ECOGRAPHY

Research

Deriving site-specific species pools from large databases

Helge Bruelheide, Borja Jiménez-Alfaro, Ute Jandt and Francesco Maria Sabatini

H. Bruelheide (https://orcid.org/0000-0003-3135-0356) ✉ (helge.bruelheide@botanik.uni-halle.de), B. Jiménez-Alfaro (https://orcid.org/0000-0001-6601-9597), U. Jandt and F. M. Sabatini (https://orcid.org/0000-0002-7202-7697), Martin Luther Univ. Halle-Wittenberg, Inst. of Biology/Geobotany and Botanical Garden, Halle, Germany, and German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Leipzig, Germany. BJ-A also at: Research Unit of Biodiversity (CSIC/UO/PA), Univ. of Oviedo, Mieres, Spain.

Defining the species pool of a community is crucial for many types of ecological analyses, providing a foundation to metacommunity, null modelling or dark diversity frameworks. It is a challenge to derive the species pool empirically from large and heterogeneous databases. Here, we propose a method to define a site-specific species pool (SSSP), i.e. the probabilistic set of species that may co-occur with the species of a target community. Using large databases with geo-referenced records that comprise full plant community surveys, our approach characterizes each site by its own species pool without requiring a pre-defined habitat classification. We calculate the probabilities of each species in the database to occur in the target community using Beals' index of sociological favourability, and then build sample-based rarefaction curves from neighbouring records with similar species composition to estimate the asymptotic species pool size. A corresponding number of species is then selected among the species having the highest occurrence probability, thus defining both size and composition of the species pool. We tested the robustness of this approach by comparing SSSPs obtained with different spatial extents and dissimilarity thresholds, fitting different models to the rarefaction curves, and comparing the results obtained when using Beals co-occurrence probabilities or presence/absence data. As an example application, we calculated the SSSPs for all calcareous grassland records in the German Vegetation Reference Database, and show how our method could be used to 1) produce grain-dependent estimations of species richness across plots, 2) derive scalable maps of species richness and 3) define the full list of species composing the SSSP of each target site. By deriving the species pool exclusively from community characteristics, the SSSP framework presented here provides a robust approach to bridge biodiversity estimations across spatial scales.

Keywords: rarefaction curves, spatial scaling, species pool, species richness, vegetation databases

Introduction

The concept of a species pool, i.e. the set of all species available to colonize a focal site (Eriksson 1993, Srivastava 1999), plays a key role in community ecology theory (Cornell and Harrison 2014, Jiménez-Alfaro et al. 2018), restoration applications (Ladouceur et al. 2018), and the integration of large- and small-scale processes

underlying the biodiversity patterns observed at the community scale (Ricklefs 1987, Cornell and Harrison 2014). The so-called 'species pool hypothesis' (Taylor et al. 1990) proposes that the number of species present in any given location depends both on the commonness and geological age of the particular habitat type, thus jointly accounting for phylogenetic and biogeographical processes (Zobel 2016). The species pool is also a key ingredient of metacommunity theory, as it is the sum of all the communities that are linked through dispersal (Leibold et al. 2004). Furthermore, an explicit definition of the species pool is the key prerequisite of null model analyses, for instance, when testing the role of species interactions or environmental filters on the taxonomic and functional diversity and composition, or phylogenetic structure of a local community (Carstensen et al. 2013, Sabatini et al. 2018). Finally, the species pool concept is tightly linked to the concept of 'dark diversity' (Pärtel et al. 2011). By focusing on the species of the pool that do not occur in a target community, dark diversity has been used as a measure of community completeness in basic (Pärtel et al. 2013) and applied (Suding 2011) ecology.

When defined as the sum of all species having a high probability to occur in a specific habitat (Eriksson 1993, Pärtel et al. 1996, Zobel 1997, Lewis et al. 2016), the species pool explicitly accounts for environmental and biotic filtering (Graves and Gotelli 1983). Even if different definitions exist (Cornell and Harrison 2014), the habitat-specific definition of species pools has two advantages. First, it rules out species whose fundamental niches do not overlap with the abiotic site conditions of a location, e.g. excluding mire species from the species pool of a dry grassland habitat within the same region. Second, it ensures that species that pass the abiotic filter are also able to tolerate the intensity of competition or disturbance at a particular site (Lewis et al. 2016). To be useful for practical purposes, however, this habitat-specific formulation needs an operational definition, but no consensus has yet arisen (Cornell and Harrison 2014, Karger et al. 2016). Possible approaches to define habitat-specific species pools consider fixed geographic units such as ecoregions or landscapes or species distribution ranges (Gotelli and Graves 1996). However, it is difficult to define a geographical area from which a species can potentially disperse into a community because the range sizes of species in the species pool (i.e. the so-called dispersion fields) differ by orders of magnitude across regions (Graves and Rahbek 2005). Similarly, considering species' distribution ranges alone does not consider species' habitat-specificity, and accurate information on species' fidelity and ecological requirements are in most case unavailable (Lessard et al. 2012).

Alternatively, the species pool can be constructed by considering all species that occur in a set of sampled communities surrounding a focal site (Lewis et al. 2016). By only considering communities which are sufficiently similar to the focal site, this approach offers an opportunity to account for habitat-specificity (Cornell and Harrison 2014). This is conceptually equivalent to calculating rarefaction curves (Gotelli and Colwell 2001) or nonparametric richness estimators (Chao 1984), but extends the focus from the estimation of pool size alone (= number of species in the species pool), to include also its composition (= identity of the species in the pool). However, the selection of a set of communities sampled around a focal site does sum up the grain sizes of plots, but it does not control for the total area covered (i.e. the extent effectively sampled). This makes the species pool scale-dependent in two different aspects, i.e. grain and extent, ranging from local pools (generally defined at the landscape scale) to regional species pools calculated at larger distances (Jiménez-Alfaro et al. 2018).

Integrating the information derived from the composition of the sites surrounding the target location and the information on species co-occurrence patterns is a promising approach (Ewald 2002, Lewis et al. 2016). Embracing a probabilistic rather than deterministic view based on co-occurrence likelihoods, in particular, could resolve two practical problems of working with biodiversity data. First, it avoids the necessity to a-priori identify sites into discrete habitats, thus allowing to calculate the species pool for communities in ecotones. Second, it is convenient when dealing with heterogeneous data sets, which are often spatially unbalanced and collected with different protocols, as it allows to explicitly account for scale-dependency of biodiversity (Chase et al. 2018). Yet, even if defining site-specific species pools using species co-occurrence information has a long tradition (Beals 1984, Ewald 2002, Münzbergová and Herben 2004, Breitschwerdt et al. 2015, Lewis et al. 2016), it requires large-scale collections of co-occurrence data, which have only recently become widely available (Chytrý et al. 2016, Bruelheide et al. 2019). Simulation tests, for instance, showed that the use of Beal's index is only reliable when the number of sampling units exceeds 40 and the number of species exceeds ten (De Caceres and Legendre 2008).

Here, we propose a new approach to define the site-specific species pool (SSSP) of a community at a given geographic location, which combines the estimation of asymptotic curves based on neighbouring records with the probabilities of recorded species to occur in the target community. This approach 1) takes advantage of the existence of large databases with geo-referenced samples and 2) allows characterizing both the size and composition of the species pool of a target site, without requiring a pre-defined habitat classification. To illustrate the proposed method, we used a national vegetation database and defined the species pools of calcareous grassland communities in Germany. We tested the robustness of the approach by comparing the impact of different parameters when selecting neighbouring records, including different spatial extents, dissimilarity thresholds and asymptotic richness estimators. We also tested for the effect of heterogeneity of plot sizes and the use of pure presence/absence information of species instead of Beals occurrence probabilities. Finally, we demonstrate how this approach could be used to explore biogeographical patterns, and produce

scalable maps of species richness at different spatial grains, as well as standardized maps showing differences in the species pool composition across the study area.

Material and methods

General approach

The conceptual workflow is outlined in Fig. 1. Briefly, we first calculated the occurrence probabilities for all species in every plot using Beals' index of sociological favourability (Beals 1984). For each target plot, we then selected neighbouring plots with similar species occurrence probabilities, and estimated the expected size S of the species pool based on richness estimators or from the asymptote of sample-based rarefaction curves. Finally, we obtained the species pool composition by selecting the S species having the highest occurrence probability.

