Estimating dark diversity and species pools: an empirical assessment of two methods

Rob J. Lewis*, Robert Szava-Kovats and Meelis Pärtel

Institute of Ecology and Earth Sciences, University of Tartu, Lai 40, Tartu 51005, Estonia

Summary

1. Species absent from a community but with the potential to establish (dark diversity) are an important, yet rarely considered component of habitat-specific species pools. Quantifying this component remains a challenge as dark diversity cannot be observed directly and must be estimated. Here, we empirically test whether species ecological requirements or species co-occurrences provide accurate estimates of dark diversity.

2. We used two spatially nested independent datasets, one comprising 3033 samples of coastal grassland vegetation from 4 m^2 and 200 m^2 plots from Scotland, UK, and another comprising 780 samples of forest vegetation plots from 30 m^2 and 500 m^2 plots from Switzerland. Dark diversity for each of the smaller scaled plots was estimated through investigating the degree of (i) similarity in ecological requirements (measured as Ellenberg values); and (ii) co-occurrence likelihood. Estimates were validated using species from the larger spatial scales. Estimates were further validated using observations from all larger scale plots surrounding a focal assemblage within a 2 km (Scottish grassland) and 10 km (Swiss forest) radius.

3. The co-occurrence method was shown to be more accurate resulting in far fewer negative mismatches (i.e. species observed but not predicted), as well as higher proportions of observed and predicted species, relative to the Ellenberg method. Of the species observed in the large-scale samples, 18% were estimated as part of the smaller scale dark diversity via the co-occurrence approach relative to 8% for the Ellenberg method for both the Scottish and Swiss data, respectively. These values increased to 67% & 60% and 32% & 35%, respectively, across all observations within a 2 km (Scottish grasslands) and 10 km (Swiss forests) radius.

4. The study demonstrates that dark diversity for a community can be successfully estimated using readily available data, through exploring species co-occurrence patterns. This work substantiates that habitat-specific species pools can be accurately quantified and should prove valuable for understanding underlying community processes and improving our knowledge of the mechanisms governing species co-existence.

Key-words: Beal's smoothing, co-occurrence, ecological requirements, Ellenberg, method comparison, species beta-niche

Introduction

Species present in a community can be viewed as a subset of all species theoretically able to exist in that community (Pärtel 2014). This idea is a refinement of the species-pool hypothesis (for a review see Cornell & Harrison 2014), in that communities maintain a theoretical suite of species that have the potential to establish and co-occur (Eriksson 1993; Zobel 1997). By definition, a species pool, irrespective of scale, is always habitat specific inasmuch as it excludes species that would be environmentally and ecologically unsuited to conditions present in the target community. Species absent from a target community, but which remain part of the 'habitat-specific' species pool constitutes dark diversity (Pärtel, Szava-Kovats & Zobel 2011). Such species maintain the potential to establish and understanding reasons for its absence is ecologically meaningful. Dark diversity therefore can serve as a valuable biodiversity

metric (Pärtel 2014). Habitat-specific species pools (hereafter species pools) have been demonstrated to help disentangle trait dissimilarity (i.e. convergence vs divergence patterns) resulting from fine-scale biotic processes (e.g. phenotypic exclusion), enhancing our understanding of community assembly (de Bello *et al.* 2012). Understanding species pools in terms of its partitionable additive components (i.e. observed and absent species) also permits measures of community completeness, that is a logistic expression (*ln(observed diversity/dark diversity)*; Pärtel, Szava-Kovats & Zobel 2013). Such an approach has previously been postulated as a better means of comparing biodiversity among regions and habitats (de Bello *et al.* 2010), and should be useful for quantifying habitat degradation and monitoring restoration success (Suding 2011).

Identifying absent species from regional, local or community pools is, however, not straightforward. It has been proposed that four major scale-dependent processes – speciation, dispersal, selection and drift – operate to govern species distribution patterns and in turn the beta-niche (i.e. the region of a species' niche that corresponds to the habitat(s) where it is found;

*Correspondence author. E-mail: robert.lewis@ut.ee

Silverton *et al.* 2006) to which individual species belong (Vellend 2010). Arbitrarily delineating species pools, through geographic species inventories, is one approach, but one that fails to consider these fundamental ecological constraints, ignoring inherent spatial and temporal variability among species pools within regions (Lessard *et al.* 2012). Furthermore, it is not possible for inventories to identify those species that constitute a community's dark diversity, which by definition cannot be directly observed. One should also consider the sample effort. Species observations typically represent snapshots of community compositions, potentially overlooking hidden species (see Pärtel 2014) that are either cryptic or dormant (Klimesova & Klimes 2007). Disentangling between dark diversity and those species that have simply been overlooked is not possible from a single observation study and requires repeat sampling.

At the macroscale, recent attention has been paid to accurate estimation of biogeographic species pools (e.g. dispersion fields; Carstensen *et al.* 2013), accounting for a species site-specific dispersion probability. However, beyond the biogeographic scale, species pool estimates must consider a species affinity to the local and regional environmental conditions, accounting for a species site-specific establishment probability. One method is to consider ecological requirements of species based on expert evaluation. Regional floras and faunas can be broadly linked to habitat types; however, more quantitative data can often be sought. For example, for plant species, Ellenberg indicator values (Ellenberg *et al.* 1991) provide a semi-quantitative measure of species environmental thresholds and distributions, linking species occurrences with the environment. First demonstrated by Pärtel *et al.* (1996), the approach considers species with similar ecological requirements to have the ability to establish in similar habitats (Fig. 1). Alternatively, another method is to use species co-occurrence patterns, patterns that have in the past proven central to the development of ecological theories (Diamond 1975; Gilpin & Diamond 1982), and continue to remain at the forefront of ecological research (Veech 2013). Here, an absent species is considered part of the species pool if it typically co-occurs with observed species present in the focal community (Ewald 2002; Münzbergová & Herben 2004; Riibak *et al.* 2015; Ronk, Szava-Kovats & Pärtel in press; Fig. 2).

