Assessing spatial and temporal biases and gaps in the publicly available distributional information of Iberian mosses

Cristina Ronquillo[‡], Fernanda Alves-Martins[‡], Vicente Mazimpaka[§], Thadeu Sobral-Souza[|], Bruno Vilela-Silva[¶], Nagore G. Medina[§], Joaquín Hortal[‡,#,¤]

‡ Dept. Biogeography & Global Change, Museo Nacional de Ciencias Naturales (MNCN-CSIC), Madrid, Spain
§ Dept. Biología (Botánica), Facultad de Ciencias, Universidad Autónoma de Madrid, Madrid, Spain
| Dept. Botanica e Ecologia, Universidade Federal de Mato Grosso (UFMT), Cuiaba, Brazil
¶ Instituto de Biologia, Universidade Federal da Bahia. 1154, R. Barão de Jeremoabo, 668 - Ondina, Salvador, Brazil
# Universidade Federal de Goiás, Goiânia, Brazil
¤ Faculdade de Ciências da Universidade de Lisboa, Lisboa, Portugal

Corresponding author: Cristina Ronquillo (cristinaronquillo@mncn.csic.es)

Academic editor: Yasen Mutafchiev

Abstract

One of the most valuable initiatives on massive availability of biodiversity data is the Global Biodiversity Information Facility, which is creating new opportunities to develop and test macroecological knowledge. However, the potential uses of these data are limited by the gaps and biases associated to large-scale distributional databases (the so-called Wallacean shortfall). Describing and quantifying these limitations are essential to improve knowledge on biodiversity, especially in poorly-studied groups, such as mosses. Here we assess the coverage of the publicly-available distributional information on Iberian mosses, defining its eventual biases and gaps. For this purpose, we compiled IberBryo v1.0, a database that comprises 82,582 records after processing and checking the geospatial and taxonomical information. Our results show the limitations of data and metadata of the publicly-available information. Particularly, ca. 42% of the records lacked collecting date information, which limits data usefulness for time coverage analyses and enlarges the existing knowledge gaps. Then we evaluated the overall coverage of several aspects of the spatial, temporal and environmental variability of the Iberian Peninsula. Through this assessment, we demonstrate that the publicly-available information on Iberian mosses presents significant biases. Inventory completeness is strongly conditioned by the recorders' survey bias, particularly in northern Portugal and eastern Spain and the spatial pattern of surveys is also biased towards mountains. Besides, the temporal pattern of survey effort intensifies from 1970 onwards, encompassing a progressive increase in the geographic coverage of the Iberian Peninsula. Although we just found 5% of well-surveyed cells of 30’ of resolution over the 1970-2018 period, they cover about a fifth of the main climatic gradients of the Iberian

Peninsula, which provides a fair – though limited – coverage. Yet, the well-surveyed cells are biased towards anthropised areas and some of them are located in areas under intense land-use changes, mainly due to the wood-fires of the last decade. Despite the overall increase, we found a noticeable gap of information in the south-west of Iberia, the Ebro river basin and the inner plateaus. All these gaps and biases call for a careful use of the available distributional data of Iberian mosses for biogeographical and ecological modelling analysis. Further, our results highlight the necessity of incorporating several good practices to increase the coverage of high-quality information. These good practices include digitalisation of specimens and metadata information, improvement on the protocols to get accurate data and metadata or revisions of the vouchers and recorders' field notebooks. These procedures are essential to improve the quality and coverage of the data. Finally, we also encourage Iberian bryologists to establish a series of re-surveys of classical localities that would allow updating the information on the group, as well as to design their future surveys considering the most important information gaps on IberBryo.

Keywords

Biodiversity data, Bryophyta, Global Biodiversity Information Facility, IberBryo, Iberian Peninsula, Inventory completeness, Wallacean Shortfall

Introduction

The current massive availability of biodiversity data is creating new opportunities to develop and test macroecological knowledge (Hampton et al. 2013, Morueta-Holme and Svenning 2018). Advances in the management (i.e. acquisition, cleaning and integration) and analysis of 'biodiversity big data' are crucial (Gandomi and Haider 2015, Devictor and Bensaude-Vincent 2016), thus promoting the emergence of new fields such as eco-informatics and biodiversity informatics (Bisby 2000, Soberón and Peterson 2004). One of the most valuable initiatives on this matter is the Global Biodiversity Information Facility (GBIF, http://www.gbif.org/), a distributed network of databases that seeks to provide open access to all biodiversity data through the internet (Saarenmaa and Nielsen 2002). The GBIF platform offers a vast amount of primary distributional information that allows outlining large-scale questions from a data-driven approach (García-Roselló et al. 2015, Franklin et al. 2017).