We first describe the details of this general procedure using the German Vegetation Reference Database (GVRD) as an example. We then explore the sensitivity of the algorithm to different options, such as varying the extent of the sample surrounding the target plot, dissimilarity thresholds and richness estimators. As a benchmark, we compare our results with those obtained when replacing occurrence probabilities with observed species occurrences or frequencies for calculating similarity to the target plot and for assessing the composition of the species pool. In addition, we test to which degree the results depend on different plot sizes or on using a particular preselected subset of potentially similar plots compared to the whole database. As a possible application, we finally illustrate how our method could be based to explore patterns in species pool composition, and create scalable maps of species pool richness. As an example, we use a vegetation dataset; however, the method can be equally applied to any type of co-occurrence data.

Example data set

As a basis for calculating SSSPs we used the German Vegetation Reference Database (GVRD Jandt and Bruelheide 2012), ID EU-DE-14 in the Global Index of Vegetation-Plot Databases, <www.givd.info>). After removing bryophyte records, these data include 2603 plant taxa sampled in 170 039 vegetation plots (representative areas of homogeneous plant communities in a given site) across different habitats, ranging from forests to grasslands, from fens to salt marshes. As a target habitat for estimating the size and composition of SSSPs, we extracted all records of dry calcareous grasslands (described as *Festuco-Brometea* in the German vegetation literature), which were previously used in a continental survey (Willner et al. 2019). We only considered those plots with known plot size and location (n = 6319). Plot sizes varied between 0.1 and 5000 m² (median 10 m², mean 21.6 m², Supplementary material Appendix 1 Fig. A1). After taxonomic harmonization, total number of species in the data set was 1067. Species number per plot varied between 1 and 103 species (median 33).

Species co-occurrence probabilities

Occurrence probabilities for all species in the target habitat were calculated using Beals' index of sociological favourability (Beals 1984) according to formula (1).

$$x_{pi} = \frac{1}{N_p} \sum_{j}^{N_p} \frac{M_{ij}}{M_j} \tag{1}$$

The probability x_{pi} for species i to occur in a vegetation plot p is calculated as the sum of joint occurrences M_{ij} across all N_p species j contained in the plot, divided by the number of plots M_j in which the species j is present. With formula (1) species i itself is included in the calculation, which differs from the usage by Münzbergová and Herben (2004) and Lewis et al. (2016), where species i is excluded. We derived the M_{ij} values from the co-occurrence matrix **M**, based on all plot records included in GVRD, thus also including co-occurrence probabilities of species that were not part of the calcareous grassland sample set. Making use of the full vegetation database allows to include species from other habitats in the pool, when they occur sufficiently frequently with species of the target habitat. From the GVRD records, we removed all species that occurred only once (i.e. singletons), except if they occurred in the calcareous grassland subset (thus allowing singletons to affect community similarity when drawing samples around a target plot, see below). In total, the **M** matrix contained the co-occurrence information of 2432 × 2432 species. As a result of formula (1), we obtained the probability x_{pi} for each species to occur in each one of the 6319 calcareous grassland plots.
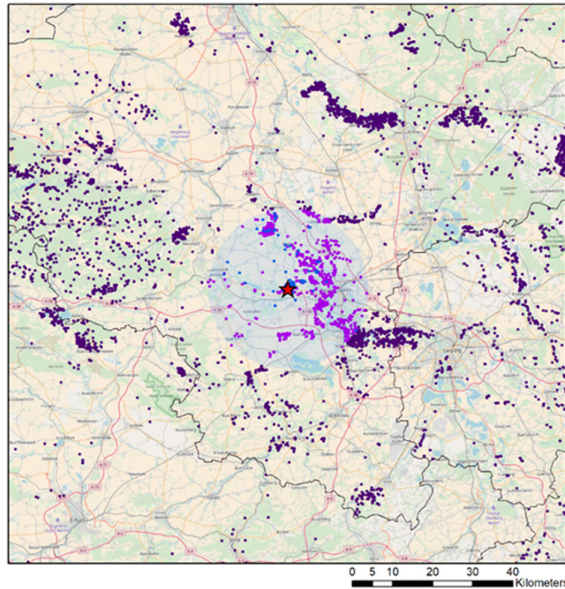
Assessing the species pool size and composition

We constructed empirical sample-based rarefaction curves for each of the 6319 target plots in the calcareous grassland data set. We fitted these curves empirically, instead of using analytical estimators for two reasons. First, because our data do not include individual species counts and stem from plots having unequal size, so that we could not calculate the typical individual-based analytical estimators of asymptotic species richness. Second, because our plots have different sampled areas, thus violating the assumptions of typical sample-based rarefaction curves, which require that all plots have an equivalent sampling-intensity (Gotelli and Colwell 2001, Hsieh et al. 2016). We built sample-based rarefaction curves combining all plots within a given geographic radius and compositional similarity from each target plot. We first sampled all plots within circles of different radii (10, 20, 50 or 100 km) around the target plot. We then selected all plots q being compositionally similar to the target plot p based on the matrix of species occurrence probabilities, using the Bray–Curtis dissimilarity index (2):

Workflow for calculating site-specific species pools

1- Calculate occurrence probabilities for all species in every plot in the database, using Beals' index of sociological favorability.

2. Select a target plot (in the example here ID 21752 in GVRD, marked with a red asterisk in the map below).

3. Select all plots within a radius (e.g., 20 km in the example below) around the target plot (blue and magenta dots in the map below).

4. From those, select those plots that are compositionally similar to the target plot based on Beals' probabilistic species composition (e.g., using a dissimilarity threshold to the target of 0.2, blue dots in the map below, n=254).

Service Layer Credits: (c) OpenStreetMap and contributors, Creative Commons-Share A SA)

★ Target Plot
● Plots similar to target
● Plots within Radius
● Other Plots

Pool size estimation

394 ──── 7a

266 ──── 7b
232 ────

160 ──── 7c
152 ────
147 ────

5. Use these plots to construct sample-based rarefaction curves by randomly drawing 1, 2, 3, ... n plots and accumulating their species and areas.

6. Fit regression lines (e.g. Arrhenius, Gompertz, Michaelis-Menten) to 5. or estimate asymptotic species richness nonparametrically (e.g., iChao2)

7. Estimate the size S of the species pool from
   a) richness estimation (iChao2)
   b) calculating the asymptote for SARs in step 6 (e.g. Gompertz, Michaelis-Menten)
or (7c) calculate species richness for a given area (e.g. 1000 m²) using the curves from step 6.

8. Derive the composition of the species pool (7a,7b), or for a given sampling area (7c) by selecting the first S species ranked by their Beals' occurrence proability. For the example here (plot # 21752) and using Gompertz's asymptote, S is 232 species (for the full list see Supplementary material Appendix 1 Table A1).
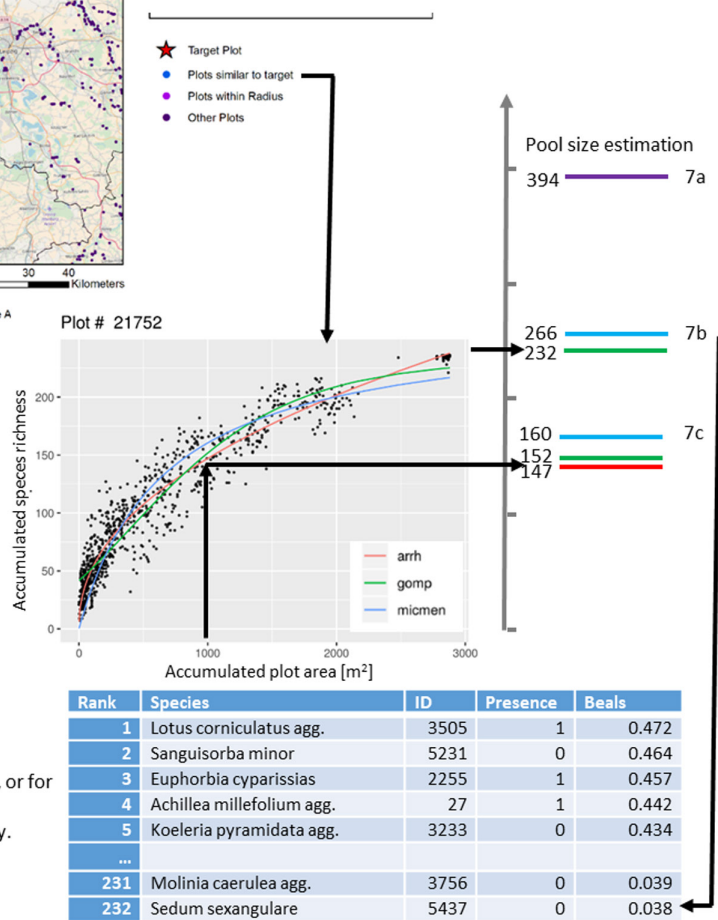
Plot # 21752