The species pools in which both these methods attempt to measure are the same and comparable, that is the realized beta-niche of plants (Silverton *et al.* 2006); however, the size of dark diversity and in turn the size of species pools can vary as a result of applying different statistically justified threshold values. For the Ellenberg method, these thresholds determine the limits of similarity (i.e. Euclidean distance in an x-dimensional ellipsoid hypervolume) based on standard deviations (Fig. 1). And, for the co-occurrence method, thresholds reflect co-occurrence probabilities (i.e. percentiles along a species distribution curve; Fig. 2). Broader thresholds, that is, larger standard deviations or small percentiles will result in larger estimates of dark diversity and *vice a* versa. The ability to adjust thresholds therefore adds great flexibility to the estimation methods. For example, population studies interested in detecting a single species would benefit from constraining the estimated size of dark diversity, resulting in greater confidence in dark diversity estimations. Alternatively, a lesser constraint may benefit community studies interested in investigating diversity patterns, proving a broad representation of a communities' dark diversity.
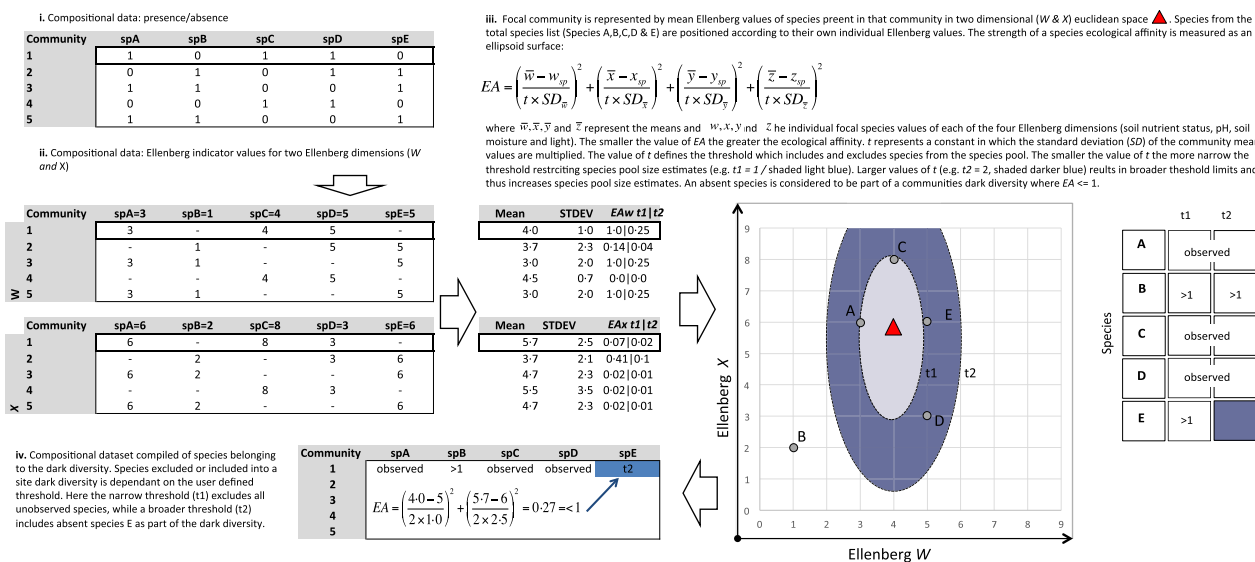

Fig. 1. Conceptual flow diagram illustrating how ecological requirements of species, (i.e. Ellenberg indicator values) are used to predict the dark diversity of a focal community. Illustrated are the sequential steps in the calculation process (i-iv) for a hypothetical assemblage (Community 1) comprising three species: species A (spA), species C (spC) and species D (spD) from a regional species list: species A, B, C, D & E. The influence of two user defined threshold values (i.e. a constant '*t*' in which the standard deviation (SD) is multiplied), one broad (t1) and one narrow (t2), is represented. For ease of interpretation, the example represents only the minimum number of dimensions of an *n*-dimensional Euclidean space, that is two dimensions (*W & X*).

i. Compositional data: presence/absence from which to calculate probability index

Community	spA	spB	spC	spD	spE
1	1	0	1	1	0
2	0	1	0	1	1
3	1	1	0	0	1
4	0	0	1	1	0
5	1	1	0	0	1

Nk

Species Freq	3	3	2	3	3

Njk

	spA	spB	spC	spD
spA				
spB	2			
spC	1	0		
spD	1	1	2	
spE	2	3	0	1

Si

Community	Richness
1	3
2	3
3	3
4	2
5	3

Beal's probability index =
$$P_{ij} = \frac{1}{S_i - I_{ij}} \sum_{k \neq j} \frac{N_{jk} I_{ik}}{N_k}$$

where: S_i = number of species at community i
I_{ij} = incidence (0, 1) of species j at community i
N_{jk} = number of joint occurrences of species j and k
I_{ik} = incidence (0, 1) of species k at community i
N_k = number of occurrences of species k

ii. Species × sample data matrix of probability indices (P_{ij}). For example, Community 1, SpE:

$$P_{ij} = \frac{1}{S_i - I_{ij}} \sum_{k \neq j} \frac{N_{jk} I_{ik}}{N_k} = \frac{1}{3} \sum_{k \neq j} \left[\frac{2 \times 1}{3}, \frac{3 \times 0}{3}, \frac{0 \times 1}{2}, \frac{1 \times 1}{3} \right] = 0.33$$

Community	spA	spB	spC	spD	spE
1	0·42	0·33	0·50	0·67	0·33
2	0·56	0·67	0·22	0·33	0·67
3	0·67	0·83	0·11	0·33	0·83
4	0·42	0·17	0·67	1·00	0·17
5	0·67	0·83	0·11	0·33	0·83

iii. For each species, a distribution curve is computed from sites where the species is actually present. User defined threshold values (i.e. percentiles) delineate between species included and excluded from the dark diversity. Here we illustrate the influence on one narrow (t1) and one broad (t2).

Conceptual illustration of step (iii) taking species E as an example. Illustrated is the influence of two threshold values, one narrow (t1) and one broad (t2).

Community	t1	t2
1		t2
2	observed	observed
3	observed	observed
4		
5	observed	observed

iv. Compositional dataset compiled of species belonging to the dark diversity. Species excluded or included into the dark diversity of a focal community is dependent on the user-defined threshold. Here the narrow threshold (t1) attributes species E to the dark diversity of community 4, while the broader threshold (t2) attributes species E to the dark diversity of community 4 and also community 1.

