MRS. MARIA SPORBERT (Orcid ID : 0000-0001-7994-8491)

DR. HELGE BRUELHEIDE (Orcid ID : 0000-0003-3135-0356)

DR. ANDRAŽ ČARNI (Orcid ID : 0000-0002-8909-4298)

DR. MILAN CHYTRÝ (Orcid ID : 0000-0002-8122-3075)

CORRADO MARCENÒ (Orcid ID : 0000-0003-4361-5200)

DR. VIGDIS VANDVIK (Orcid ID : 0000-0003-4651-4798)

Article type : Research article

Duccio Rocchini

Co-ordinating Editor: Duccio Rocchini

Assessing sampling coverage of species distribution in biodiversity databases

Running title: Sampling coverage by box-counting

Maria Sporbert [1,2 *], Helge Bruelheide [1,2], Gunnar Seidler [1], Petr Keil [1,3], Ute Jandt [1,2], Gunnar Austrheim [4], Idoia Biurrun [5], Juan Antonio Campos [5], Andraž Čarni [6,7], Milan Chytrý [8], János Csiky [9], Els De Bie [10], Jürgen Dengler [2,11,12], Valentin Golub [13], John-Arvid Grytnes [14], Adrian Indreica [15], Florian Jansen [16], Martin Jiroušek [8,17], Jonathan Lenoir [18], Miska Luoto [19], Corrado Marcenò [5], Jesper Erenskjold Moeslund [20], Aaron Pérez-Haase [21], Solvita Rūsiņa [22], Vigdis Vandvik [23,24], Kiril Vassilev [25], Erik Welk [1,2]

[1]Institute of Biology / Geobotany and Botanical Garden, Martin Luther University Halle-Wittenberg, Halle, Germany

[2]German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Leipzig, Germany

[3]Institute of Computer Science / Biodiversity Synthesis, Martin Luther University Halle-Wittenberg, Halle, Germany

[4]Department of Natural History, University Museum Norwegian University of Science and Technology, Trondheim, Norway

[5]Department Plant Biology and Ecology, University of the Basque Country UPV/EHU, Bilbao, Spain

[6]Scientific Research Centre of the Slovenian Academy of Sciences and Arts, Jovan Hadži Institute of Biology, Ljubljana, Slovenia
[7]School for Viticulture and Enology, University of Nova Gorica, Nova Gorica, Slovenia
[8]Department of Botany and Zoology, Faculty of Science, Masaryk University, Brno, Czech Republic
[9]Institute of Biology / Ecology, University of Pécs, Hungary
[10]Research Institute for Nature and Forest, Biotope Diversity, Brussels, Belgium
[11]Vegetation Ecology Group, Institute of Natural Resource Management (IUNR), Zurich University of Applied Sciences (ZHAW), Wädenswil, Switzerland
[12]Plant Ecology, Bayreuth Center of Ecology and Environmental Research (BayCEER), University of Bayreuth, Bayreuth, Germany
[13]Institute of Ecology of the Volga River Basin, Russian Academy of Sciences, Togliatti, Russia
[14]Department of Biological Sciences, University of Bergen, Bergen, Norway
[15]Department of Silviculture, Transilvania University of Brasov, Brasov, Romania
[16]Faculty of Agricultural and Environmental Sciences, University of Rostock, Germany
[17]Department of Plant Biology, Faculty of AgriSciences, Mendel University, Brno, Czech Republic
[18]UR "Ecologie et Dynamique des Systèmes Anthropisés" (EDYSAN, UMR 7058 CNRS-UPJV), Université de Picardie Jules Verne, Amiens, France
[19]Department of Geosciences and Geography, University of Helsinki, Helsinki, Finland
[20]Department of Bioscience - Biodiversity and Conservation, Aarhus University, Rønde, Denmark
[21]Department of Evolutionary Biology, Ecology and Environmental Sciences, University of Barcelona, Barcelona, Spain
[22]Faculty of Geography and Earth Sciences, University of Latvia, Riga, Latvia
[23]Department of Biological Sciences, University of Bergen, Bergen, Norway
[24]Bjerknes Centre for Climate Research, University of Bergen, Bergen, Norway
[25]Institute of Biodiversity and Ecosystem Research / Plant and Fungal Diversity and Resources, Bulgarian Academy of Sciences, Department of, Sofia, Bulgaria

* Corresponding author: tel +49 345 55 26287; maria.sporbert@botanik.uni-halle.de

Funding information: Graduiertenförderung Sachsen-Anhalt (scholarship to MS), with additional support through institutional funds of Martin Luther University Halle-Wittenberg, Czech Science Foundation (project no. 19-28491X to MC)

Abstract

Aim: Biodiversity databases are valuable resources for understanding plant species distributions and dynamics, but they may insufficiently represent the actual geographic distribution and climatic niches of species. Here we propose and test a method to assess sampling coverage of species distribution in biodiversity databases in geographic and climatic space.

Location: Europe.

Methods: Using a test selection of 808,794 vegetation plots from the European Vegetation Archive (EVA), we assessed the sampling coverage of 564 European vascular plant species across both their geographic ranges and realized climatic niches. Range maps from the Chorological Database Halle (CDH) were used as background reference data to capture species geographic ranges and to derive species climatic niches. To quantify sampling coverage, we developed a box-counting method, the Dynamic Match Coefficient (DMC), which quantifies how much a set of occurrences of a given species matches with its geographic range or climatic niche. DMC is the area under the curve measuring the match between occurrence data and background reference (geographic range or climatic niche) across grids with variable resolution. High DMC values indicate good sampling coverage. We applied null models to compare observed DMC values with expectations from random distributions across species ranges and niches.

Results: Comparisons with null models showed that, for most species, actual distributions within EVA are deviating from null model expectations and are more clumped than expected in both geographic and climatic space. Despite high interspecific variation, we found a positive relationship in DMC values between geographic and climatic space, but sampling coverage was in general more random across geographic space.

Conclusion: Because DMC values are species-specific and most biodiversity databases are clearly biased in terms of sampling coverage of species occurrences, we recommend using DMC values as covariates in macroecological models that use species as the observation unit.

Keywords: Chorological Database Halle (CDH), climatic niche, Dynamic Match Coefficient (DMC), European Vegetation Archive (EVA), macroecology, multi-scale, realized niche, sampling bias, spatial scale, species range, vascular plant, vegetation-plot databases.