Rank	Species	ID	Presence	Beals
1	Lotus corniculatus agg.	3505	1	0.472
2	Sanguisorba minor	5231	0	0.464
3	Euphorbia cyparissias	2255	1	0.457
4	Achillea millefolium agg.	27	1	0.442
5	Koeleria pyramidata agg.	3233	0	0.434
...				
231	Molinia caerulea agg.	3756	0	0.039
232	Sedum sexangulare	5437	0	0.038

Figure 1. Workflow for estimating the size S and the composition of the site-specific species pool of an exemplary vegetation record (ID 21752 in GVRD). All plots within a radius of 20 km (magenta dots) that have a similarity to the target plot higher than 0.2 based on the Beals' occurrence probability of all species (blue dots, n = 254) were used to construct sample-based rarefaction curves and obtain the site-specific species pool (SSSP) of the target plot. The list of species in this particular SSSP is found in Supplementary material Appendix 1 Fig. A5.

$$d_{pq} = \frac{\sum_i^m |x_{pi} - x_{qi}|}{\sum_i^m (x_{pi} + x_{qi})} \qquad (2)$$

where m is the total number of species, x_{pi} and x_{qi} is the occurrence probability of species i in plot p and q, respectively. We varied the dissimilarity threshold of plots in the neighbourhood, using $d_{pq} = 0.1$, 0.2, 0.5 and 1, the latter being equivalent of using no threshold at all. We retained only those vegetation records in the selection that were more similar than these thresholds, with T being the number of plots finally selected. Testing different geographical radii and similarity thresholds allowed for accounting for the impact of the geographical filter and degree of habitat specificity on the species pool, respectively. We calculated the Bray–Curtis dissimilarity not on the species' cover in a plot but on their occurrence probability x_{pi} (formula 1), thus applying formula 2 across all m = 2432 non-singleton species in GVRD. Basing the dissimilarity on Beals' index had not only the effect to combine plots that potentially shared the same species, but also made the calculation of d_{pq} independent of species richness, and plot size. For an example of a plot selection with r = 20 km and $d_{pq} = 0.2$ see Fig. 1.

We then built rarefaction curves, by plotting accumulated species richness against accumulated area for different random combinations of 2, 3, …. n plots, which is conceptually similar to type IIIB of the six species-area curves described by (Scheiner 2003). As it is hugely time-consuming to calculate all possible combinations for large sample sets n (in Fig. 1, n = 254), we drew the number of plots to be combined from a series of 10 intervals equally spaced on a logarithmic scale, together with the endpoint, i.e. all plots together (e.g. 1, 2, 3, 6, 12, 22, 40, 74, 137, 253, 254 for the case of Fig. 1). We randomly drew 99 combinations of each number of plots (or the maximum number of possible combinations if that number was < 99) and used their pooled species richness and sampled accumulated areas for building the rarefaction curve. To test on how strongly the construction of the rarefaction curve depended on the number of selected plots, we varied the required minimum number of selected plots from 5, 10, 20 to 50, but found little differences in the outcome (results not shown). Thus, in the following we applied a threshold of at least 10 plots surrounding a target plot to build sample-based rarefaction curves.

We compared three main ways to assess the size of SSSPs, based on a) nonparametric estimations based on incidence data, b) on asymptotic rarefaction curves where the asymptote gives the number of species in the SSSP and c) non-asymptotic curves with predefined habitat areas (= sampling grains) as cut-off values.

The asymptotic b) and non-asymptotic c) curves were based on different models (Dengler 2009) that allow setting a common grain (i.e. estimate species richness for a certain area); the power function (Arrhenius 1921):

$$S = kA^z \qquad (3)$$

where S is the species number, the parameter k is species richness per unit area (i.e. 1 m²) and z the slope of the curve; the Gompertz function:

$$S = ae^{-b1b2^A} \qquad (4)$$

where a describes asymptotic richness and $b1$ and $b2$ determine intercept and slope of the curve; the Michaelis–Menten function (Kluth and Bruelheide 2004):

$$S = \frac{aA}{b + A} \qquad (5)$$

where a describes asymptotic species richness and b is the sampling effort required to exactly observe half of the asymptote value a.

The regression lines for the three models for the exemplary plot are shown in Supplementary material Appendix 1 Fig. A1. For every target plot, we compared the goodness of fit of the three models (formulas 3–5) by calculating the Akaike information criterion (AIC) (Dengler 2009) and plotted the distribution of the AIC values across all 6319 plots.

Finally, the nonparametric approach a) was based on the improved lower bound nonparametric richness estimator for incidence data developed by Chiu et al. (2014), even if our data violate the assumption of all units having an equal sampling effort (since plot sizes are largely different). The iChao2 index extends the work of (Chao 1984), and is defined as:

$$\hat{S} = \text{Chao2} + \frac{(T-3)}{4T} \frac{Q_3}{Q_4} \times \max\left[Q_1 - \frac{(T-3)}{2(T-1)} \frac{Q_2 Q_3}{Q_4}, 0 \right] \qquad (6)$$

where T is the number of sampling units, Q_k is the number of species that are observed in exactly k sampling units, $k = 0, 1, …, T$, based on presence/absence data, and Chao2 is defined as:

$$\text{Chao2} = \begin{cases} S_0 + (T-1)Q_1^2 / (2TQ_2) & \text{if } Q_2 > 0 \\ S_0 + (T-1)Q_1(Q_1 - 1) / (2T) & \text{otherwise} \end{cases} \qquad (7)$$

where S_0 is the total number of species observed (i.e. the sum of all species in the selection of T plots). We calculated the iChao2 index using the R package 'SpadeR' (Chao et al. 2019).

To obtain the concrete list of species that belonged to a species pool, we selected as many species from the list of decreasing occurrence probability as the expected species pool size S. For each plot, in practice, we first ranked all species based on their probability of occurrence (= Beals' index of sociological favourability). We then estimated the size S of

the species pool for each plot according to formula 4, 5 or 6, which we use as a cut-off to produce the list of all the species in the pool for each plot, as illustrated in Supplementary material Appendix 1 Fig. A5.

Effect of plot size, species occurrences, spatial extent and plot pre-selection

In contrast to Brown et al. (2019), we do not have a 'true' reference as a benchmark. We therefore compared the results based on our methods with those obtained when running three additional tests.

First, we checked how varying plot sizes affected the estimation of SSSPs. We did this by comparing our results to those obtained assuming all plots had a plot size equivalent to the mean plot size across the dry grassland dataset (21.6 m^2).

Second, we compared the results obtained using Beals' occurrence probabilities with those based on species presence/absence (p/a) data. We tested both for the effect of using p/a data to identify the plots with a similar species composition, and for the effect of using species frequencies to rank species when deriving a target plot's species pool composition. To standardize the conditions in these comparisons, we used a radius of r = 20 km for selecting similar plots. For p/a data we used a Sørensen dissimilarity (which is the presence/absence equivalent of Bray–Curtis dissimilarity) threshold of 0.75, as it resulted in a similar number of neighbouring plots and observed total number of species as using a Bray–Curtis dissimilarity threshold of 0.2 on Beals' occurrence probabilities. We then compared the estimation of the SSSPs using the three different models (iChao2, Gompertz, Michaelis-Menten). We also tested whether the Beals and p/a approach resulted in a different distance decay in the dissimilarity of the species composition within the selected plots around a target plot. This was achieved by relating all pairwise dissimilarities to geographic distances and calculating the Mantel statistics. We then compared the distribution of the Mantel correlation coefficients of all target plots based on Bray–Curtis dissimilarity and Beals co-occurrence probabilities with those based on Sørensen dissimilarity and p/a. As Beals occurrence probabilities are known to depend on the overall frequency of a species in the database (De Caceres and Legendre 2008), we plotted the number of times a species was included in a SSSP against the overall frequency of the species in the database. We did this both when using Beals occurrence probabilities and species frequencies to derive the species composition of SSSPs.