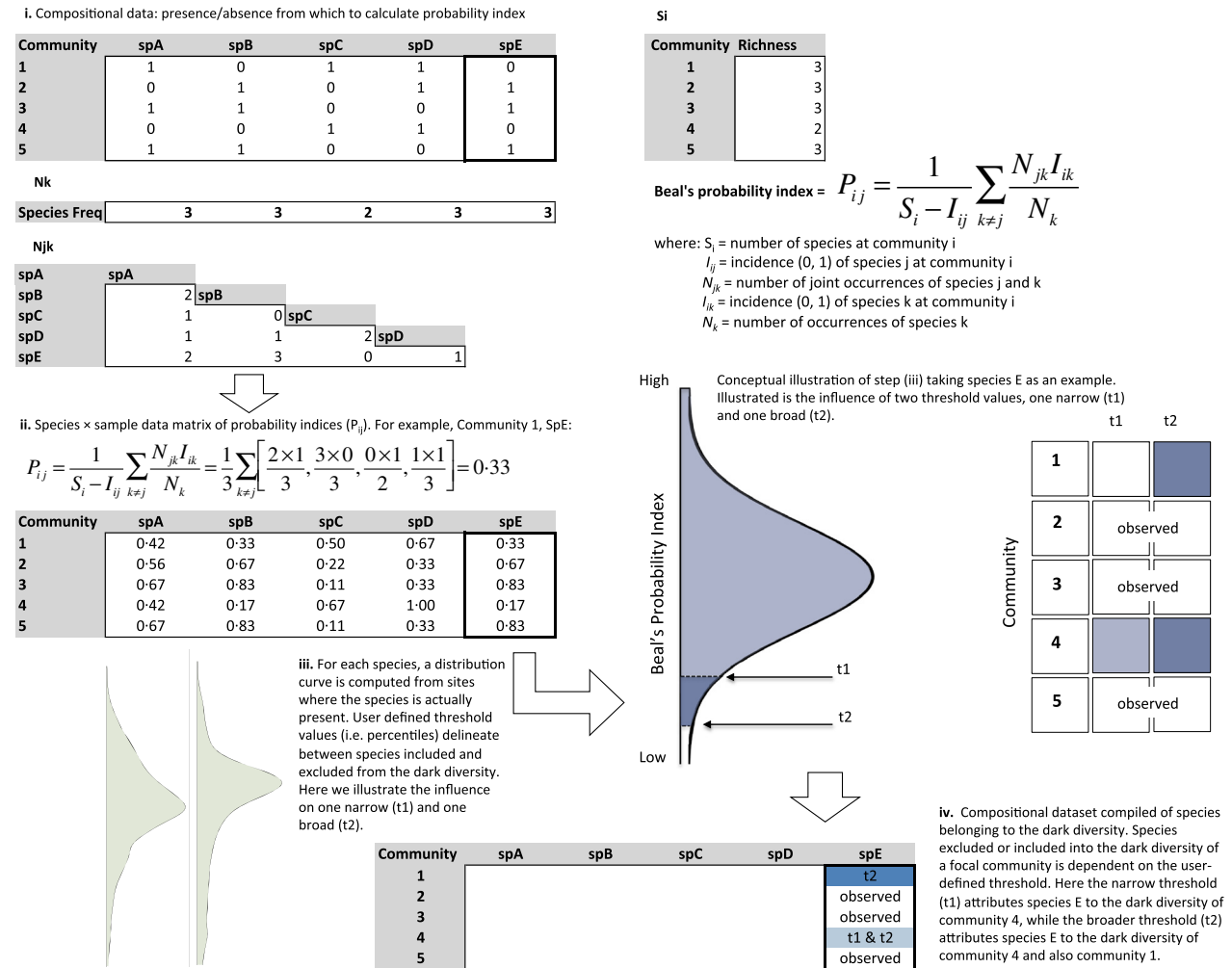| Community	spA	spB	spC	spD	spE
1					t2
2					observed
3					observed
4					t1 & t2
5					observed

Fig. 2. Conceptual flow diagram illustrating how species co-occurrence patterns calculated using Beal's smoothing multivariate transformation can be used to predict the dark diversity of a focal community. Illustrated are the sequential steps in the calculation process (i–iv), including the influence of two user defined threshold values, one broad (t1) and one narrow (t2).

While both of these analytical methods mediate the difficulties faced when measuring species pools (i.e. identifying species with high dispersal and establishment probability but are otherwise absent), an empirical evaluation of these methods has yet to be explored. Given that dark diversity can for most taxa only be estimated and not measured directly, a robust evaluation of methods and statistically justified thresholds seems an obvious and necessary pre-requirement before the dark diversity concept can be used with confidence for ecological and conservation purposes. Using two independent hierarchical nested data sets, we aim to empirically assess the accuracy of species ecological requirements and co-occurrences as two methods for estimating dark diversity and in turn quantifying species pools.

Methods

DATA

We used plant species data sampled from two very different vegetation types: coastal grassland vegetation of Scotland and forest vegetation of Switzerland, hereafter referred to as SG and SF, respectively. The data resulted from the Scottish Coastal Survey (1975–1977; Shaw, Hewett & Pizzey 1983) and a systematic sample of Swiss forests (Wohlgemuth et al. 2008). Both reflect a nested hierarchical sampling design, measuring nested community compositions from small to larger spatial scales. We use data from the smallest scale (SG = 4 m² & SF = 30 m²) to estimate dark diversity and data from the largest scale (SG = 200 m² & SF = 500 m²) to validate our estimations. For the respective habitat types, the 200 m² and 500 m² scales were suitably large enough to be representative of the community while not so large as to encompass different community types, maintaining environmental homogeneity. Plots where fewer than 3 species were observed at the smaller scale were removed from the analyses. In turn, the SG data set comprised 3033 nested plots encompassing 465 plant species, while the SF data set comprised 726 nested plots encompassing 410 plant species.

ESTIMATING DARK DIVERSITY

Our evaluation focused on two analytical methods employed to estimate dark diversity from compositional data: (i) ecological requirements of species, using Ellenberg indictor values; and (ii) species co-occurrence patterns.

(i) Ellenberg indicator values

Ellenberg indicator values originate from Central European flora reflecting plant species preferences in respect to several environmental factors. For the Scottish data, values were adjusted for British plants (Hill *et al.* 1999) for four environmental factors soil nutrient status, pH, soil moisture and light. A plant community habitat preference was determined through pinpointing its position in a four-dimensional hyperspace by calculating an average indicator value for all species in the community, in relation to the four habitat factors, each representing an individual axis. Because, Ellenberg indicator values only indicate the position of species optima along environmental gradients and contain no information on niche width, the ecological amplitude of species must be considered to prevent severe underestimation of the species pool. We therefore compiled species pool estimates through defining species ecological amplitudes as a chosen constant t multiplied by the standard deviations (SD) of the mean inside a four-dimensional (w, x, y, z) ellipsoid hypervolume.

$$EA = \left(\frac{\overline{w} - w_{sp}}{t \times SD_{\overline{w}}}\right)^2 + \left(\frac{\overline{x} - x_{sp}}{t \times SD_{\overline{x}}}\right)^2 + \left(\frac{\overline{y} - y_{sp}}{t \times SD_{\overline{y}}}\right)^2 + \left(\frac{\overline{z} - z_{sp}}{t \times SD_{\overline{z}}}\right)^2,$$

where $\overline{w}, \overline{x}, \overline{y}$ and $\overline{z}$ represent the means and w, x, y and z the individual focal species values of each of the four Ellenberg dimensions (soil nutrient status, soil pH, soil moisture and light). t represents a constant in which the standard deviation (SD) of the community mean values is multiplied. Species pools were in turn estimated with different threshold distances through changing the value of the constant t (0·5, 1·0, 1·5 and 2·0). We selected two measures indicative of what we term 'broad' and 'narrow' thresholds following results derived from exploring dissimilarity in dark diversity size estimates from the two methods (see below). The broad and narrow thresholds govern the estimated size of dark diversity, being either less constrained (i.e. larger) if broad, or, more constrained (i.e. smaller) if narrow. For the Ellenberg method, species were included as part of the dark diversity component based on values of EA being less than or equal to 1 for four environmental proxies: soil nutrient status, pH, soil moisture and light (for a two-dimensional example see Fig. 1).