1 Introduction

Large-scale biodiversity databases (e.g. Global Biodiversity Information Facility (GBIF), Edwards, Lane, & Nielsen, 2000; Botany Information and Ecology Network (BIEN), Enquist, Condit, Peet, Schildhauer, & Thiers, 2009; sPlot, Bruelheide et al., 2019) are valuable resources for understanding species distributions and dynamics. Possible applications include broad-scale analyses across species or community types (e.g. Bruelheide et al., 2018; Jiménez-Alfaro et al., 2018), species distribution

models (SDM) (Gomes et al., 2018; Wasof et al., 2015); and monitoring biodiversity changes over time (Bertrand et al., 2011; Jandt, von Wehrden, & Bruelheide, 2011). For broad-scale analyses covering the entire range of species, the quality of the sampling coverage across a given species range or throughout its realized niche is crucial. Hence, consistent data distribution is highly desirable across both the geographic and environmental space (Broennimann & Guisan, 2008; Pearman, Guisan, Broenniman, & Randin, 2008; Troia & McManamay, 2016). However, biodiversity databases often suffer from sampling gaps and biases limiting their application potential. Because of the uneven collection effort (Daru et al., 2018; Soria-Auza & Kessler, 2007; Speed et al., 2018) often caused by difficult access to some areas (Sousa-Baena, Garcia, & Peterson, 2014), broad regions of the world remain poorly sampled. Even comprehensive databases of species occurrences in well-surveyed regions are prone to geographic (Yang, Ma, & Kreft, 2013) and taxonomic biases (Pyke & Ehrlich, 2010; Soberón, Jiménez, Golubov, & Koleff, 2007). In an in-depth evaluation, Meyer, Weigelt, & Kreft (2016) found severe geographical bias in the GBIF database (Edwards et al., 2000), concluding that data limitations are rather the rule than the exception for most species and regions.

Species distribution models (SDM) are commonly used for macroecological niche analyses. They represent the estimation of species occurrence probabilities based on observed geographic distributions. Thereby, SDMs are sensitive to poor sampling coverage, especially if spatial bias results in climatically biased sampling (Fourcade, Engler, Rödder, & Secondi, 2014). In such situations, SDMs tend to misestimate species climatic niches (Titeux et al., 2017). Thus, for reliable analyses of biodiversity distribution patterns, sampling coverage needs to be representative for both the climatic and geographic space (Hortal, Jiménez-Valverde, Gómez, Lobo, & Baselga, 2008; Troia & McManamay, 2016). Unbiased sampling is typically obtained by meeting two interrelated requirements: sufficient sample size and even coverage of geographical and environmental gradients. Towards coarser spatial resolution, good coverage is easier to achieve and, as a consequence, sampling bias typically decreases. Consequently, the negative impact of sampling bias is clearly related to spatial grain. Several studies have analyzed the importance of spatial scaling in niche studies (e.g. Pearman et al., 2008; Soberón et al., 2007; Hortal, Borges, & Gaspar, 2006). Recently, procedures have been developed to assess the completeness of a spatial dataset at different spatial resolutions in geographic space (*KnowBR,* Lobo et al., 2018; *downscale*, Marsh, Barwell, Gavish, & Kunin, 2018). At large spatial extent, climate is among the most important factors determining species distributions (Woodward, 1986). However, although including climate seems straightforward, until now, few studies have accounted for how evenly occurrence data cover species ranges in climatic space (e.g. Bruelheide et al., 2018). To our knowledge, no study has explicitly tested the degree to which the spatial distribution of occurrences represents the geographical range as well as the climatic niche of the sampled species.

Here we test the spatial and climatic coverage of plant occurrence data using an example dataset of the European Vegetation Archive (EVA). EVA is a key macroecological resource that incorporates information from 57 countries on approximately 1.5 million vegetation plots containing more than 10,000 vascular plant species (Chytrý et al., 2016). EVA data are used for various research objectives, yet the degree of unevenness in sampling effort across Europe's geographic and environmental space is unclear. A species distribution database covering EVA's spatial extent, but otherwise independent from EVA, is the Chorological Database Halle (CDH) (Welk et al., unpubl.). CDH stores georeferenced information (range polygons and point occurrences) on the distribution range of more than 1,200 European vascular plant species. Species distribution data from CDH have already been used in several biodiversity studies (e.g. Csergő et al., 2017; San-Miguel-Ayanz, de Rigo, Caudullo, Houston Durrant, & Mauri, 2016; Schleuning et al., 2016) and as basis for biogeographical experiments on plant range limits (Bütof et al., 2012; Hofmann, Bütof, Welk, & Bruelheide, 2013; Welk, Welk, & Bruelheide, 2014). Here, we made use of expert-based range maps stored in CDH to extract information on both species geographic ranges and climatic niches and assess the sampling coverage of species occurrences stored in EVA across each of these two backgrounds (geographic and climatic).

To quantify sampling coverage, we developed the Dynamic Match Coefficient (DMC), a measure based on the area-under-the-curve (AUC) derived from threshold-independent box-counting statistics across variable spatial grains. We compared the observed DMC values with the values of plots randomly distributed across the species range and niche. Thereby, we produced an expected null reference distribution (Nunes & Pearson, 2017) within both the geographic and climatic space for a given sampling effort (sample size) and corresponding to the observed species frequency in the database. This enabled us to evaluate the observed plot distribution in geographic space (DMC_{GEO}) and climatic space (DMC_{CLIM}) in comparison to expectations of randomly distributed plots across the species range and realized climatic niche. We tested four hypotheses on sampling coverage of species occurrences across both the geographic and climatic space:

(1) Sampling coverage within the climatic space depends strongly on good sampling coverage across the geographic space because climatic conditions are spatially autocorrelated. We expect a positive correlation between sampling coverage in the geographic and climatic space.

(2) Sampling coverage is less representative in the climatic space than in the geographic space. The reason is the asymmetric transferability between points in the climatic and geographic space: a single point within the climatic space might translate to several geographic locations, while a single geographic location can only translate to one point in the climatic space. An increase in sampling coverage within the geographic space might thus be without positive effect on sampling coverage within the climatic space.

(3) Given the general sampling issues of biodiversity databases mentioned above and the heterogeneous nature of their source data, we expect that sampling coverage of the realized niches of plant species by such data is largely imperfect because of an underdispersed (clumped) distribution of species observations within the geographic space and supposedly also within the climatic space.
(4) Finally, for a given range size and macroclimatic niche size, we expect sampling coverage to increase with increasing sample size.

2 Material and Methods

We assessed the sampling coverage of European vascular plant species ranges (using species range data from the Chorological Database Halle, CDH) by a test selection of species occurrence data taken from vegetation plots from the European Vegetation Archive (EVA, Chytrý et al., 2016). We did this both in the geographic space (distribution range data from CDH) and in the climatic space (realized climatic niche space derived from CDH geographical distributions). We focused on species presence data (i.e. locations of vegetation plots in which the focal species was recorded) and examined the relationship between the geographic and climatic sampling coverage, as well as interspecific variability. The study area comprised all European countries plus Turkey, Georgia, Armenia and Azerbaijan (Figure 1a).