Finally, we assessed the effect of pre-selecting plots belonging to the same habitat when estimating SSSPs. All previous results were based on a subset of dry grassland (Festuca-Brometea) records from the whole German vegetation-plot database. We asked how pool size and composition depend on this preselection and compared the previous results with those obtained when using all records in the database to construct sample-based rarefaction curves. We used the same thresholds reported above (0.2 for Bray–Curtis dissimilarities

based on Beals occurrence probabilities and 0.75 for Sørensen dissimilarity based on p/a). We then plotted pool sizes against each other.

Applications: scalable species richness and species pool composition maps

We show two possible applications of how the approach based on Beals occurrence probabilities can be used for the exploration of biogeographical patterns. First, we produced a set of scalable maps of species richness at different spatial grains, fitting different models to the sample-based rarefaction curves: Arrhenius, Gompertz and Michaelis-Menten (formulas 3, 4 and 5, respectively) across a range of spatial grains, i.e. 1, 10, 100, 1000 m^2. In other words, after creating sample-based rarefaction curves for each plot, we used these curves to predict the species richness of the respective plot assuming increasing sampled areas. For visualization purpose, we averaged estimated species richness over hexagonal areas of approximately 850 km^2 and plotted the result at the scale of entire Germany.

Second, we produced a standardized map showing differences in the species pool composition of calcareous grasslands across Germany. We also visualized the differences in species composition by subjecting the site-specific species pool to non-metric multidimensional scaling (NMDS), using three dimensions. For this purpose, we used the full list of the 1067 species in the calcareous grassland records with their probability of occurrence, and not the S species with highest ranks. The degree to which the SSSPs differed on the three NMDS dimensions was visualized by ternary colour gradients, using the 'tricolore' package (Schöley and Kashnitsky 2018). All programming was done in R (ver. 3.6.1, R Development Core Team). The code used for calculating sample-based rarefaction curves, and the plot-specific species pool is available as an R package and maintained in GitHub (<https://github.com/idiv-biodiversity/SpeciesPool>).

Results

Comparing different richness estimators

Across all plots of the example dataset of dry grassland plot records, the average number of records used to construct the site-specific species pools (SSSPs) from rarefaction curves was 334, ranging from 10 to 1847. The number of species in the SSSPs differed between the three models, with mean pool sizes of 202, 262 and 399 for Gompertz, Michaelis-Menten and iChao2, respectively (Fig. 2). Both asymptotic curves were highly correlated with each other and resulted in sizes of the SSSP that were about 63% (Gompertz) and 66% (Michaelis-Menten) smaller than those estimated with iChao2. The goodness of fit of the three models was very similar, as shown by the high correlation of the plots' AIC. However, the Arrhenius function
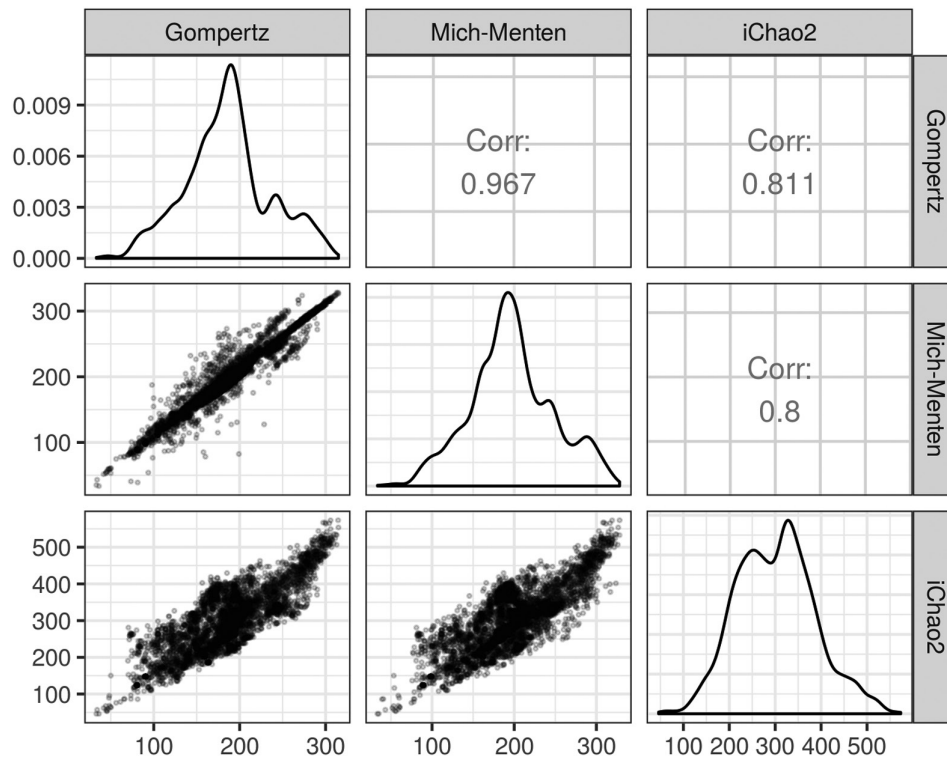
Figure 2. Correlation of the different methods of richness estimation of the species pools of all calcareous grassland plots (n = 6319) in the German Vegetation Reference Database (GVRD), using three different approaches (i.e. Gompertz, Michaelis-Menten and iChao2, corresponding to formulas 4, 5 and 6, respectively).

outperformed the Gompertz and Michaelis-Menten functions, as seen in the ΔAIC values when comparing the three models across all 6319 plots (Supplementary material Appendix 1 Fig. A2). The difference in AIC values becomes more prominent in plots with large species pools, which are those with high AIC values (Supplementary material Appendix 1 Fig. A2).

Comparing different extents for sampling neighbour plots

The size of the species pool varied only moderately when changing the extent of the sample area from 10 to 20 km radius, or when changing the Bray–Curtis dissimilarity threshold from $d_{pq} = 0.1$ to 0.2 (Fig. 3). Radii > 20 km or dissimilarity thresholds > 0.2 increased the number of plots used for SSSP size estimation, especially when using the Michaelis-Menten asymptotic curve or the iChao2 estimator. When using the most liberal thresholds (radius > 100 km radius and $d_{pq} = 1$ Bray–Curtis dissimilarity), the SSSP size approached 292, 458 and 730 for Gompertz, Michaelis-Menten and iChao2, respectively. However, none of these different parameters resulted in completely unrealistic SSSP sizes. Variation of the radius had much greater effects than variation of dissimilarity thresholds, probably as a consequence of pre-selecting dry calcareous grassland plots only. For this reason, in the following we focus on r = 20 km and $d_{pq} = 0.2$.

Effect of plot size, species occurrences, spatial extent and plot pre-selection

Using equal plot sizes had no effect on the estimation of the species pool size using the nonparametric richness estimator of iChao2, which is independent of plot size (Supplementary material Appendix 1 Fig. A3A), while mean species pool sizes increased by 31.4 and 37.1 species when using Gompertz and Michaelis-Menten, respectively (Supplementary material Appendix 1 Fig. A3B–C).

