reichstein, M., Mahecha, M.D., Carvalhais, N., 2009. Novel data mining strategies for exploring biogeochemical cycles and biosphere–atmosphere interactions, workshop report. iLEAPS Newsletter 8, 40–41.
Roisin, P., 1969. Le domaine phytogéographique atlantique d'Europe. Editions J. Duculot, Gembloux.
Roweis, S., Saul, L., 2000. Nonlinear dimensionality reduction by locally linear embedding. Science 290 (5500), 2323–2326.
Sammon, J., 1969. A nonlinear mapping for data structure analysis. IEEE Transactions on Computers 18 (5), 401–409.
Schmidtlein, S., 2004. Coarse-scale substrate mapping using plant functional response types. Erdkunde 58, 137–155.
Scholes, R.J., Mace, G.M., Turner, W., Geller, G.N., Jürgens, N., Larigauderie, A., Muchoney, D., Walther, B.A., Mooney, H.A., 2008. Towards a global biodiversity observing system. Science 321, 1044–1045.
Tautenhahn, S., Heilmeier, H., Götzberger, L., Klotz, S., Wirth, C., Kühn, I., 2008. On the biogeography of seed mass in Germany—distribution patterns and environmental correlate. Ecography 31, 457–468.
Tenenbaum, J., de Silva, V., Langford, J., 2000. A global geometric framework for nonlinear dimensionality reduction. Science 290 (5500), 2319–2323.
ter Braak, C., 1995. Ordination. In: Jongman, R., ter Braak, C., van Tongeren, O. (Eds.), Data Analysis in Community and Landscape Ecology. Cambridge University Press, Cambridge, UK, pp. 91–169.
Torgerson, W., 1952. Multidimensional scaling I: theory and method. Psychometrika 17, 401–419.
van der Maaten, L., Hinton, G., 2008. Visualizing data using t-SNE. Journal of Machine Learning Research 9, 2431–2456 (Nov).
van der Maaten, L., Hinton, G., 2012. Visualizing non-metric similarities in multiple maps. Machine Learning (to appear). .
Weinberger, K., Sha, F., Zhu, Q., Saul, L., 2007. Graph Laplacian regularization for large-scale semidefinite programming. Advances in Neural Information Processing Systems, vol. 19, pp. 1489–1496.