2.1 Background data on species geographic range and climatic niche

The Chorological Database Halle (CDH) stores information on distribution ranges of about 17,000 vascular plant taxa. For 5,583 taxa, maps were compiled based on published distribution range maps (Meusel, Jäger, & Weinert, 1965; Meusel, Jäger, Rauschert, & Weinert, 1978; Meusel & Jäger, 1992), national and floristic databases and further maps from floristic literature (see bibliographic details in Index Holmiensis, Tralau, 1969-1981; Lundqvist & Nordenstam, 1988; Lundqvist, 1992; Lundqvist & Jäger, 1995-2007). CDH data can be requested for research objectives via http://chorologie.biologie.uni-halle.de/choro/. We retrieved from CDH the available geographical information for the distribution ranges of 1,200 European vascular plant species in electronic format (range polygons and point occurrences) in October 2015. The species range information was processed as raster layers of 2.5-min cell resolution, which is about 15 km² in Central Europe (Figure 1a). The multi-dimensional climatic space (climatic niche) was determined by principal components analysis (PCA) of 19 bioclimatic variables from Worldclim with 2.5-min cell resolution (Hijmans, Cameron, Parra, Jones, & Jarvis, 2005) (for detailed information see Appendix S1 in the Supporting Information).

2.2 Vegetation plots

A test selection of vegetation plots was provided by the European Vegetation Archive in October 2015, containing information on 10,082 species from 933,228 vegetation plots. This selection included all the plots that were available in EVA at that time. Data for intraspecific taxa such as subspecies were merged at the species level. Further, we matched species names and checked for synonyms according to (i) the taxonomic reference list for Germany (German SL version 1.2, Jansen & Dengler, 2008) and (ii) all taxonomic reference lists available via the R package 'taxize' (Chamberlain & Szöcs, 2013; Chamberlain et al., 2018). We excluded trees, bryophytes, lichens, fungi, algae and species exotic to Europe. We also excluded 67,200 vegetation plots with location uncertainty larger than 10 km and 417 species that occurred in less than 10 plots.

After matching EVA and CDH species, 808,794 vegetation plots contained at least one of the 564 vascular plant species (herbs, dwarf shrubs and shrubs) with available digitized geographic distribution data in CDH. A list of these species and all the databases that provided vegetation plot data can be found in Appendices S2 and S3 in the Supporting Information. The 808,794 vegetation plots from EVA were heterogeneously distributed across the study area in the geographic space. While some geographic regions were represented very well and with high density (e.g. the Czech Republic, the Netherlands), other regions were represented sparsely (e.g. Norway, Sweden, Finland, Belarus, parts of Russia; Figure 1a). In contrast to geographic space, the study area was well represented by EVA vegetation plots in climatic space, except some marginal parts of the climatic background space (Figure 1b). The maximum density of species was 396 species per 2.5 min raster cell in geographic space (Figure 2a) and 528 species per cell in climatic space (Figure 2b). Stacked CDH ranges of the 564 study species covered 98.5% of the study area in geographic space (154,455 raster cells of 2.5-min in total) (Figure 2a) and 100% in climatic space (9,931 cells in total; Figure 2b).

2.3 Dynamic Match Coefficient (DMC) - a measure of plot sampling coverage across spatial scales

Sampling bias is mainly a result of two interrelated issues: insufficient number of samples and inadequate sample distribution. The impact of sampling bias is related to spatial scale (spatial extent and grain size) and should decrease with increasing grain size. The spatial arrangement of sampling locations could be evaluated by classical methods of point pattern analysis (Boots & Getis, 1988; Wiegand & Moloney, 2013). However, there are two main issues related to the spatial pattern in the ecological domain of the data of interest. First, because of the generally irregular, often non-contiguous geometry of plant distribution ranges, traditional Euclidean geometry often fails to estimate characteristics of point patterns correctly (Pentland, 1984). Second, species ranges and niches cannot be regarded as merely geometric phenomena. Spatio-temporal population processes often

result in complex range structures of genetic diversity, demographic performance and abundance (Peterson et al., 2011; Ricklefs, 2004).

To measure how well, i.e. how uniform vs. clustered and simultaneously how dense or scarce vegetation plots containing the focal species are located across the species` range or niche, we developed a measure inspired by fractal dimension analysis (Hall & Wood, 1993), which we call the Dynamic Match Coefficient (DMC). The DMC represents a measure of cell matches between a point pattern and spatial layers that are iterated across different raster cell resolutions (grain sizes), from fine to coarse (Figure 3). Here, 20 iterative scaling steps were used, which resulted in a maximum achievable DMC of 2000 (20 × 100% match). The obtained values were standardized to 0-1. For all species, the starting grain size in geographic space was $1/20^{th}$ of the respective species maximum North-South and East-West range extent. Hence, the initial grain size was smaller for small-range species (e.g. 50 km × 20 km for *Centaurea deustiformis*) than for large-range species (e.g. 211 km × 273 km for *Plantago major*) (see Appendices S2 and S4.1 in the Supporting Information for distribution of initial grain sizes in DMC calculations). Among the chosen starting grain sizes for the geographic space, even the finest grid cells (50 km × 20 km) are at a spatial resolution where climate conditions are considered the most important (Pearson & Dawson, 2003). The scaling procedure used in the climatic space was similar to that in the geographic space. Here the initial grain size was derived as the $1/20^{th}$ fraction of the respective species maximum niche extent along the first two PCA axes. High DMC values indicate high sampling coverage, i.e. a more regular distribution and density of EVA vegetation plots across a species distribution range or within its realized climatic niche. In contrast, low DMC values indicate underdispersed sampling coverage, i.e. clumped distribution and/or inappropriately low density of EVA vegetation plots across a species distribution range or within its realized climatic niche (Figure 3).

Figure 4 shows how the DMC approach works for the geographic and climatic space and for two contrasting species: *Hieracium murorum*, a species with clumped distribution in EVA plots, and *Calluna vulgaris*, a species with a more regular distribution in EVA plots, both in the species range and in the realized climatic niche (Figure 4a). Range size and the number of vegetation plots are similar in both species. The cell match ratio between species range and EVA vegetation plots was calculated in 20 iterations from fine to coarse raster cell resolution for both species in the geographic and climatic space (Figure 4b). The cell match ratio at the 20 single raster steps was summed up, and this sum is what we term the final DMC value of a species in the geographic space (DMC_{GEO}) and climatic space (DMC_{CLIM}). For *Hieracium murorum*, DMC values reached 0.42 and 0.58 for the geographic (DMC_{GEO}) and climatic (DMC_{CLIM}) space, respectively. For *Calluna vulgaris*, DMC values reached 0.74 for both the geographic (DMC_{GEO}) and climatic (DMC_{CLIM}) space.