Advances in big biodiversity data tools and computational power are continually increasing the potential offered by this information (Bisby 2000, Maldonado et al. 2015, Devictor and Bensaude-Vincent 2016, Wüest et al. 2019). However, managing the vast amount of data is challenging due to its large volume and the high variability, velocity and variety in the creation, veracity and value of data (Gandomi and Haider 2015, Devictor and Bensaude-Vincent 2016, Wüest et al. 2019). Data pre-processing is key to reach adequate levels of quality and reliability of the records that are finally analysed (Calabrese 2019). The more common limitations of biodiversity data are related to

georeferencing and taxonomy (Soberón and Peterson 2004, Wieczorek et al. 2004, Yesson et al. 2007, Sousa-Baena et al. 2014, Isaac and Pocock 2015) and data cleaning processes have an important role in their solution (Chapman 2005, Gandomi and Haider 2015, Maldonado et al. 2015, Gueta and Carmel 2016, Calabrese 2019).

Once these issues are handled, the subsequent task would be to assess the quality of data as a whole. In the particular case of macroecology and biogeography, this means addressing the gaps and biases associated to large-scale databases (Hortal et al. 2007, Beck et al. 2013, Engemann et al. 2015, Amano et al. 2016, Meyer et al. 2016), which compromise the description of biodiversity patterns (Hortal et al. 2008, Boakes et al. 2010 , Yang et al. 2013, Beck et al. 2014, Hortal et al. 2015, Morueta-Holme and Svenning 2018). By evaluating and describing how these limitations affect the geographic distribution of species – the so-called Wallacean shortfall (Lomolino 2004) – it is possible to enhance the insights obtained with these data and also design research seeking to fill in the gaps in this knowledge (Rocchini et al. 2011, Hortal et al. 2015, Morueta-Holme and Svenning 2018, Wetzel et al. 2018). Essentially, the Wallacean shortfall is due to uneven sampling effort through space and time, typically caused by the historical patterns of collecting and analysing data (Hortal et al. 2007, Hortal et al. 2008, Sastre and Lobo 2009, Hortal et al. 2015, Isaac and Pocock 2015, Maldonado et al. 2015, Amano et al. 2016). To overcome this shortfall, we need to evaluate and quantify the survey coverage of biodiversity data along space, time, environment and taxonomy (Hortal et al. 2008, Boakes et al. 2010, Meyer et al. 2015, Meyer et al. 2016, Troia and McManamay 2016).

The extent of the Wallacean shortfall varies considerably amongst taxonomic groups (Amano et al. 2016, Troia and McManamay 2016), depending on the historical interest on the survey or study of each one of them. While the study of diversity patterns at large scales has been mainly focused on vascular plants and vertebrates (Mutke and Barthlott 2005, Aranda et al. 2015), bryophytes have been considered just on a few occasions (Mutke and Geffert 2010, Geffert et al. 2013, Hespanhol et al. 2015, Mateo et al. 2016, Berdugo et al. 2018). Therefore, although the knowledge on this highly-diverse group of organisms has been developed over a long historical period (Magill 2010), especially in Europe (Mutke and Geffert 2010), the quality of moss distributional data has been scarcely assessed (Callaghan and Ashton 2008, Mutke and Geffert 2010, Aranda et al. 2011, Meyer et al. 2016). As a result, the coverage of its spatial and temporal distributional information is poorly-known and may, indeed, reflect the historical pattern of surveys, rather than the actual diversity of this group (Mutke and Barthlott 2005).

Here we aim to assess and quantify the knowledge on the publicly-available distributional information on Iberian mosses, defining its eventual biases and gaps. To do this, we compile an extensive Iberian moss database, process its records to filter those with adequate quality and then analyse their coverage. Specifically we aim to: (i) assess the overall quality of moss records in the Iberian Peninsula; (ii) evaluate their substrate, altitudinal, temporal and spatial coverage; (iii) analyse their inventory completeness; and (iv) assess the adequacy of well-surveyed areas to recover the responses of biodiversity to climatic and land-use changes.