Selecting neighbouring plots based on species presence/absences, rather than Beals' occurrence probabilities, resulted in very similar total numbers of species and species pool sizes (Supplementary material Appendix 1 Fig. A4). Sizes of the species pools were on average estimated to be only slightly larger when based on presence/absence as compared to Beals, with a difference of 11.0, 3.7 and 4.2 species for iChao2, Gompertz and Michaelis-Menten, respectively. However, the distance decay of compositional dissimilarity was much steeper when selecting neighbouring plots based on presence/absence as compared to Beals occurrence probabilities (Supplementary material Appendix 1 Fig. A5). There were also differences in the composition of the species pools, when species were included either according to decreasing Beals occurrence probability or to decreasing frequency in the selected plots around the target plot (Supplementary material Appendix 1 Fig. A6). In the first case, the proportion of species predicted to occur either in
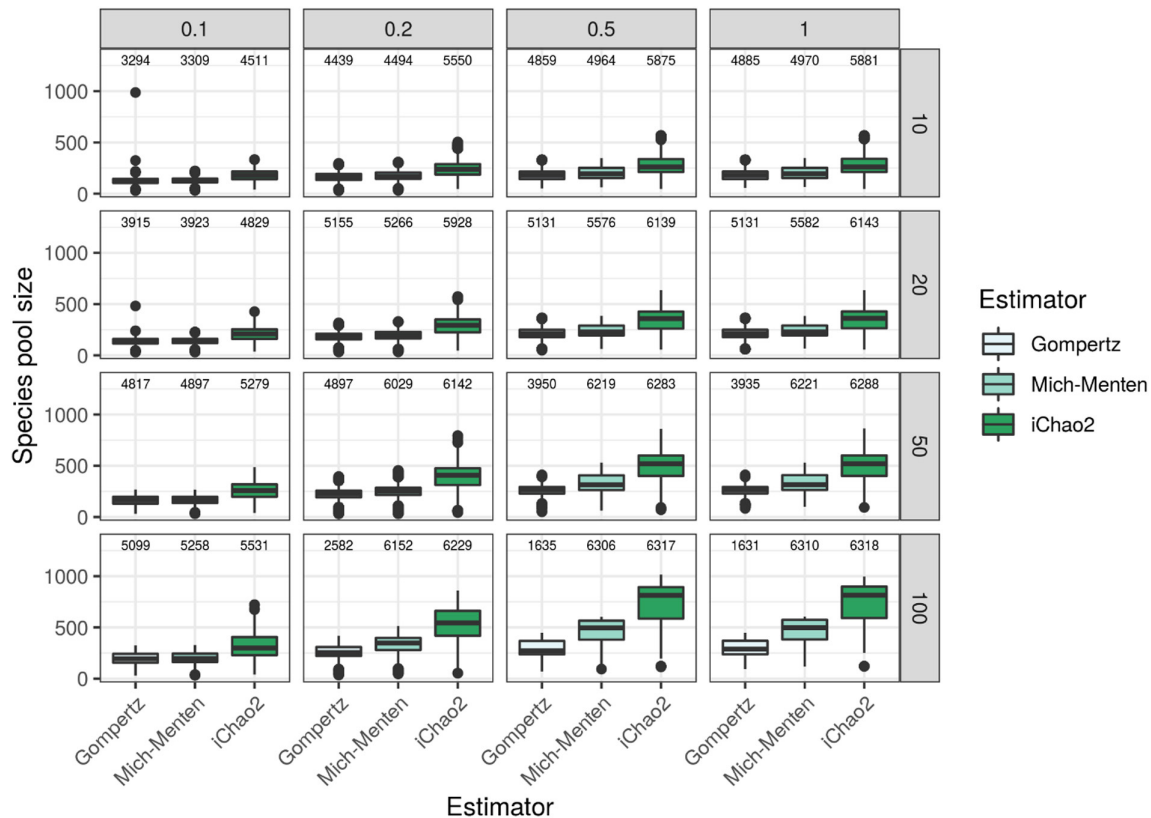
Figure 3. Boxplots of sizes of the site-specific species pools (SSSPs) for the three different approaches (i.e. Gompertz, Michaelis-Menten and iChao2). Rows show different extents of the sample area with r = 10, 20, 50 or 100 km around the target plot and columns show different Bray–Curtis dissimilarity thresholds for plots in the sample area to be included in the calculations, with d_{pq} = 0.1, 0.2, 0.5 and 1. The numbers in the graphs show the numbers of plots on which the calculations were based.

all species pools or never was higher than when using presence/absence data. Using presence/absence data, instead, systematically reduced the selection of frequent species, while favouring rare species. This resulted in a higher total number of species included in any species pool of 874 as compared to 829 for presence/absence and Beals, respectively. While with p/a the species frequency in the species pools was closer to the species frequencies in the total dataset (Supplementary material Appendix 1 Fig. A6C), basing the approach on Beals resulted in a more selective inclusion of species in the species pool (Supplementary material Appendix 1 Fig. A6B).

When using all 170 039 plots in the German Vegetation Reference Database (GVRD) for selecting the similar neighbours of each of the 6319 dry grassland plots, estimated species pool sizes increased for all methods (Supplementary material Appendix 1 Fig. A7). However, while the estimation based on Beals occurrence probabilities increased by 56.7, 52.0 and 53.7 species for iChao2, Gompertz and Michaelis-Menten, respectively, species pools were much larger when based on presence/absence, with increases of 88.4, 73.6 and 76.8, respectively. Thus, the presence/absence approach admitted more species from vegetation types other than dry grasslands into the species pool as compared to the Beals approach.

Applications: scalable species richness and species pool composition maps

Figure 4 shows the maps of estimated SSSP sizes in Germany. All three models predicted the largest pools for the region of the Kyffhäuser Mountains in Thuringia and Saxony-Anhalt (east Germany) and the Swabian Alb in Baden-Württemberg. Large pools were also found in southern Lower Saxony, North Hesse and the Franconian Alb in Bavaria. Except for differing in the absolute sizes of the SSSPs, the three models resulted in very consistent geographic patterns.

The scalable maps of species richness at different spatial grains (Fig. 5), showed regional differences between different models, especially at small grain. At the grain size of 1 m², richness ranged from 3 to 49 and from 8 to 90 when obtained from the Arrhenius and Gompertz model, respectively. In contrast, richness predicted from the Michaelis-Menten model ranged only between 1 and 6 species. In the Arrhenius model, the hotspot of highest richness shifted from the Swabian Alb to central Germany and finally to the Franconian Alb with increasing grain size. The Gompertz model identified richness hotspots in the Swabian Alb at the grain of 1 and 10 m², while the Kyffhäuser region was richest at the grain of 1000 m². At large grain size, the maps obtained from the Michaelis-Menten model converged with those of the Gompertz model. Here,
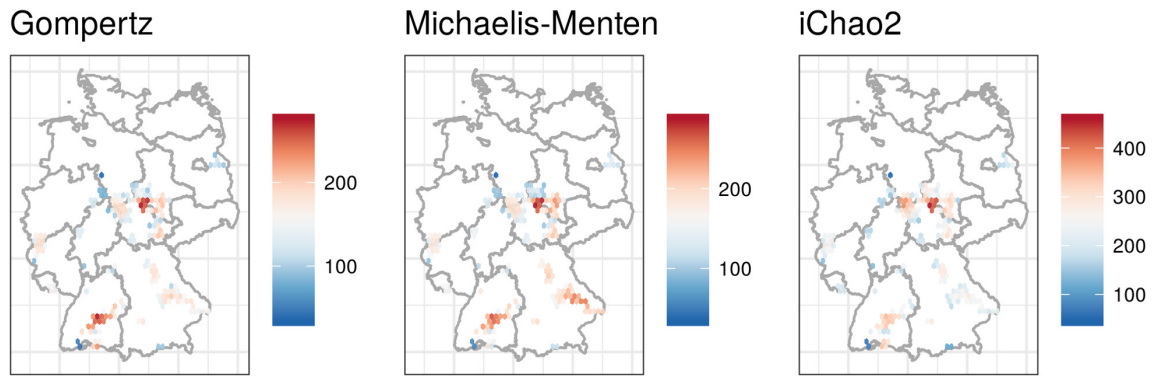
Figure 4. Maps of the site-specific species pool (SSSP) sizes for German calcareous grasslands, using either the asymptotes of sample-based rarefaction curves obtained fitting the Gompertz or Michaelis-Menten functions or by using the iChao2 nonparametric asymptotic richness estimator (formulas 4, 5 and 6, respectively). All calculations are based on a 20 km radius and a Bray–Curtis dissimilarity threshold of $d_{pq} = 0.2$.

predicted species richness from the Arrhenius model was extremely high, with up to 429 species on an area of 1000 m^2.