2.4 Observed vs. expected distributions

In order to quantify how far the observed DMC deviates from an expected random distribution, we applied a null model simulation (Nunes & Pearson, 2017) for each species. We randomly distributed a number of species occurrences for each species (n = number of plots containing the species) across its geographic range and climatic niche. We calculated the DMC_{GEO} and DMC_{CLIM} values for 100 such random distributions in the geographic and climatic space, respectively, and compared the simulated DMC distribution with the observed value. To quantify the deviation of the observed DMC value from the median of the simulated ideal random distribution (DMC_{NULL}) we calculated a DMC ratio as:

$$DMC\ ratio\ =\ \frac{(DMC\ NULL - DMC\ observed)}{DMC\ observed}$$

A high DMC ratio corresponds to an underdispersed distribution of the EVA plots containing the species, while a low DMC ratio corresponds to a more random distribution. A negative ratio corresponds to an overdispersed distribution.

2.5 Effect of sample size on the DMC value

We analysed the effect of sample size (number of EVA plots containing a given species) on DMC values while accounting for range size (or niche size) by applying linear models with DMC_{GEO} (or DMC_{CLIM}) values as the response variable, sample size as the main explanatory variable and range size (resp. niche size) as a covariate to correct for potential confounding effects of range size or niche size. In a first step, for each species, the percentage match of the species range (derived from CDH) by the respective EVA vegetation plots where the species occurred was calculated at 2.5-min raster cell resolution. Multiple occurrences per raster cell were reduced to presence-absence data per species and 2.5-min raster cell. In the second step, species ranges and the respective vegetation plots were projected into the climatic space. The study area in the climatic space is well represented by its first two PCA axes, which explain 88.0% of the data variance (for details see Appendix S1 in Supporting Information). Finally, the percentage of a species climatic niche matched by vegetation plots where the species occurred was calculated as the ratio of PCA cells of the respective EVA vegetation plots where the species occurred to all raster cells matched by the species range in the PCA space (species percentage match of its range and niche by EVA vegetation plots is provided in Appendix S2 in the Supporting Information).

3 Results

Overall, sampling coverage of European vascular plant species ranges by EVA vegetation plots was more complete within the geographic space than within the climatic space (Figure 5), i.e. consistently higher DMC values were within the geographic space (DMC_{GEO}). The mean of DMC_{GEO} was slightly

higher than that of DMC_{CLIM}, with values of 0.56 and 0.49, respectively. Species DMC_{GEO} values ranged from 0.08 to 0.94. For half of the species the DMC_{GEO} was between 0.48 and 0.65 (25^{th} and 75^{th} percentile). DMC_{CLIM} values ranged from 0.08 to 0.82 and for half of the species the DMC_{CLIM} was between 0.40 and 0.60 (25^{th} and 75^{th} percentile). We found a highly significant positive correlation (Spearman´s rho = 0.768; $p < 0.001$) between species geographic DMC values (DMC_{GEO}) and their climatic DMC values (DMC_{CLIM}) (Figure 5). DMC_{CLIM} values were higher than DMC_{GEO} values for only 119 species (21.1%), while 445 species (78.9%) had higher DMC_{GEO} values than DMC_{CLIM} values. Furthermore, some species showed a high deviation in DMC values between the geographic and climatic space. For instance, *Arabis alpina* was more randomly sampled within the climatic space (DMC_{CLIM}: 0.55) than within the geographic space (DMC_{GEO}: 0.24), while this was the opposite for *Vinca major* (DMC_{GEO}: 0.63, DMC_{CLIM}: 0.29). In general a positive relationship between species range size and niche size could be observed (Spearman´s rho = 0.805; $p <0.001$; Appendix S4.2 in Supporting Information).

3.1 Deviation of the observed DMC from the expected random distribution

We found a positive correlation between the observed DMC values and the expected DMC values, based on our null model, for both the geographic space (weaker, Spearman´s rho = 0.389; $p <0.001$) and the climatic space (stronger, Spearman´s rho = 0.824; $p <0.001$) (Figures 6a and 6b). Importantly, a large majority (92.0%) of the observed species distributions in EVA were significantly underdispersed in both the geographic and climatic space. This is indicated by the position of most of the points above the 1:1 line, especially in the climatic space. Exceptionally, for a small number of species in the geographic space (43 species, 7.6%) (Figure 6a) and for two species in the climatic space (Figure 6b), the observed DMC values were higher than the null random expectation, indicating overdispersion.

For each species, we calculated the deviation of the observed DMC values from the null model DMC values in geographic and climatic space. While a low deviation of the observed DMC values from the null expectation indicates a more regular distribution of occurrences for a given species across its reference range or realized climatic niche, a high deviation indicates an underdispersed (more clumped) distribution. We found a positive correlation for the deviation of observed DMC values from the null model DMC values between geographic and climatic space (Spearman´s rho = 0.615; $p <0.001$). Despite a higher variability, DMC deviation from the null model was on average slightly lower in geographic space (min_{DEV_GEO}: -0.31, max_{DEV_GEO}: 2.47, $median_{DEV\_GEO}$: 0.46) than in climatic space (min_{DEV_CLIM}:-0.10, max_{DEV_CLIM}:2.09, $median_{DEV\_CLIM}$:0.47, see Figure 7).

3.2 Effect of sample size on DMC values

In geographic space, the percentage match of species ranges by EVA vegetation plots containing the same species (measured as the percentage of the range containing the EVA plots at 2.5-min raster cell resolution) ranged from 0.01% to 67.6%. For half of the species, the percentage match was between 0.5% and 2.3% (25^{th} and 75^{th} percentile), with a mean of 1.1% in the geographic space. In the climatic space, the percentage match of species niches by EVA vegetation plots ranged from 0.5% to 72.7% and for half of the species the percentage match was between 7.6% and 22.1% (25^{th} and 75^{th} percentile), with a mean of 14.1%. The applied linear models revealed a positive effect of sample size (vegetation plots) on DMC values while accounting for range size or niche size in both the geographic space (multiple R^2: 0.212) and climatic space (multiple R^2: 0.571). We found a significantly positive correlation between the percentage match of the species range by EVA plots in both the geographic space (Spearman´s rho= 0.726; p <0.001) and climatic space (Spearman´s rho= 0.901; p <0.001) (Figure 8a and b). Furthermore, we encountered a significantly negative relationship between percentage match of species ranges by EVA vegetation plots and deviation from the null model in the geographic space (Spearman´s rho= -0.601; p <0.001) and climatic space (Spearman´s rho= -0.651; p <0.001) (Figure 8c and d). Apart from this, a significantly positive correlation between the percentage match of the species range by EVA plots in the geographic space and climatic space could be found (Spearman´s rho= 0.865; p <0.001; Appendix S4.3 in Supporting Information).