(ii) Co-occurrence patterns

The probability of species co-occurrences was estimated using Beals index (Beals 1984), following the methods of Münzbergová & Herben (2004):

$$P_{ij} = \frac{1}{S_i - I_{ij}} \sum_{k \neq j} \frac{N_{jk} I_{ik}}{N_k}$$

where S_i is the number of species at site i, I_{ij} is the incidence (0, 1) of species j at community i.

N_{jk} is the number of joint occurrences of species j and k, I_{ik} is the incidence (0, 1) of species k at site i, and N_k is the number of occurrences of species k. For a given species, predicting dark diversity for sites where the species is absent required a species-specific threshold. Frequent species will systematically be assigned a higher Beals index values relative to less frequent species. This was achieved by modelling co-occurrence probabilities from sites where the focal species was actually present and setting threshold limits to the absolute minimum (*sensu* Münzbergová & Herben 2004). To exclude bias through outliers common in co-occurrence probability distributions (Botta-Dukat 2012), we also modelled the 1st, 5th and 10th percentile as a means of determining species-specific threshold limits. Species with co-occurrences probabilities outside the per-

centile range were excluded from dark diversity estimates (Fig. 2). Co-occurrence thresholds operated to the same effect as the thresholds distances assigned to the Ellenberg analysis, that is it determined whether species were included or excluded from the dark diversity. As with the Ellenberg method, two threshold values were selected indicative of 'broad' and 'narrow'. Threshold selection was based on minimum dissimilarity in dark diversity sizes (see below).

STATISTICAL ANALYSIS

Adjusting the dark diversity size

We performed a dissimilarity analysis to ensure we selected threshold values for both the Ellenberg and co-occurrence methods that result in reasonably similar sizes of dark diversity. This initial procedure was necessary to ensure a fair comparison of the two methods. Large differences in dark diversity sizes between the two methods can lead to spurious interpretations. Therefore, data from the small-scale plots were used to estimate species pools following the methods detailed above, systematically substituting threshold values. Congruence between methods with differing thresholds was tested through a ratio of species pool size estimates from co-occurrence over that of the Ellenberg approach:

$$lr1 = \log \frac{\sum pool_{co-occurrence}}{\sum pool_{Ellenberg}}$$

Values closest to zero identified threshold values that result in the lowest dissimilarity between the two methods. We selected two threshold values with a median $lr1$ closest to zero. Each threshold represented a narrow filter (i.e. minimizing the species pool) and a broad filter (i.e. maximizing the species pool) and used to investigate the method performance.

Empirical dark diversity performance testing

To evaluate the accuracy of the two methods,, we validated the species pool estimates derived from modelling 465 species (SG) and 410 species (SF) recorded at the smaller scale plots, with species observed at the larger scale 200 m^2 (SG) and 500 m^2 (SF) plots. Species new to the data, that is species only observed at the larger scales, were removed from the analyses, as these cannot be estimated. Data were decomposed using the following notation: a = number of species observed at the larger scale and also predicted from modelling the smaller scale compositions, b= number of species observed at the larger-scale but not predicted (negative mismatch), while c = number of species predicted but not observed at the larger scale (positive mismatch). Expressed in terms of matching and mismatching components, we use ternary plots to examine the congruence between the two methods. Differences between 'a', 'b' and 'c' components derived from the two methods were statistically tested using paired t-tests. Where necessary, data were transformed to meet assumptions of normality and homogeneity of variance.

Because it is unlikely that all species belonging to a regional-scale habitat-specific species pool to be observed at the upper scales of the data sets (i.e. 200 m^2 and 500 m^2, for SG and SF, respectively), we extended the above analyses to include a further four scales from the centre of each focal assemblage (SG = 0·2, 0·5, 1 and 2 km radius and, SF = 4, 6, 8 and 10 km radius). These scales allowed us to evaluate the proportions of positive mismatches that would otherwise be observed if sampling effort were increased, that is species that belong to the dark diversity. This was achieved by compiling species lists of observed species in all larger scale plots within an increasing radius (km) of each

focal assemblage. Still, even here, large saturation of observed relative to the species pool is not expected, as even within these broader scales, we use a limited number of sample plots and not complete inventories. All statistical analyses were done in R version 3·0·2 (R Development Core Team 2014).

Results

ADJUSTING THE DARK DIVERSITY SIZE

As can be expected, size estimates of dark diversity were shown to be sensitive to the user defined threshold values. For SG, congruence between the two methods was closest where threshold values equalled 2 standard deviations of the mean

(Ellenberg) and a 1 percentile limit (co-occurrence); median lr1 = 0·05 (Fig. 3a & Table S1). While for SF similarity in pool size was high where threshold values equalled 1·5 standard deviations of the mean (Ellenberg) and a zero percentile (co-occurrence); median lr1 = −0·18 (Fig. 3b & Table S1). These threshold values determined the 'broad' filter, increasing the size of dark diversity and in turn the size of the species pool (SG *median pool size* = 63 & 60 and SF = 72 & 84 for the co-occurrence and Ellenberg methods, respectively). Threshold values 1·5 and 1 standard deviations of the mean (Ellenberg) and a 10th percentile limit (co-occurrence) delineated the 'narrow' filters for SG and SF, respectively, decreasing the size of dark diversity and in turn the size of the species pool (SG *median pool size* = 26 & 30 and SF = 26 & 29 for the co-occurrence and
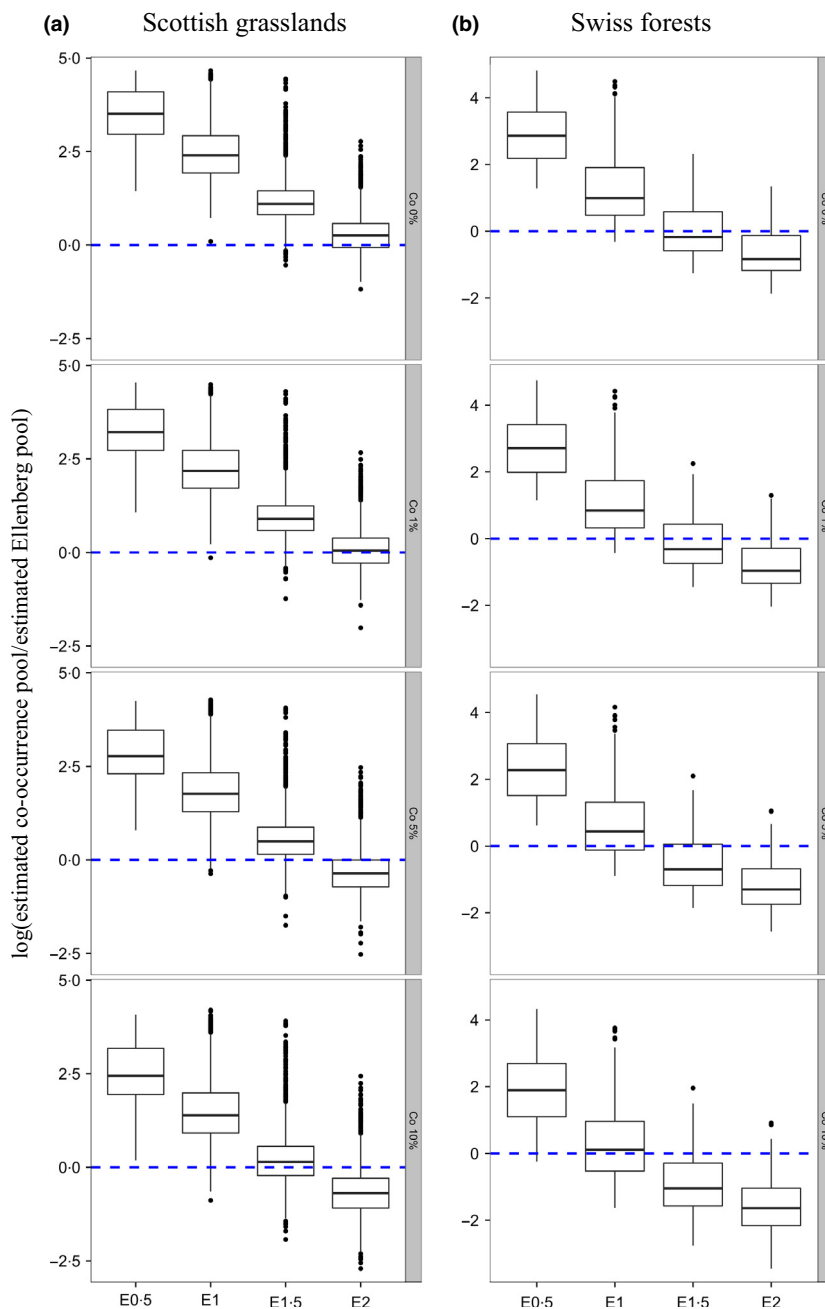
Fig. 3. Box plots displaying disparity in estimated species pool size between the both methods (co-occurrence & Ellenberg) for (a) Scottish grasslands and (b) Swiss forests. Presented are all combinations of tested thresholds between the co-occurrence (0, 1, 5, & 10 percentiles) and the Ellenberg (0·5, 1, 1·5 and 2 standard deviations of the mean.