Materials and Methods

Pre-processing of occurrence data

We downloaded 97,597 records of mosses (keyword Phylum: Bryophyta) for the Iberian Peninsula – defined as mainland Portugal and Spain, plus the Balearic Islands, Andorra and Gibraltar – from GBIF (GBIF 2018a, accessed 8 August 2018 for Spain and Portugal and GBIF 2018b, accessed 9 October 2018 for Andorra and Gibraltar). We also retrieved 5,876 occurrences from two PhD dissertations that comprised geographically-extensive surveys, encompassing several Spanish provinces and climatic zones (Cezón and Muñoz 2013, Medina et al. 2015). Records from Medina et al. (2015) – that include previously-surveyed areas in Galicia and Asturias from Albertos (2001) – were published in GBIF afterwards and they are now available in Medina and Ronquillo (2020). In total, the version 0.1 of our database (hereafter called IberBryo) held 103,473 unprocessed raw records. We will consider only good-quality occurrences for our analysis, i.e. those that represent an individual organism collected from certain location (i.e. latitude and longitude) and at a given time, such as, at least, calendar year (Troia and McManamay 2016; see also Hortal and Lobo 2005). In order to check and improve the quality of IberBryo records, we performed a data cleaning protocol (Fig. 1, Suppl. material 12) addressing the three main issues that may affect the quality of biological records: geospatial location, taxonomical identification and temporal allocation.

Geospatial validation. We checked the coordinates of all records following their available geographic location through 'point-in-polygon' test at province/district level with QGIS Development Team (2019) software and Global Administrative Areas (2018) country layers. Records that presented numerical sign errors were manually corrected, based on their locality description. Those placed on the sea, less than 10 km from the coast, were relocated at the nearest coastal place. Then, we georeferenced records without coordinates that presented a specific 'named place' (Wieczorek et al. 2004) in the locality description through geocoding using the corresponding official national gazetteers (as the geographic centre or locality centroids): *"Nomenclátor de Municipios y Entidades de población"* and *"Nomenclátor Básico"* of Instituto Geográfico Nacional (IGN) for Spanish records; *"Servicio de Localização Toponímica del Grupo Crise Rede de Informaçao de Situações de Emergencia"* for Portuguese records and *"Nomenclàtor Oficial del Govern d'Andorra"* for Andorran records. Finally, we discarded records lacking coordinates and outliers whose locality description was missing or inaccurate and those located on the sea more than 10 km from the coast.

Taxonomic validation and standardisation. We checked all species names (extracted from GBIF fields "scientific_name" and "genus" + "species") to remove fossil specimens, misidentifications, wrong country locations or insufficient taxon rank identification. Records were reviewed following the checklists in Casas et al. (2006), Hill et al. (2006), Ros et al. (2013), Hodgetts (2015), Sotiaux and Vanderpoorten (2017) and *Flora Briofítica*

Ibérica (Guerra et al. 2006, Brugués et al. 2007, Guerra et al. 2010, Guerra et al. 2014, Brugués and Guerra 2015, Guerra et al. 2018) under the expert supervision of one of us (VM). Subsequently, we unified the list of species names (correction of spelling, synonyms and authority standardisation) according to Hill et al. (2006) and Ros et al. (2013). For the assignation of the species name, we gave priority to the most recent checklist, except for taxa that have further experienced taxonomic or nomenclatural changes: for example, *Bartramia stricta* (Müller 2014), *Orthotrichum* (Plášek et al. 2015, Lara et al. 2016), *Codonoblepharon forsteri* (Goffinet et al. 2004, Mazimpaka and Lara 2014) and *Oxystegus tenuirostris* (Alonso et al. 2016, Alonso et al. 2018).

Year validation. We excluded all the occurrences without collecting date information at year level in the IberBryo v1.1 database to perform the climatic and land-use coverage analyses (see below), although we kept them in the IberBryo v1.0.

Assessing survey coverage

Once all records had been pre-processed, we assessed the overall coverage of the spatial, temporal and environmental variability of the Iberian Peninsula provided by the inventories contained in IberBryo. All analyses were performed in R (R Development Core Team 2019 v 3.6) and RStudio (RStudio Team 2019 v 1.2) environment and coverage maps were customised in RWizard version 4.3 (Guisande et al. 2014). See the relation of scripts used in Suppl. material 13.