4 Discussion

4.1 Plot sampling coverage across spatial scales

In line with the general positive relationship between range size and niche size (see Appendix S4.2 in Supporting Information), we assumed that (1) a species will be well sampled throughout its multidimensional climatic niche (reaching high DMC_{CLIM} values) only if it is well sampled throughout its geographic range (high DMC_{GEO} values). The demonstrated positive correlation between DMC_{CLIM} and DMC_{GEO} confirms the first hypothesis. However, the relationship was far from perfect, since there are also species that are well sampled within the geographic space (reaching high DMC_{GEO} values) but less well sampled in the climatic space (reaching low DMC_{CLIM} values), and vice versa. Exceptions from the suggested positive relationship can arise especially due to high spatial heterogeneity in climatic conditions, e.g. in mountain regions (Hirst, Griffin, Sexton, & Hoffmann, 2017; Köckemann, Buschmann, & Leuschner, 2009).

Because of the one-to-n relationship between climatic and geographic data points we expected (2) a sparser species sample coverage (lower DMC values) in the climatic space. Accordingly, we found that the sampling coverage (DMC value) of species distribution in EVA was more random in the geographic space (DMC_{GEO}) than in the climatic space (DMC_{CLIM}) for 77.9% of the studied species. This more random sampling coverage in geographic space is explainable by the niche–biotope duality

(Hutchinson, 1978). The same combination of climate factors can occur in only one location in geographic space, but will more likely occur in several localities with increasing spatial extent (Colwell & Rangel, 2009; Soberón & Nakamura, 2009). However, the rules that define the niche–biotope duality are not reciprocal (Colwell & Rangel, 2009; Soberón & Nakamura, 2009), and the climatic niche of a species might be fully captured even if only a part of its geographic distribution was sampled (Guisan, Petitpierre, Broennimann, Daehler, & Kueffer, 2014). This seems to be the case for 22.9% of the studied species that occupy ranges with highly heterogeneous climatic conditions (e.g. in mountain regions as mentioned above). For those species, the sampling coverage was higher in the climatic space (DMC_{CLIM}) than in geographic space (DMC_{GEO}).

Large-scale biodiversity databases consist of heterogeneous, non-systematically sampled datasets with underdispersed observations within the geographic space and supposedly also within the climatic space. We therefore expected (3) the sampling coverage of species geographic ranges and climatic niches to be largely imperfect due to sampling biases. Accordingly, we found limited sampling coverage for most of the studied species. In almost all cases, the observed species distributions in EVA significantly underrepresented both the species geographic range and climatic niche space. It is achievable to identify species which are poorly represented in biodiversity databases relative to their geographic ranges or realized climatic niches (Boakes et al., 2010; Hoffmann et al., 2014). Since the observed and expected DMC values were highly positively correlated, the applied null model approach supports the usefulness of the presented DMC metric to assess sampling bias in the distribution of species occurrences in biodiversity databases.

We assumed that (4) on condition that range size and climatic niche size are correlated, sampling coverage increases with increasing sample size. The applied linear models revealed a positive effect of sample size on DMC values while accounting for range size and niche size, which supports our fourth hypothesis. Nevertheless, especially for the geographical space, high percentage cover of species range by the EVA plots cannot directly indicate high DMC values. In general, the correlation of percentage match of a species range by the EVA plots at 2.5-min raster cell resolution with DMC values was highly positive in geographic space. Nevertheless, there were species with higher percentage match that only reached lower DMC values while there were also species with lower percentage match that reached higher DMC values. Our results show that the number and thereby the density of observations across a species distribution range remains crucial. On the one hand, too small number of plots representing a species distribution range may be a sample of insufficient size even if the plots are distributed randomly (as suggested by the null model calculations). On the other hand, even a large number of vegetation plots may underrepresent a species range if their spatial distribution is underdispersed. Consequently, both clumping and density of occurrence observations have to be considered, computed and estimated simultaneously to evaluate the representativeness of biodiversity databases.

4.2 Possible applications of the DMC

Occurrence data and distribution maps for species of various taxa are increasingly being made available from biodiversity databases (e.g. Map Of Life, Jetz, McPherson, & Guralnick (2012); The IUCN Red List, IUCN (2019); Euro+Med Plantbase, Euro+Med (2019); The PLANTS Database, USDA, NRCS (2019)).

(I) Our DMC approach enables evaluation and comparison of the coverage of occurrence data across irregular and even non-contiguous background spaces. Thus, it helps identifying species with a suitable representation of their range / niche by existing point samples. In species distribution modelling, uneven or inconsistent representation of environmental gradients by occurrence records can strongly influence the model accuracy (Tessarolo, Rangel, Araújo, & Hortal, 2014), which can result in limited applicability for climate change predictions (Araújo & Guisan, 2006; Titeux et al., 2017).

(II) The DMC value calculation is applicable in both the climatic and geographic space and can help evaluate the coverage of species samples for species distribution modelling. Using such information derived from the DMC metric inside the modelling framework of SDM is likely to improve SDM predictive performance. Nevertheless, independent information on species geographic distribution is needed to correctly evaluate point sampling coverage for SDM studies. It is not recommended to generate range models based on sampling data of unknown coverage. While $DMC_{(GEO)}$ values generated this way might be used to gather information on species geographic point sampling quality, $DMC_{(CLIM)}$ values might be highly biased. Without independently generated distribution information, $DMC_{(CLIM)}$ values are not applicable for SDM evaluation. Since observed and expected DMC values (see the applied null model approach) were highly positively correlated, the deviation from the expected DMC is a suitable measure for the representativeness of species occurrence data. A high deviation corresponds to an underdispersed distribution of plots, while a low deviation corresponds to a more random distribution of plots and a negative deviation corresponds to an overdispersed distribution of plots.

(III) Data limitations (i.e. lack of fine-resolution data of species occurrences over large spatial extents) will remain the norm for most species and regions, and best-possible use should be made of limited information (Hoffmann et al., 2014; Meyer et al., 2016). Here, based on the curves resulting from the DMC calculations it would be possible to determine the raster cell resolution where results of the analyses are least vulnerable to errors due to the existing sampling gaps by calculating the inflection point of the DMC curve. Nevertheless, one must be aware that the achievable raster cell resolution always depends on the spatial extent of the study (e.g. regional, continental or global scale) (Hartley & Kunin, 2003; Pearson & Dawson, 2003; Willis & Whittaker, 2002).

(IV) The efficacy of database platforms strongly depends on the completeness of species inventories and the survey coverage across space and the environment (Hortal et al., 2008; Troia & McManamay, 2016), therefore it is necessary to continue surveys in undersampled areas (Beck et al., 2012;

Engemann et al., 2015). Here, results of the DMC analyses can be used to identify these undersampled areas and help focus search efforts for data information in relevant literature or further databases. This would be possible by selecting undersampled parts of the niche and translate them back to the geographical space. Furthermore, the results of DMC analyses can be used to guide future botanical explorations and practical fieldwork, to make new sampling in geographical and climate spaces cost-efficient.

(V) Including both the DMC metrics as covariates in any model with species as the observational unit may help to account for potential confounding effects due to the varying sampling coverage of the sampled species distribution within both the climatic and geographic space. Since DMC values are species-specific, they can be included as weights in macroecological analyses and models, where well-represented species might be weighted higher than less-well represented species. Nevertheless, it might be necessary to apply re-sampling methods (e.g. Lengyel, Chytrý, & Tichý, 2011) to prevent spatial autocorrelation in model residuals.