Ellenberg methods, respectively). These restricted threshold values also resulted in close to zero lr1 values: SG lr1 = +0·14 and SF lr1 = +0·11 (Fig. 3; Table S1).

EMPIRICAL DARK DIVERSITY PERFORMANCE TESTING

The co-occurrence method outperformed the Ellenberg method irrespective of the data set used with significantly larger proportions of species observed and predicted resulting from the co-occurrence method, relative to the Ellenberg (Fig. 4, Table 1). Negative mismatches (i.e. observed but not predicted) were frequent among the Ellenberg method, while large proportions of positive mismatches (i.e. predicted but not observed) prevailed across both methods (Fig. 4). These patterns were statistically validated. Negative mismatches and indeed positive mismatches resulting from the Ellenberg

method were statistically and significantly more frequent relative to the co-occurrence across both data sets (Table 1). Restricting size estimates of dark diversity (i.e. narrow thresholds) resulted in greater proportions of negative mismatches for both methods (Fig. 4b, d, f, h). However, significantly fewer prevailed among estimates resulting from the co-occurrence method (Table 1). While neither method resulted in large proportions of component '*a*', that is where species predicted from smaller scale observations were in turn observed at the larger scale, increasing the observation scale did increase predictive success. Using only broad threshold values, SG predictive success increased from 18% at 200 m^2 to 67% (2 km scale) using the co-occurrence method and from 8% to 32% using the Ellenberg method (Fig. 5a). Similarly, SF predictive success increased from 18% at 500 m^2 to 60% (10 km scale) using the co-occurrence method and from 8% to 35% using the Ellenberg method (Fig. 5b).

Discussion

Dark diversity, that is species theoretically able to exist in a community but otherwise absent, is an important component of species pools. Unlike, its counter-part 'observed diversity', dark diversity cannot be directly measured, irrespective of spatial scale. This presents a challenge for deriving 'complete' estimates of a species pool, as it is vital the absent component be accurately accounted. Such difficulties may in part explain the relatively limited advance in the species pool concept since its formalization (Eriksson 1993; Pärtel *et al.* 1996). Most often, studies define the species pool subjectively or arbitrarily, using regional or local species inventories (Williams 1947; White & Hurlbert 2010), or by pooling local-scale observations (Belote, Sanders & Jones 2009; Chase *et al.* 2011; Kraft 2011). Such unstandardized approaches inherently suffer from bias (Hortal *et al.* 2008), fail to consider species ecological affinity to a focal assemblage and, importantly, ignore entirely dark diversity. Through the use of two independent and disparate data sets, the present study focused on two analytical approaches: Ecological requirements of species taken as Ellenberg indicator values, and species co-occurrence patterns, which both alleviate the difficulties in accurately constructing species pools. Both methods consider species ecological affinities and both account for dark diversity. However, our results suggest the two methods are not equal in terms of overall performance accuracy.

The results of this study suggest that the co-occurrence approach, relative to the Ellenberg method, to be more consistent in successfully predicting species observed with increasing sampling scale. The disparity between the two methods was especially clear when species observations were extended across larger Euclidean distances up to 2 and 10 km (depending on the analysed data set) from the focal assemblage, indicating a large proportion of positive mismatches through the co-occurrence approach accounted for local dark diversity and not an overestimation of the species pool. In fact, in this study, disparity between method performances cannot be linked to over- and/or underestimations of species pools since
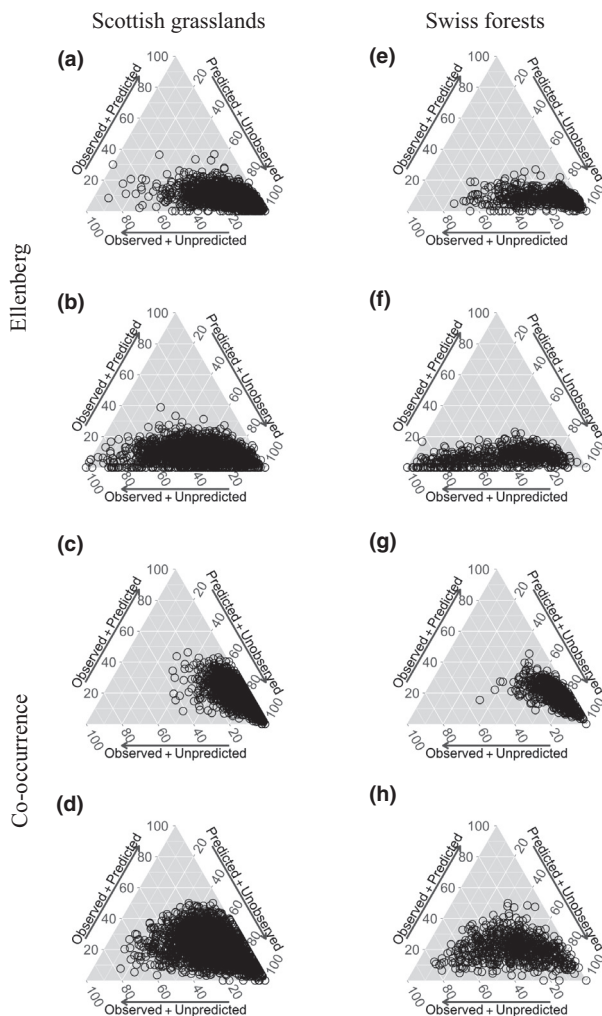