The different regions with limestone bedrock in Germany did not only differ in size, but also in composition of the species pools (Fig. 6). The NMDS ordination returned a very low stress (0.057). One gradient extended from blue in the W (Eifel Mts, Rhineland-Palatinate and Northrhine-Westphalia) to yellow in the NE (Odertal in Brandenburg) and SE (Franconian Alb in Bavaria), reflecting a gradient of climatic continentality. Another gradient with increasing proportion of pink in the Kyffhäuser region (Thuringia, Saxony-Anhalt) and Swabian Alb (Baden Württemberg) mainly reflected increasing proportion of species with submediterranean distribution ranges (Supplementary material Appendix 1 Table A2).

Discussion

This study proposes an approach to empirically define site-specific species pools on large databases containing community data. Previous approaches based on community datasets were generally constrained to select an arbitrary pool of samples (Belote et al. 2009, Chase et al. 2011, Kraft et al. 2011) or to pre-define a subset of data corresponding to a specific habitat or region (Jiménez-Alfaro et al. 2018, Sabatini et al. 2018). In contrast, the method presented here provides a standardized protocol for resampling a database in order to provide a SSSP which is not influenced by a-priori definitions of regions or habitats. The method can be directly applied to large vegetation databases (Lopez-Gonzalez et al. 2012, Chytrý et al. 2016, Bruelheide et al. 2019) but also to other community data sets with geo-referenced samples and similar properties.

Pool sizes based on Beals probabilities versus presence/absence

A key feature of our approach is the use of Beals' index to calculate probabilities of species occurrences in a specific site using the community composition recorded in such site.

This idea has already been applied in other attempts to define the species pool (Ewald 2002, Lewis et al. 2016). However, we based pool calculations on statistically defined richness estimators, which is a major improvement compared to arbitrary cut-off values (Lewis et al. 2016). Deriving Beals' index from the co-occurrence information of a much larger data basis, rather than only from the set of communities for which we want to assess the species pool, reduces the risk of missing species that potentially build the species pool but for some reasons were not recorded in the set of records used. Using Beals occurrence probabilities also allows to assess the composition of the species pool if the pool size is larger than the total number of species sampled in the surrounding of a target plot, which is often the case, for instance, when using Chao's (1984) nonparametric richness estimator. Of course, all vegetation records in a database may not reflect the complete flora of a region and rare species of a regional checklist may still be missing. As rare species would also have a low probability to occur in a given plot, they may not be included in the species pool. However, studies focusing on rare species can directly compare their occurrence probabilities across sites, irrespective of their inclusion in the SSSP.

Using Beals occurrence probabilities also has the desirable property of downplaying the role of accidental species. This became obvious when we compared the species pool compositions to those obtained by using species presence/absence, where the inclusion of species in the species pools followed more closely their overall frequency in the dataset. When using Beals occurrence probabilities for selecting plot records for SSSP size estimations, species with high frequencies were much more often represented in the SSSP. This resulted in a much weaker distance decay, compared to using presence/absence data. Together, these results show that using Beals occurrence probability returns SSSP compositions that are much less dependent on the way neighbour plot records around a target plot are chosen and on their spatial distribution.

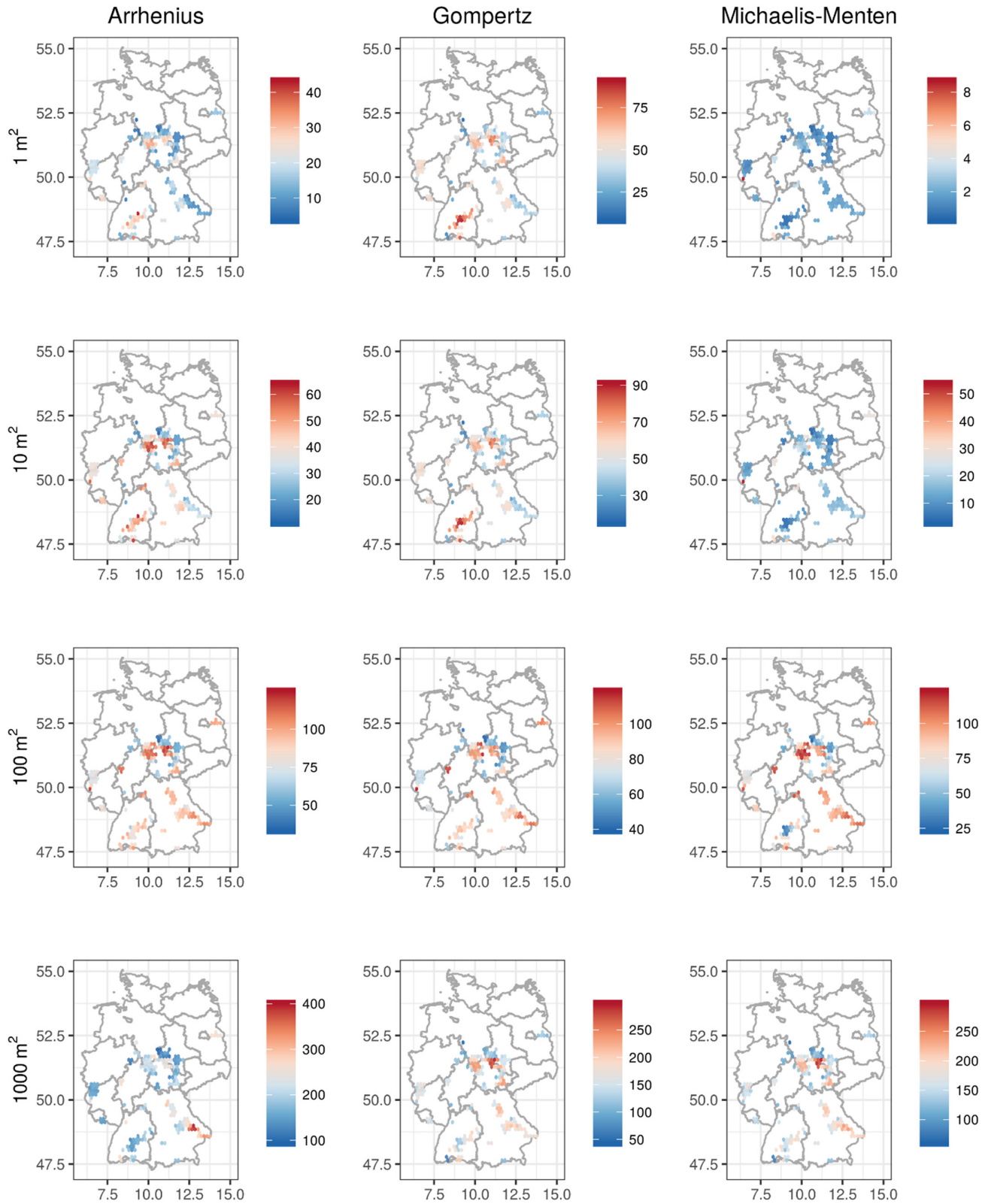Basing the probability of species occurrences on databases with large extent may result in including species whose