Acknowledgements

We thank all scientists who collected vegetation-plot data in the field, the custodians of vegetation-plot databases represented in EVA and the EVA database managers Stephan Hennekens, Borja Jiménez-Alfaro and Ilona Knollová whose contributions were essential for this broad-scale study.

Author contributions

EW and MS developed the DMC concept, with considerable input by GS and HB. MS wrote the first draft of the manuscript, with considerable input by EW, HB, PK and UJ. MS and GS harmonized data retrieved from EVA and CDH. GS wrote R code for DMC calculation. PK wrote R code for the null model application for DMC calculations. MS carried out statistical analyses and produced the graphs. All other authors contributed data. All authors contributed to writing the manuscript.

Data accessibility

The R code for DMC calculation with an application example is available from Figshare Digital Repository: <https://doi.org/10.6084/m9.figshare.7924934.v2>.

References

Araújo, M. B. & Guisan, A. (2006). Five (or so) challenges for species distribution modelling. *Journal of Biogeography, 33,* 1677–1688.

Beck, J., Ballesteros-Mejia, L., Buchmann, C. M., Dengler, J., Fritz, S. A., Gruber, B., ..., Dormann, C. F. (2012). What's on the horizon for macroecology? *Ecography, 35,* 673–683.

Bertrand, R., Lenoir, J., Piedallu, C., Riofrío-Dillon, G., de Ruffray, P., Vidal, C., …, Gégout, J.-C. (2011). Changes in plant community composition lag behind climate warming in lowland forests. *Nature, 479*, 517–520.

Boakes, E. H., McGowan, P. J. K., Fuller, R. A., Chang-qing, D., Clark, N. E., O'Connor, K., & Mace, G. M. (2010). Distorted views of biodiversity: spatial and temporal bias in species occurrence data. *PLOS Biology*, *8*, e1000385.

Boots, B. N., & Getis, A. (1988). *Point pattern analysis (Vol. 8).* Newbury Park, CA, US: Sage Publications Inc.

Broennimann, O., & Guisan, A. (2008). Predicting current and future biological invasions: both native and invaded ranges matter. *Biology Letters*, *4*, 585–589.

Bruelheide, H., Dengler, J., Jiménez-Alfaro, B., Purschke, O., Hennekens, S., Chytrý, M., …, Winter, M. (2019). sPlot – a new tool for global vegetation analyses. *Journal of Vegetation Science*, *30*, 161-186.

Bruelheide, H., Dengler, J., Purschke, O., Lenoir, J., Jiménez-Alfaro, B., Hennekens, S. M., …, Jandt, U. (2018). Global trait–environment relationships of plant communities. *Nature Ecology & Evolution*, *2*, 1906–1917.

Bütof, A., von Riedmatten, L.R., Dormann, C.F., Scherer-Lorenzen, M., Welk, E., & Bruelheide, H. (2012). The responses of grassland plants to experimentally simulated climate change depend on land use and region. *Global Change Biology, 18*, 127–137.

Chamberlain, S. A., & Szöcs, E. (2013). taxize - taxonomic search and retrieval in R. *F1000 Research, 2,* 191.

Chamberlain, S. A., Szöcs, E., Foster, Z., Arendsee, Z., Boettiger, C., Ram, K., Baratomeus, I., …, O'Donnell, J. (2018). *taxize: Taxonomic information from around the web. R package version 0.9.3.*

Chytrý, M., Hennekens, S. M., Jiménez-Alfaro, B., Knollová, I., Dengler, J., Jansen, F., …, Yamalov, S. (2016). European Vegetation Archive (EVA): an integrated database of European vegetation plots. *Applied Vegetation Science, 19,* 173–180.

Colwell, R. K., & Rangel, T. F. (2009). Hutchinson's duality: the once and future niche. *Proceedings of the National Academy of Sciences of the United States of America*, *106*, 19651–19658.

Csergő, A. M., Salguero-Gómez, R., Broennimann, O., Coutts, S. R., Guisan, A., Angert, A. L., ..., Buckley, Y. M. (2017). Less favourable climates constrain demographic strategies in plants. *Ecology Letters, 20,* 969–980.

Daru, B. H., Park, D. S., Primack, R. B., Willis, C. G., Barrington, D. S., Whitfeld, T. J. S., ..., Davis, C. C. (2018). Widespread sampling biases in herbaria revealed from large-scale digitization. *New Phytologist, 217*, 939–955.

Edwards, J. L., Lane, M. A., & Nielsen, E. S. (2000). Interoperability of biodiversity databases: Biodiversity information on every desktop. *Science, 289*, 2312–2314.

Engemann, K., Enquist, B. J., Sandel, B., Boyle, B., Jørgensen, P. M., Morueta–Holme, N., …, Svenning, J.-C. (2015). Limited sampling hampers "big data" estimation of species richness in a tropical biodiversity hotspot. *Ecology and Evolution*, *5*, 807–820.

Enquist, B. J., R. Condit, B. Peet, M. Schildhauer, B. Thiers, and BIEN working group. (2009). The Botanical and Information Ecology Network (BIEN): Cyberinfrastructure for an integrated botanical information network to investigate the ecological impacts of global climate change on plant biodiversity. Available at http:// www.iplantcollaborative.org/sites/default/files/ BIEN_White_Paper.pdf

Euro+Med (2019, February 1). Euro+Med PlantBase – the information resource for Euro-Mediterranean plant diversity. http://ww2.bgbm.org/EuroPlusMed/.

Fourcade, Y., Engler, J. O., Rödder, D., & Secondi, J. (2014). Mapping species distributions with MAXENT using a geographically biased sample of presence data: a performance assessment of methods for correcting sampling bias. *PLOS ONE*, *9*, e97122.

Gomes, V. H. F., Ijff, S. D., Raes, N., Amaral, I. L., Salomão, R. P., de Souza Coelho, L., …, ter Steege, H. (2018). Species distribution modelling: contrasting presence-only models with plot abundance data. *Scientific Reports (2018), 8,* 1003.

Guisan, A., Petitpierre, B., Broennimann, O., Daehler, C., & Kueffer, C. (2014). Unifying niche shift studies: insights from biological invasions. *Trends in Ecology & Evolution*, *29*, 260–269.

Hall, P., & Wood, A. (1993). On the performance of box-counting estimators of fractal dimension. *Biometrika, 80,* 246–252.

Hartley, S., & Kunin, W. E. (2003). Scale dependency of rarity, extinction risk, and conservation priority. *Conservation Biology, 17*, 1559-1570.

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., & Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. I*nternational Journal of Climatology, 25,* 1965–1978.