Fig. 4. Ternary diagrams illustrating proportions of observed & predicted, predicted and unobserved (positive mismatch) and observed and not predicted (negative mismatch) for species pool predictions from Scottish grasslands (a:d) and Swiss forests (e:h). Presented are results from two tested methods: Ellenberg (a, b, e, f) and Co-occurrence (c, d, g, h) representing threshold values maximizing the species pool (broad filter; a, c, e, g) and minimizing the species pool (narrow filter; b, d, f, h).

Table 1. Results of paired t-tests statistically comparing co-occurrence and Ellenberg methods in estimating dark diversity of Scottish grasslands (SG) and Swiss forest (SF). Presented are the mean differences (x diff; co-occurrence–Ellenberg), t and p statistics from comparisons of individual three part compositions where a = species observed & predicted, b = observed and not predicted (negative mismatch) and c = predicted and unobserved (positive mismatch), for both broad and narrow threshold filters

SG		Broad			Narrow		
	Component	x diff	t	*P*	x diff	t	*P*
Co-occurrence: Ellenberg	a	0·61	80·13	<0·001	0·79	76·67	<0·001
	b	−0·96	−86·44	<0·001	−0·58	−70·64	<0·001
	c	−0·03	−2·64	0·008	0·02	1·31	0·19
SF	Component	x diff	t	*P*	x diff	t	*P*
Co-occurrence: Ellenberg	a	0·61	27·91	<0·001	0·94	32·63	<0·001
	b	−0·66	−29·43	<0·001	−0·4	−31·17	<0·001
	c	−0·11	−4·18	<0·001	0·02	0·52	0·6

we determined and applied threshold values that result in convergent dark diversity sizes. The latter is an important point. Threshold values applied here were not designed to optimize overall estimation performance, as would be for a specific 'case study'. Using ecological requirements of species thus simply resulted in more instances where estimates failed to account for locally observed species relative to the co-occurrence approach.

The relatively better performance of co-occurrence estimates of dark diversity can be explained in part by the community composition data from which they are derived. As is well understood, compositional data typically display greater dissimilarity with increasing distance, that is they exhibit spatial autocorrelation. This autocorrelation in a broad-sense captures information on abiotic, biotic and disturbance factors; thus, the co-occurrence patterns themselves are the response of the prevailing ecological conditions and emphasizing ecological control (Whittaker 1956; Bray & Curtis 1957; Hutchinson 1957). This makes the co-occurrence approach exceptionally holistic in determining compatible absent species, that is dark diversity, principally because estimates are based on fairly accurate representations of a species realized niche. This however, is also a limiting factor. Species are not always in equilibrium with the abiotic environment (Svenning & Skov 2004), yet estimating species pools through species concurrences assumes that they are. Co-occurrence of 'leading' and 'trailing' edge populations (Svenning & Sandel 2013) may thus results in an overestimation of dark diversity, estimating species to be present which are otherwise environmentally non-adapt.

On the contrary, the relatively poorer performance of the Ellenberg approach might be explained through its estimates being based on broader representations of a species' realized niche. The Ellenberg method considers a species part of a community's dark diversity if an absent species shares similar environmental tolerances to those observed in a community. However, the ability to first establish in that community is not considered. Using co-occurrence patterns, the likelihood of a species to establish among a community given both abiotic and biotic constraints is inherently considered. However, species establishment probability derived from Ellenberg values

considers primarily the abiotic environment and even then limited to mean values of just a few specific ecological gradients. The approach thus disregards important mechanisms that govern species community assembly (e.g. biotic exclusion, propagule dispersal and population growth) and in turn fails to fully consider a vital pre-requisite when estimating species pools, that is to focus only on those species with a reasonable probability to disperse to the study site (Zobel, van der Maarel & Dupré 1998; Pärtel, Szava-Kovats & Zobel 2011, 2013).

One major challenge in quantifying dark diversity and species pools is how the relative scales (i.e. regional, local and community pools) are delineated. Here, the confines of our data defined our regional pool as all species with the ability to disperse to and establish in a given focal assemblage, but limited to the gamma-diversity of the entire survey area. Applying either method across large geographic scales can result in an overestimation of populations belonging to the dark diversity. For example, for species co-occurrences, generalist species that possess broad geographic ranges and which frequently occur can increase the probability that a species historically constrained to one region to be included into a communities' dark diversity in a region outside the species range. Although not strictly incorrect, it must first be asked whether realized or fundamental ranges are of interest as well as consideration of populations that may not necessarily be at equilibrium with the abiotic environment. Therefore, at broad scales prior to the application of co-occurrence or even ecological requirements of species to estimate dark diversity, it may be beneficial first to employ additional filters. This can be achieved through the application of concentric circles around a study site (*sensu* Graves & Gotelli 1983; geographic species pool), or by considering only species that adhere to similar historical and environmental conditions, that is the biogeographic species pool (Carstensen *et al.* 2013), see Ronk, Szava-Kovats & Pärtel (in press). This may even provide a possible explanation for the large proportion of predicted and unobserved species (i.e. positive mismatch) apparent in this regional-scale study, irrespective of the estimation method and vegetation type.