Figure 5. Richness maps of German dry calcareous grasslands at different spatial grains, i.e. at 1, 10, 100, 1000 m² (rows), using three different models of fitting the sample-based rarefaction curves, i.e. Arrhenius, Gompertz and Michaelis-Menten (columns), corresponding to formulas 3, 4 and 5, respectively. All calculations are based a 20 km radius and a Bray–Curtis dissimilarity threshold of $d_{pq} = 0.2$.
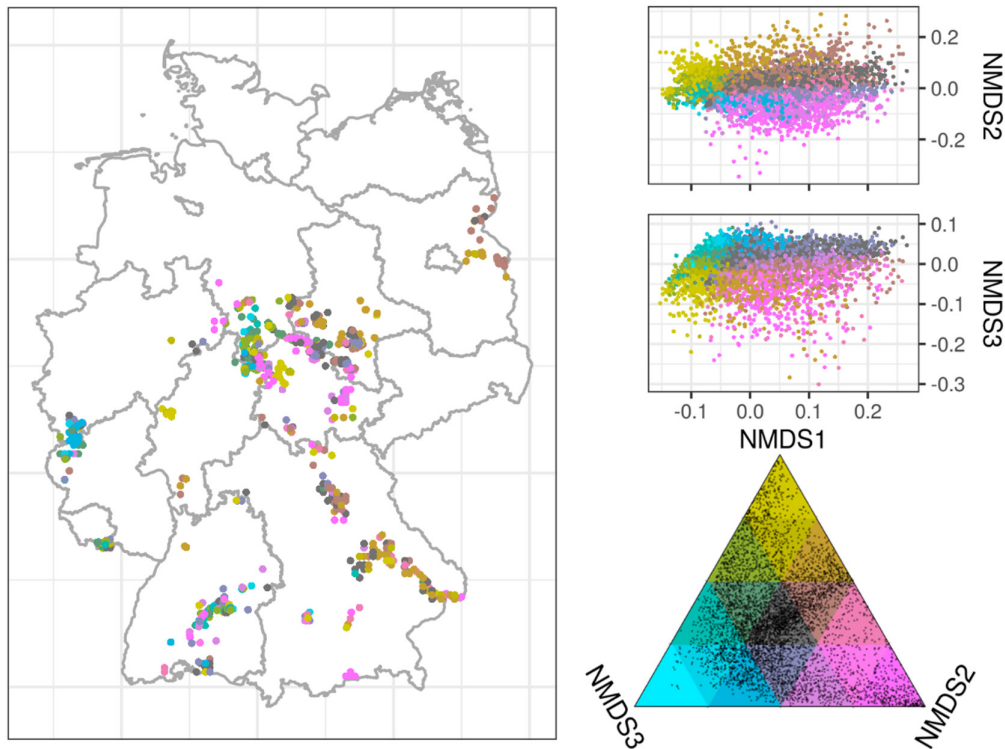
Figure 6. Spatial structure of the site-specific species pool (SSSP) composition of dry calcareous grasslands in Germany. The colours reflect the scores of the first three axes of a non-metric multidimensional scaling (NMDS) based on the occurrence probabilities of all 1067 species in all 6319 dry calcareous grassland plots. To create color-codes, we transformed the scores on the three NMDS axes to ranks, scaled their sum to unity and transformed these values to rgb scale. For the species scores on the NMDS axes see Supplementary material Appendix 1 Table A2.

location is outside their geographical distribution range. In principle, we mitigated this problem by using circular windows to limit the extent from which samples are selected. A proper choice of window size, however, critically depends on the dispersal abilities of the organisms in a specific study system. Still, one species may be predicted to co-occur with another species of the target site even if both species do not occur in a particular selection window. We do not see this as a disadvantage but as a powerful predictive element, since it may indicate that the species may potentially be able to colonize the target site even if its propagules have not yet arrived in the surrounding of this location. By experimentally planting plant individuals into plots in which they were absent Breitschwerdt et al. (2015) demonstrated that co-occurrence information based on Beals' index successfully predicts the establishment success of new colonizers. In extreme cases, our predictions of SSSPs may even be used for predicting the potential of non-native species to colonize a target plot, based on the co-occurrence pattern of these non-native species elsewhere in the database.

We also showed that using Beals occurrence probabilities does not require an a priori classification of plots into homogenous habitats. Further benefits of using a dissimilarity measure based on Beals' index instead of p/a for selecting the plot records for SSSP size estimations were a much lower distance decay and a much lower dependence on the preselection of

the records used. Allowing all plots from the whole GVRD database to be potentially included in the selection resulted in much higher SSSP size estimation when using presence/absence as compared to Beals occurrence probabilities. From this we conclude that basing the dissimilarity threshold on Beals probabilities is advantageous for heterogeneous data that do not belong to the same vegetation type. Plots are grouped exclusively based on their species composition, which results in as many different habitat definitions as the plot records considered. One might even argue that this procedure makes the species pool definition habitat-independent. While this is theoretically true, in practice habitat definitions are needed for further analyses of site-specific species pools, such as describing geographical patterns or comparing different habitats (Willner et al. 2019), as we showed in our application on dry calcareous grasslands in Germany. It is important to point out that using the results of the same approach in other data sets will depend on data coverage. In general, using a database in which the habitat type of interest is scarcely represented may not give realistic results.

The role of the different parameters to construct SSSPs

We are aware that our standardization requires expert decisions on 1) the area around the target plot from which records are considered to construct the SSSP (radius r),

2) the dissimilarity threshold of records from this selection to be included (Bray–Curtis dissimilarity d_{pq}) and 3) the minimum number of records in the plot selection around the target plot needed for the construction of sample-based rarefaction curves. The decision on these thresholds is to some degree arbitrary and case-specific, as different extents and dissimilarity thresholds affect the degree of the habitat-specificity of the species pool definition. The larger the extent and the larger the dissimilarity threshold, the less habitat-specific the SSSP is defined. However, we feel that this not a disadvantage but provides flexibility when dealing with different community types, which may vary in their definability. We also demonstrated that the approach is robust to using plots with varying plot sizes. The increase in estimated SSSP sizes when using the mean plot size ($21.6\,m^2$), as compared to using the exact sizes, is explained by a general decrease in the slope of the asymptotic rarefaction curves. This happened because the majority of plots was smaller than the mean size (median size was $10\,m^2$), and thus, the accumulated exact plots sizes started on average with smaller grain sizes. However, we do not consider an increase in SSSP sizes of 31.4 or 37.1 a large change, compared to using different radii of plot selection or dissimilarity thresholds.

Applications: scalable species richness and species pool composition maps

Finally, our approach allows scaling species richness across spatial grains, thus explicitly considering the scale-dependency of biodiversity estimations. While nested plots series are advantageous, at least when based on the same methodology and grain sizes, such data are only available for some selected types of vegetation (e.g. see the GrassPlot database, Dengler et al. 2018). In contrast, the SSSP approach can be applied to any type of data set as long as the samples have a sufficient spatial coverage to allow the creation of sample-based rarefaction curves. Compared to nested plot designs (type I species–area curves in Scheiner 2003), as commonly used in vegetation science (Dengler 2009), the SSSP approach consistently predicts higher species pool sizes. The underlying reason is a higher spatial extent on which the sample-based rarefaction curves are built. Given the same total area covered, dispersed samples are likely to contain more species than contiguous samples. Compared to basing the size-specific pool size estimation on a single nested plots series, however, our approach mitigates the risk of excluding species whose abiotic and biotic niche matches the conditions of that site but never arrived at this particular location due to limitations in dispersal.

We compared different mathematical functions to model sample-based rarefaction curves, and found all of them to have advantages and disadvantages. While the Arrhenius power function (formula 2), which was found to perform best in nested plots series (Dengler 2009), predicted very variable hotspot patterns at different grain sizes and unrealistically high richness values at a grain size of $1000\,m^2$, the asymptotic Michaelis-Menten and

Gompertz functions gave more consistent results. Here, the Michaelis-Menten function predicted unrealistically low richness values at the smallest grain of $1\,m^2$, probably because such small plot sizes are rare in the database. In contrast, the Gompertz function did not encounter this problem, but has the potential disadvantage of returning unrealistic results at the smallest grain sizes, because it does not approach a species number of 0 when area becomes minimal (formula 5). Extrapolating predictions at scales outside of the range observed in the data, however, always demands extreme caution. Even with this restriction, the SSSP approach provides a robust approach to bridge biodiversity estimations across different scales. Correcting for plot size when comparing species richness at a specific spatial grain is a clear advantage of our method, because survey protocols and plot sizes have varied over the years, since the early days in vegetation science. Thus, we now have a powerful tool to make predictions of species richness, which is independent of sampling resolution.