Hirst, M. J., Griffin, P. C., Sexton, J. P., & Hoffmann, A. A. (2017). Testing the niche-breadth–range-size hypothesis: habitat specialization vs. performance in Australian alpine daisies. *Ecology, 98*, 2708–2724.

Hoffmann, A., Penner, J., Vohland, K., Cramer, W., Doubleday, R., Henle, K., …, Häuser, C. L. (2014). Improved access to integrated biodiversity data for science, practice, and policy - the European Biodiversity Observation Network (EU BON). *Nature Conservation*, *6*, 49–65.

Hofmann, M., Bütof, A., Welk, E., & Bruelheide, H. (2013). Relationship between fundamental and realized niches in terms of frost and drought resistance. *Preslia*, *85*, 1–17.

Hortal, J., Borges, P. A., & Gaspar, C. (2006). Evaluating the performance of species richness estimators: sensitivity to sample grain size. *Journal of Animal Ecology, 75*, 274–287.

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

Hutchinson, G. E. (1978). *An introduction to population ecology.* New Haven, CT, US: Yale University Press.

IUCN (2019, February 1). The IUCN Red List of Threatened Species. Version 2018-2. http://www.iucnredlist.org.

Jandt, U., von Wehrden, H., & Bruelheide, H. (2011). Exploring large vegetation databases to detect temporal trends in species occurrences. *Journal of Vegetation Science, 22,* 957-972.

Jansen, F., & Dengler, J. (2008). GermanSL – Eine universelle taxonomische Referenzliste für Vegetationsdatenbanken in Deutschland. *Tuexenia, 28,* 239– 253.

Jetz, W., McPherson, J. M., & Guralnick, R. P. (2012). Integrating biodiversity distribution knowledge: toward a global map of life. *Trends in Ecology and Evolution, 27*, 151-159.

Jiménez-Alfaro, B., Girardello, M., Chytrý, M., Svenning, J.-C., Willner, W., Gégout, J.-C., ..., Wohlgemuth, T. (2018). History and environment shape species pools and community diversity in European beech forests. *Nature Ecology & Evolution*, *2*, 483–490.

Köckemann, B., Buschmann, H., & Leuschner, C. (2009). The relationships between abundance, range size and niche breadth in Central European tree species. *Journal of Biogeography*, *36*, 854–864.

Lengyel, A., Chytrý, M., & Tichý, L. (2011). Heterogeneity-constrained random resampling of phytosociological databases. *Journal of Vegetation Science, 22,* 175–183.

Lobo, J. M., Hortal, J., Yela, J. L., Millán, A., Sánchez-Fernández, D., García-Roselló, E., …, Guisande, C. (2018). KnowBR: An application to map the geographical variation of survey effort and identify well-surveyed areas from biodiversity databases. *Ecological Indicators*, *91*, 241–248.

Lundqvist, J. & Nordenstam, B. (1988). *Index Holmiensis vol. 6.* Swedish Museum of Natural History, Stockholm.

Lundqvist, J. (1992). *Index Holmiensis vol. 7.* Swedish Museum of Natural History, Stockholm.

Lundqvist, J. & Jäger, E. J. (1995-2007). *Index Holmiensis vol. 8-10.* Swedish Museum of Natural History, Stockholm.

Marsh, C. J., Barwell, L. J., Gavish, Y., & Kunin, W. E. (2018). downscale: an R package for downscaling species occupancy from coarse-grain data to predict occupancy at fine-grain sizes. *Journal of Statistical Software*, *86*.

Meusel, H., Jäger, E. J. & Weinert, E. (1965). *Vergleichende Chorologie der zentraleuropäischen Flora, Karten, Band I.* VEB Gustav Fischer Verlag.

Meusel, H., Jäger, E. J., Rauschert, S. & Weinert, E. (1978). *Vergleichende Chorologie der zentraleuropäischen Flora, Karten, Band II.* VEB Gustav Fischer Verlag.

Meusel, H. & Jäger, E. J. (1992). *Vergleichende Chorologie der zentraleuropäischen Flora, Karten, Band III.* Gustav Fischer Verlag.

Meyer, C., Weigelt, P., & Kreft, H. (2016). Multidimensional biases, gaps and uncertainties in global plant occurrence information. *Ecology Letters*, *19*, 992–1006.

Nunes, L. A. & Pearson, R. G. (2017). A null biogeographical test for assessing ecological niche evolution. *Journal of Biogeography, 44*, 1331–1343.

Pearman, P. B., Guisan, A., Broennimann, O., & Randin, C. F. (2008). Niche dynamics in space and time. *Trends in Ecology & Evolution, 23*, 149–158.

Pearson, R. G. & Dawson, T. P. (2003). Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful? *Global Ecology and Biogeography*, *12*, 361–371.

Pentland, A. P. (1984). Fractal-based description of natural scenes. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, *6*, 661–674.

Peterson, A. T., Soberón, J., Pearson, R. G., Anderson, R. P., Martínez-Meyer, E., Nakamura, M., & Araújo, M. B. (2011). *Ecological niches and geographic distributions (MPB-49).* Princeton, NJ, US: Princeton University Press.

Pyke, G. H. & Ehrlich, P. R. (2010). Biological collections and ecological/environmental research: a review, some observations and a look to the future. *Biological Reviews, 85*, 247–266.

Ricklefs, R. E. (2004). A comprehensive framework for global patterns in biodiversity. *Ecology Letters*, *7*, 1–15.

San-Miguel-Ayanz, J., de Rigo, D., Caudullo, G., Houston Durrant, T., & Mauri, A. (Eds.) (2016). *European atlas of forest tree species*. Luxembourg, LU: Publication Office of the European Union. DOI: 10.2788/038466

Schleuning, M., Fründ, J., Schweiger, O., Welk, E., Albrecht, J., Albrecht, M., ..., Hof, C. (2016). Ecological networks are more sensitive to plant than to animal extinction under climate change. *Nature Communications, 7*, 13965.

Soberón, J. & Nakamura, M. (2009). Niches and distributional areas: Concepts, methods, and assumptions. *Proceedings of the National Academy of Sciences of the United States of America, 106*, 19644–19650.

Soberón, J., Jiménez, R., Golubov, J., & Koleff, P. (2007). Assessing completeness of biodiversity databases at different spatial scales. *Ecography, 30*, 152–160.

Soria-Auza, R. W., & Kessler, M. (2007). The influence of sampling intensity on the perception of the spatial distribution of tropical diversity and endemism: a case study of ferns from Bolivia. *Diversity and Distributions*, *14*, 123–130.

Sousa-Baena, M. S., Garcia, L. C., & Peterson, A. T. (2014). Completeness of digital accessible knowledge of the plants of Brazil and priorities for survey and inventory. *Diversity and Distributions, 20*, 369–381.

Speed, J. D. M., Bendiksby, M., Finstad, A. G., Hassel, K., Kolstad, A. L., & Prestø, T. (2018). Contrasting spatial, temporal and environmental patterns in observation and specimen based species occurrence data. *PLOS ONE*, *13*, e0196417.