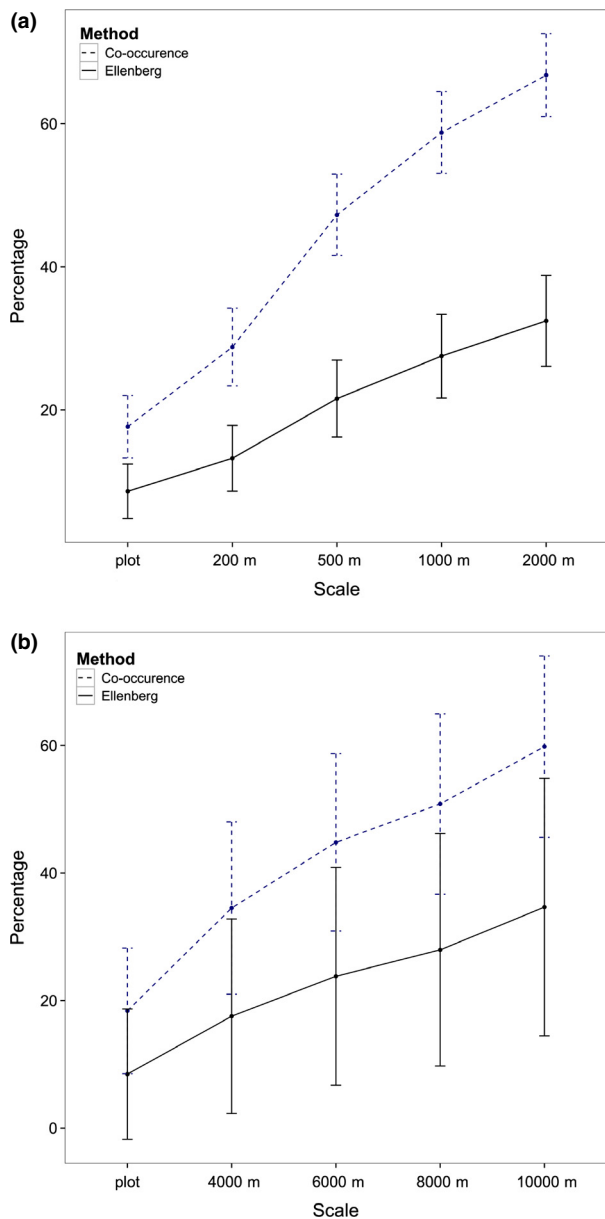For two distinctly separate vegetation types, we report significantly more negative mismatches (i.e. observed &

Fig. 5. Cumulative per cent of observed and predicted species (± standard error) with increasing distance from the focal plot (Scottish grasslands (a) 200–2000 m; Swiss forests (b) 4000–10 000 m). Observed species at these scales are compound observations across all 200 m^2 (Scottish grasslands) and 500 m^2 (Swiss forests) plots within the defined spatial extent.

unpredicted) resulting from the Ellenberg method than the co-occurrence method. Nevertheless, in some cases, the use of ecological requirements of species may still provide a suitable means of estimating absent species. For example, in this study, we focus on Ellenberg indicator values that are based on individual expert knowledge about species ecological requirements (Ellenberg *et al.* 1991). Although generally considered good estimates of real environmental conditions, some issues have undergone discussion as to whether Ellenberg indicator values are limited to certain parts of a given ecological gradient, vegetation type (Diekmann 2003) and region of origin (Godefroid

& Elías 2007), which together can limit the utility of using Ellenberg values for macro-ecological species pool estimates. Alternatively, use of Ellenberg might perform better where habitat heterogeneity in the data is high (e.g. along a successional gradient), providing broad ranges of Ellenberg values. In such cases, a variation of the approach used here has been shown to perform well working with nationwide habitat types (Pärtel *et al.* 1996).

Furthermore, although effective in this study, the success of co-occurrence patterns in detecting absent species is strongly dependent on large compositional data sets from which patterns can be analysed. The higher a species frequency the higher the likelihood of successful predictions (Münzbergová & Herben 2004), with 40 sampled units considered the minimum threshold in order to achieve good correlation between predicted and modelled species (De Cáceres & Legendre 2008). The use of this method can therefore be limited where data are sparse or where species incidence is low, that is where suites of species are temporally hidden (Pärtel 2014) or are simply rare. In fact, disentangling unsampled species with dark diversity is problematic from a single temporal observation, irrespective of the method. Nevertheless, techniques are available whereby sampling effort can be tested, for example species accumulation and rarefaction curves, Jackknife and Chao estimates (Gotelli & Colwell 2001; Cao *et al.* 2007). Estimating dark diversity for undersampled studies is therefore problematic for the co-occurrence approach, and in such instances, regional species lists and ecological requirements of species might prove a useful alternative.

Our results are clear in that the co-occurrence outperforms the Ellenberg method. Nevertheless, given our empirical approach, it remains difficult to make generalizations concerning how the two methods might respond to different ecological factors. For example, as already discussed, generalist species may result in overestimating dark diversity, but to what extent remains unanswered. Similarly, how the two methods might respond to rarity, diversity, homogeneity/ heterogeneity among the data, as well as variation in pool size, sample size, functional composition and dominant assembly processes cannot be inferred empirically. Explorative simulations would help in addressing these types of questions, continuing to increase our understanding of the types of ecological factors that enhance or limit successful dark diversity estimates.

Summary

Arguably, absent species can be as ecologically meaningful as present species, and in part, only tradition dictates focus on one rather than the other. Dark diversity has potential to unveil valuable information, revealing underlying community processes, and to enhance our understanding of the distribution and function of species (de Bello *et al.* 2012; Riibak *et al.* 2015). However, estimating dark diversity is challenging and not straightforward. This study brings new empirical evidence highlighting the efficacy of two methods applied towards accurately estimating dark diversity. We demonstrate through

empirical hierarchically sampled data sets that both species ecological requirements and species co-occurrence patterns provide an opportunity to examine the dark diversity of ecological systems. In particular, our results highlight that using species co-occurrence for estimating dark diversity and in turn species pool calculations, ultimately provides far greater accuracy than using species ecological requirements for macro-scale assessment of ecological communities. While further work is clearly required to fully understand responses of dark diversity estimates to various ecological structures, processes and patterns reflected in community data, this study does indeed demonstrate that accurate estimates of dark diversity are achievable.

Acknowledgements

Data accessibility

SG data used for these analyses are archived in the Dryad digital repository doi: 10.5061/dryad.gv506. SF data are freely available under the R package 'dave' (Wildi 2013).

References

Beals, E.W. (1984) Bray-Curtis ordination – an effective strategy for analysis of multivariate ecological data. *Advances in Ecological Research*, **14**, 1–55.

de Bello, F., Lavorel, S., Gerhold, P., Reier, Ü. & Pärtel, M. (2010) A biodiversity monitoring framework for practical conservation of grasslands and shrublands. *Biological Conservation*, **143**, 9–17.

de Bello, F., Price, J.N., Münkemüller, T., Liira, J., Zobel, M., Thuiller, W. *et al.* (2012) Functional species pool framework to test for biotic effects on community assembly. *Ecology*, **93**, 2263–2273.

Belote, R.T., Sanders, N.J. & Jones, R.H. (2009) Disturbance alters local-regional richness relationship in Appalachian forest. *Ecology*, **90**, 2940–2947.

Botta-Dukat, Z. (2012) Co-occurrence-based measure of species' habitat specialization: robust, unbiased estimation in saturated communities. *Journal of Vegetation Science*, **23**, 201–207.

Bray, R.J. & Curtis, J.T. (1957) An ordination of the upland forest communities of southern Wisconsin. *Ecological Monographs*, **27**, 325–349.

Cao, Y., Hawkins Charles, P., Larsen David, P. & Van Sickle, J. (2007) Effects of sample standardization on mean species detectabilities and estimates of relative differences in species richness among assemblages. *The American Naturalist*, **170**, 381–395.

Carstensen, D.W., Lessard, J.-P., Holt, B.G., Krabbe Borregaard, M. & Rahbek, C. (2013) Introducing the biogeographic species pool. *Ecography*, **36**, 1310–1318.