We note that our approach for calculating site-specific species pools opens the way to many applications. By bridging the gaps across scales, the SSSP approach sheds light on the multi-scale effect of some environmental filters (e.g. as climate) on species richness and coexistence patterns (Levin 1992, Bruelheide et al. 2018). In these regards our approach complements previous attempts of producing scalable maps of species richness (Keil and Chase 2019). As such, our method may help highlighting regional anomalies, i.e. areas where species accumulation is more or less steep than predicted on the basis of species pools alone, which could inform conservation priority-setting (Jenkins et al. 2013). Finally, an important feature of the SSSP is the provision of the species composition of the species pool, which can be used to compare patterns of β-diversity at different scales and to relate them to different drivers. Information on the compositional identity of species pools, may also be combined with functional trait data, to analyse the species pool functionally, for instance relating the degree of filtering between the species pool and the local community to dispersal, establishment and competition traits. This may shed light on those assembly processes that shape the composition of local communities.

Data availability statement

The data used are part of the German Vegetation Reference Database and can be directly accessed there.

References

Arrhenius, O. 1921. Species and area. – J. Ecol. 9: 95–99.

Beals, E. W. 1984. Bray–Curtis ordination: an effective strategy for analysis of multivariate ecological data. – Adv. Ecol. Res. 14: 1–55.

Belote, R. T. et al. 2009. Disturbance alters local–regional richness relationships in Appalachian forests. – Ecology 90: 2940–2947.

Breitschwerdt, E. et al. 2015. Do newcomers stick to the rules of the residents? Designing trait-based community assembly tests. – J. Veg. Sci. 26: 219–232.

Brown, J. J. et al. 2019. A novel method to predict dark diversity using unconstrained ordination analysis. – J. Veg. Sci. 30: 610–619.

Bruelheide, H. et al. 2018. Global trait–environment relationships of plant communities. – Nat. Ecol. Evol. 2: 1906–1917.

Bruelheide, H. et al. 2019. sPlot – a new tool for global vegetation analyses. – J. Veg. Sci. 30: 161–186.

Carstensen, D. W. et al. 2013. Introducing the biogeographic species pool. – Ecography 36: 1310–1318.

Chao, A. 1984. Nonparametric estimation of the number of classes in a population. – Scand. J. Stat. 11: 265–270.

Chao, A. et al. 2019. SpadeR: species-richness prediction and diversity estimation with R. – R package ver. 0.1.1, <https://CRAN.R-project.org/package=SpadeR>.

Chase, J. M. et al. 2011. Using null models to disentangle variation in community dissimilarity from variation in α-diversity. – Ecosphere 2: art24.

Chase, J. M. et al. 2018. Embracing scale-dependence to achieve a deeper understanding of biodiversity and its change across communities. – Ecol. Lett. 21: 1737–1751.

Chiu, C. H. et al. 2014. An improved nonparametric lower bound of species richness via a modified good-turing frequency formula. – Biometrics 70: 671–682.

Chytrý, M. et al. 2016. European Vegetation Archive (EVA): an integrated database of European vegetation plots. – Appl. Veg. Sci. 19: 173–180.

Cornell, H. V. and Harrison, S. P. 2014. What are species pools and when are they important? – Annu. Rev. Ecol. Evol. Syst. 45: 45–67.

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

Dengler, J. 2009. Which function describes the species–area relationship best? A review and empirical evaluation. – J. Biogeogr. 36: 728–744.

Dengler, J. et al. 2018. GrassPlot – a database of multi-scale plant diversity in Palaearctic grasslands. – Phytocoenologia 48: 331–347.

Eriksson, O. 1993. The species-pool hypothesis and plant community diversity. – Oikos 68: 371–374.

Ewald, J. 2002. Multiple controls of understorey plant richness in mountain forests of the Bavarian Alps. – Phytocoenologia 32: 85–100.

Gotelli, N. and Graves, G. 1996. Null models in ecology. – Smithsonian Inst. Press.

Gotelli, N. J. and Colwell, R. K. 2001. Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. – Ecol. Lett. 4: 379–391.

Graves, G. R. and Gotelli, N. J. 1983. Neotropical land-bridge avifaunas: new approaches to null hypotheses in biogeography. – Oikos 41: 322–333.

Graves, G. R. and Rahbek, C. 2005. Source pool geometry and the assembly of continental avifaunas. – Proc. Natl Acad. Sci. USA 102: 7871–7876.

Hsieh, T. et al. 2016. iNEXT: an R package for rarefaction and extrapolation of species diversity (Hill numbers). – Methods Ecol. Evol. 7: 1451–1456.

Jandt, U. and Bruelheide, H. 2012. German vegetation reference database (GVRD). – Biodivers. Ecol. 4: 355–355.

Jenkins, C. N. et al. 2013. Global patterns of terrestrial vertebrate diversity and conservation. – Proc. Natl Acad. Sci. USA 110: E2602–E2610.

Jiménez-Alfaro, B. et al. 2018. History and environment shape species pools and community diversity in European beech forests. – Nat. Ecol. Evol. 2: 483–490.

Karger, D. N. et al. 2016. Delineating probabilistic species pools in ecology and biogeography. – Global Ecol. Biogeogr. 25: 489–501.

Keil, P. and Chase, J. M. 2019. Global patterns and drivers of tree diversity integrated across a continuum of spatial grains. – Nat. Ecol. Evol. 3: 390.

Kluth, C. and Bruelheide, H. 2004. Using standardized sampling designs from population ecology to assess biodiversity patterns of therophyte vegetation across scales. – J. Biogeogr. 31: 363–377.

Kraft, N. J. B. et al. 2011. Disentangling the drivers of β diversity along latitudinal and elevational gradients. – Science 333: 1755–1758.

Ladouceur, E. et al. 2018. Native seed supply and the restoration species pool. – Conserv. Lett. 11: e12381.

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

Lessard, J.-P. et al. 2012. Inferring local ecological processes amid species pool influences. – Trends Ecol. Evol. 27: 600–607.

Levin, S. A. 1992. The problem of pattern and scale in ecology: the Robert H. MacArthur award lecture. – Ecology 73: 1943–1967.

Lewis, R. J. et al. 2016. Estimating dark diversity and species pools: an empirical assessment of two methods. – Methods Ecol. Evol. 7: 104–113.

Lopez-Gonzalez, G. et al. 2012. ForestPlots. Net-managing permanent plot information across the tropics. – Biodivers. Ecol. 4: 95–103.

Münzbergová, Z. and Herben, T. 2004. Identification of suitable unoccupied habitats in metapopulation studies using co-occurrence of species. – Oikos 105: 408–414.

Pärtel, M. et al. 1996. The species pool and its relation to species richness: evidence from Estonian plant communities. – Oikos 75: 111–117.

Pärtel, M. et al. 2011. Dark diversity: shedding light on absent species. – Trends Ecol. Evol. 26: 124–128.

Pärtel, M. et al. 2013. Community completeness: linking local and dark diversity within the species pool concept. – Folia Geobot. 48: 307–317.

Ricklefs, R. E. 1987. Community diversity – relative roles of local and regional processes. – Science 235: 167–171.

Sabatini, F. M. et al. 2018. Beta-diversity of central European forests decreases along an elevational gradient due to the variation in local community assembly processes. – Ecography 41: 1038–1048.

Scheiner, S. M. 2003. Six types of species–area curves. – Global Ecol. Biogeogr. 12: 441–447.

Schöley, J. and Kashnitsky, I. 2018. tricolore: a flexible color scale for ternary compositions. – R package ver. 1.2.0, <https://CRAN.R-project.org/package=tricolore>.

Srivastava, D. S. 1999. Using local–regional richness plots to test for species saturation: pitfalls and potentials. – J. Anim. Ecol. 68: 1–16.

Suding, K. N. 2011. Toward an era of restoration in ecology: successes, failures, and opportunities ahead. – Annu. Rev. Ecol. Evol. Syst. 42: 465–487.

Taylor, D. R. et al. 1990. On the relationship between r/K selection and environmental carrying capacity: a new habitat templet for plant life history strategies. – Oikos 239–250.

Willner, W. et al. 2019. Formalized classification of the semi-dry grasslands of central and eastern Europe. – Preslia 91: 25–49.

Zobel, M. 1997. The relative of species pools in determining plant species richness: an alternative explanation of species coexistence? – Trends Ecol. Evol. 12: 266–269.

Zobel, M. 2016. The species pool concept as a framework for studying patterns of plant diversity. – J. Veg. Sci. 27: 8–18.