Tessarolo, G., Rangel, T., Araújo, M. B., & Hortal, J. (2014). Uncertainty associated with survey design in Species Distribution Models. *Diversity and Distributions, 20,* 1258–1269.

Titeux, N., Maes, D., Daele, T. V., Onkelinx, T., Heikkinen, R. K., Romo, H., …, Luoto, M. (2017). The need for large-scale distribution data to estimate regional changes in species richness under future climate change. *Diversity and Distributions, 23*, 1393–1407.

Tralau, H. (1969-1981). *Index Holmiensis vol. 1-5.* Swedish Museum of Natural History, Stockholm.

Troia, M. J. & McManamay, R. A. (2016). Filling in the GAPS: evaluating completeness and coverage of open-access biodiversity databases in the United States. *Ecology and Evolution, 6*, 4654–4669.

USDA, NRCS. (2019, February 1). The PLANTS Database. National Plant Data Team, Greensboro, NC 27401-4901 USA. http://plants.usda.gov.

Wasof, S., Lenoir, J., Aarrestad, P. A., Alsos, I. G., Armbruster, W. S., Austrheim, G.,..., Decocq, G. (2015). Disjunct populations of European vascular plant species keep the same climatic niches. *Global Ecology and Biogeography*, *24*, 1401–1412.

Welk, A., Welk, E., & Bruelheide, H. (2014). Biotic interactions overrule plant responses to climate, depending on the species' biogeography. *PLoS ONE, 9*, e111023.

Wiegand, T. & Moloney, K. A. (2013). *Handbook of Spatial Point-Pattern Analysis in Ecology.* Boca Raton, FL, US: CRC Press.

Willis, K. J. & Whittaker, R. J. (2002). Species Diversity-Scale Matters. *Science, 295*, 1245–1248.

Woodward, F. I. (1986). *Climate and plant distribution.* Cambridge, UK: Cambridge University Press.

Yang, W., Ma, K., & Kreft, H. (2013). Geographical sampling bias in a large distributional database and its effects on species richness–environment models. *Journal of Biogeography*, *40*, 1415–1426.

Figures

Figure 1 Distribution of the 808,794 vegetation plots (green dots) extracted from EVA (European Vegetation Archive). Only plots with at least one of the 564 study species are shown. The study species merged distributions based on CDH are represented by grey cells. White areas (large water bodies, glaciers, and deserts) represent regions where none of the studied species occurs. (a) Distribution of vegetation plots in the geographic space. (b) Distribution of vegetation plots in climatic space represented by its first two PCA axes (74.1% and 13.9% variance explained by PC1 and PC2, respectively), where PC1 and PC2 were negatively and positively related to temperature and precipitation, respectively.

Figure 2 Study species data density in the geographic and climatic space. (a) Data density on species geographic ranges of 564 vascular plant species included in this study in 2.5-min resolution raster. White areas (large water bodies, glaciers, and deserts) represent regions where none of the studied species occurs. (b) Data density on climatic niches of 564 species in the respective common climatic space represented by its first two PCA axes (74.1% and 13.9% variance explained by PC1 and PC2, respectively), where PC1 and PC2 were negatively and positively related to temperature and precipitation, respectively.

Figure 3 Dynamic Match Coefficient (DMC) calculated for two example species X and Y with different plot distributions but similar ranges and climatic niches. DMC measures sampling coverage from fine resolution to coarse resolution as the area under the curve (AUC). Scaling for species X, with clumped plots (10 red dots) in the species range or climatic niche (grey background), results in a low DMC value. Scaling for species Y, with more regularly distributed plots (10 blue dots) in the species range or climatic niche (grey background), results in a high DMC value.

Figure 4 The DMC scaling approach applied to the distribution of EVA vegetation plots inside species ranges in geographic space and inside species niches in climatic space (grey cells). (a) The distribution of EVA plots containing *Hieracium murorum* (left, red) and *Calluna vulgaris* (right, blue). (b) Four selected scaling steps from fine to coarse raster-cell resolution in geographic space (left-hand four panels in each set) and climatic space (right-hand four panels in each set). (c) The resulting DMC curves along 20 scaling steps, where the cell match ratio is the percentage of grey raster cells (species range or climatic niche) matched by a vegetation plot containing the species. In all cases, the maximum achievable DMC is 1 (100% cell match in all scaling steps). DMC values reached 0.42 and 0.58 for the geographic (DMC_{GEO}) and climatic (DMC_{CLIM}) space for *Hieracium murorum* and 0.74 for both the geographic (DMC_{GEO}) and climatic (DMC_{CLIM}) space for *Calluna vulgaris*.

Figure 5 Scatterplot and Spearman correlation coefficients (rho) of the relationship between DMC values in geographic space (DMC_{GEO}) and DMC values in climatic space (DMC_{CLIM}) for 564 plant species. Low DMC values indicate an underdispersed (more clumped) distribution of species occurrences in EVA vegetation plots, while high DMC values indicate a homogenous distribution in EVA vegetation plots, in the geographic range or realized climatic niche of a species.

Figure 6 Scatterplots and Spearman correlation coefficients (rho) of the relationships between the observed DMC and expected DMC derived by null models for (a) geographic space and (b) climatic space. Dots are medians; lines are inter-quartile ranges of the simulations from the null model. Colour gradient represents the percentage match of a species range by EVA vegetation plots in the geographic space (match at 2.5-min raster cell resolution) or climate space (ratio of PCA cells matched by EVA plots to all species-specific raster cells matched by the geographic range data in the PCA space).

Figure 7 Scatterplot and Spearman correlation coefficients (rho) of the relationship between the deviation of the observed DMC values from null model DMC values in the geographic space (DEV_{GEO}) and in climatic space (DEV_{CLIM}). Low deviation of the observed DMC values from the null expectation indicates a more regular distribution of occurrences for a given species across its reference range or realized climatic niche, a high deviation indicates an underdispersed (more clumped) distribution.

Figure 8 Scatterplots and Spearman correlation coefficients (rho) of the relationships between percentage match of species ranges by EVA vegetation plots and (a) observed DMC in geographic space (DMC_{GEO}); (b) observed DMC in climatic space (DMC_{CLIM}); (c) deviation of observed DMC values from null model DMC values in geographic space (DEV_{GEO}); (d) deviation of observed DMC values from null model DMC values in climatic space (DEV_{CLIM}).

Supporting Information

Appendix S1 Climatic resampling procedure and background PCA niche space of the study area.

Appendix S2 Information on the 564 species included in this study.

Appendix S3 Information on the 59 databases that provided vegetation plots included in this study.

Appendix S4 Information on initial grain size in DMC calculations; correlation between percentage match of species ranges by EVA vegetation plots in geographic vs. climatic space; correlation between species range sizes and niche sizes.