Chase, J.M., Kraft, N.J.B., Smith, K.G., Vellend, M. & Inouye, B.D. (2011) Using null models to disentangle variation in community dissimilarity from variation in α diversity. *Ecosphere*, **2**, art24.

Cornell, H.V. & Harrison, S.P. (2014) What are species pools and when are they important? *Annual Review of Ecology, Evolution, and Systematics*, **45**, 45–67.

De Cáceres, M. & Legendre, P. (2008) Beals smoothing revisited. *Oecologia*, **156**, 657–669.

Diamond, J.M. (1975) Assembly of species communities. *Ecology and Evolution of Communities* (eds M.L. Cody & J.M. Diamond), pp. 342–444. Harvard University Press, Cambridge, MA.

Diekmann, M. (2003) Species indicator values as an important tool in applied plant ecology – a review. *Basic and Applied Ecology*, **4**, 493–506.

Ellenberg, H., Weber, H.E., Dull, R., Wirth, V., Werner, W. & Paulissen, D. (1991) Zeigerwerte von Pflanzen in Mitteleuropa. *Scripta Geobot*, **18**, 1–248.

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

Ewald, J. (2002) A probabilistic approach to estimating species pools from large compositional matrices. *Journal of Vegetation Science*, **13**, 191–198.

Gilpin, M.E. & Diamond, J.M. (1982) Factors contributing to non-randomness in species co-occurrences on islands. *Oecologia*, **52**, 75–84.

Godefroid, S. & Elías, D.D. (2007) Can Ellenberg's indicator values for Mediterranean plants be used outside their region of definition? *Journal of Biogeography*, **34**, 62–68.

Gotelli, N.J. & Colwell, R.K. (2001) Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecology Letters*, **4**, 379–391.

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

Hill, M.O., Mountford, J.O., Roy, D.B. & Bunce, R.G.H. (1999) Ellenberg indicator values for British plants, ECOFACT, vol 2, technical annex. ITE Monkswood, Huntingdon, Department of the Environment, Transport and the Regions, London, UK.

Hortal, J., Jiménez-Valverde, A., Gómez, J.F., Lobo, J.M. & Baselga, A. (2008) Historical bias in biodiversity inventories affects the observed realized niche of the species. *Oikos*, **117**, 847–858.

Hutchinson, G.E. (1957) Concluding remarks. *Cold Spring Harbor Symposia on Quantitative Biology*, **22**, 15–427.

Klimesova, J. & Klimes, L. (2007) Bud banks and their role in vegetative regeneration – a literature review and proposal for simple classification and assessment. *Perspectives in Plant Ecology, Evolution and Systematics*, **8**, 115–129.

Kraft, N.J. (2011) Disentangling the drivers of β diversity along latitudinal and elevational gradients. *Science*, **333**, 1755–1758.

Lessard, J.P., Belmaker, J., Myers, J.A., Chase, J.M. & Rahbek, C. (2012) Inferring local ecological processes amid species pool influences. *Trends in Ecology and Evolution*, **27**, 600–607.

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

Pärtel, M. (2014) Community ecology of absent species: hidden and dark diversity (R. Kalamees, Ed.). *Journal of Vegetation Science*, **25**, 1154–1159.

Pärtel, M., Szava-Kovats, R. & Zobel, M. (2011) Dark diversity: shedding light on absent species. *Trends in Ecology and Evolution*, **26**, 124–128.

Pärtel, M., Szava-Kovats, R. & Zobel, M. (2013) Community completeness: linking local and dark diversity within the species pool concept. *Folia Geobotanica*, **48**, 307–317.

Pärtel, M., Zobel, M., Zobel, K. & van der Maarel, E. (1996) The species pool and its relation to species richness: evidence from Estonian plant communities. *Oikos*, **75**, 111–117.

R Development Core Team (2014) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria http://www.R-project.org/.

Riibak, K., Reitalu, T., Tamme, R., Helm, A., Gerhold, P., Znamenskiy, S. *et al.* (2015) Dark diversity in dry calcareous grasslands is determined by dispersal ability and stress-tolerance. *Ecography*, **38**, 713–721.

Ronk, A., Szava-Kovats, R. & Pärtel, M. (in press) Applying the dark diversity concept to plants at the European scale. *Ecography*, DOI: 10.1111/ecog.01236.

Shaw, M.W., Hewett, D.G. & Pizzey, J.M. (1983) *Scottish Coastal Survey*. Institute of Terrestrial Ecology. Bangor Research Station, Bangor, Gwynedd.

Silvertown, J., Dodd, M., Gowing, D., Lawson, C. & McConway, K. (2006) Phylogeny and the hierarchical organization of plant diversity. *Ecology*, **87**, S39–S49.

Suding, K.N. (2011) Toward an era of restoration in ecology: successes, failures, and opportunities ahead. *Annual Review of Ecology, Evolution, and Systematics*, **42**, 465–487.

Svenning, J.-C. & Sandel, B. (2013) Disequilibrium vegetation dynamics under future climate change. *American Journal of Botany*, **100**, 1266–1286.

Svenning, J.-C. & Skov, F. (2004) Limited filling of the potential range in European tree species. *Ecology Letters*, **7**, 565–573.

Veech, J.A. (2013) A probabilistic model for analysing species co-occurrence. *Global Ecology and Biogeography*, **22**, 252–260.

Vellend, M. (2010) Conceptual synthesis in community ecology. *The Quarterly Review of Biology*, **85**, 183–206.

White, E.P. & Hurlbert, A.H. (2010) The combined influence of the local environment and regional enrichment on bird species richness. *American Naturalist*, **175**, 35–43.

Whittaker, R.H. (1956) Vegetation of the Great Smoky Mountains. *Ecological Monographs*, **26**, 1–80.

Wildi, O. (2013) *Data Analysis in Vegetation Ecology*, 2nd edn. Wiley-Blackwell, Chichester.

Williams, C.B. (1947) The generic relations of species in small ecological communities. *Journal of Animal Ecology*, **16**, 11–18.

Wohlgemuth, T., Moser, B., Braendli, U.-B., Kull, P. & Schuetz, M. (2008) Diversity of forest plant species at the community and landscape scales in Switzerland. *Plant Biosystems*, **142**, 604–613.

Zobel, M. (1997) The relative role of species pools in determining plant species richness. An alternative explanation of species coexistence? *Trends in Ecology and Evolution*, **12**, 266–269.

Zobel, M., van der Maarel, E. & Dupré, C. (1998) Species pool: the concept, its determination and significance for community restoration. *Applied Vegetation Science*, **1**, 55–66.

Supporting Information

Additional Supporting Information may be found in the online version of this article.

Appendix S1. Dissimilarity in dark diversity size estimates.

Appendix S2. R code for computing dark diversity. (Correction added on 12 November 2015, after first online publication: A legend has been removed as it was duplicated. Some minor amendments were made to Appendix 2).