Substrate coverage. Due to the absence of habitat-type information in most of the records, we were only able to assess the coverage of ecological substrates by checking in specialised references all the taxa that thrive in each type of substrate. First, we made a simplified reclassification based on BRYOATT (Hill 2007), assigning each species to the following five substrate classes: rock, epiphytic, soil, aquatic and decaying vegetation. This reclassification includes information of the frequency of use for each species as follows: [1] Rare substrate [2] Occasional substrate [3] Normal substrate. Then, for taxa not included in this guide, the information was extracted from Dierssen (2001), Casas et al. (2006), Garilleti and Albertos (2012) and *Flora Briofitica Ibérica* (Guerra et al. 2006, Brugués et al. 2007, Guerra et al. 2010, Guerra et al. 2014, Brugués and Guerra 2015, Guerra et al. 2018).

Altitudinal coverage. We applied a Kolmogorov-Smirnov test to assess whether the altitudinal range, covered by moss occurrences, represented the altitudinal patterns of the study area. We attributed altitudinal data to each occurrence using a digital elevation model (DEM) of the study area at a spatial resolution of 30 arc-seconds, extracted from GMTED2010 (U.S. Geological Survey 2010) and the Iberian altitudinal patterns were calculated for all DEM data.

Temporal coverage. We represented the historical accumulation of new species (excluding infraspecific taxa) recorded in IberBryo and the number of records gathered by calendar years. Then we evaluated the relationship between number of records and newly-observed species per year through Spearman correlations. We defined different

periods of data collection for the following analyses, based on the information provided by the curve and the main historical periods happening in the Iberian countries.

Spatial coverage and survey completeness. We calculated basic metrics of spatial coverage (number of records, observed richness and completeness) for all Iberian grid cells at two different resolutions, 5' (~65 km^2) and 30' (~2500 km^2), using the R package 'KnowBR' v 2.0 (Lobo et al. 2018). Metrics were calculated for each of the periods of data collection previously identified, as well as for the whole time series and the complete IberBryo v1.0 database (including occurrences without collecting date). We quantified inventory completeness in grid cells of 30' resolution as a metric of survey quality coverage. Completeness for each grid cell was calculated by adjusting the species accumulation curve (i.e. accumulated number of species by records) to the Michaelis-Menten equation (Clench 1979, Soberón and Llorente 1993) and calculating the percentage of the moss flora of each cell predicted by the curve represented in the inventories (see Lobo et al. 2018). Cells with percentages of completeness higher than 80% and ≥ 10 records were considered as well-surveyed, while those with 70-80% were considered moderately well-surveyed and those with less than 70% as poorly-surveyed cells. These thresholds are arbitrary, based on our general knowledge on the survey process and our experience on surveying Iberian mosses. Therefore, qualifying cells as well-surveyed does not mean that their inventory is complete (or nearly complete), but rather that the species missing from the inventory are locally rare and/or inconspicuous. Although these absences are part of the moss assemblage of the grid cell, we assume that their importance for the diversity of local moss communities during the historical period represented by the surveys has been minimal.

We also obtained the location of the main bryology centres of Spain and Portugal. This selection was based on the more frequent affiliation centres collected on SCOPUS publications with the keywords "Bryophyte", "moss", "musgo" or "briofito". We also extracted the location of recently-published PhD theses on bryophytes from Hespanhol et al. (2015) and checked the presence of this information in IberBryo. This allowed us to discuss and compare whether the spatial coverage results were biased by the spatial location of bryological research sources.

Climatic coverage. We assessed the coverage of the climatic variability of the Iberian Peninsula provided by the set of well-surveyed grid cells. To do this, we characterised the climatic environmental space of the Iberian Peninsula, based on the 19 bioclimatic variables from WorldClim 2.0 (Fick and Hijmans 2017) at 10' resolution, aggregating them into the 30' resolution cells of our study area. We performed a PCA to reduce the dimensionality of these data, obtaining two significant PCA axes that represent the main climatic gradients within Iberia and calculated the frequency of climate conditions in the Iberian Peninsula, based on the scores. Then, we quantified the overlap between the climatic space covered by the well-sampled cells and the climatic environmental space of the whole study area through the Schoener's D index (Schoener 1974, Broennimann et al. 2012). Briefly, this index provides a measure of the overlap of two environmental envelopes, from 0 to 1 (complete overlap); in this particular case, Schoener's D value provides a measure of the proportion of the Iberian climatic variability covered by the

well-sampled cells, as measured by the climatic PCA axes. We applied a Kolmogorov-Smirnov test to verify whether the distribution of climates shows statistically significant differences between all grid cells and well-sampled cells. We also quantified the 'rarity index' of these Iberian climate types as a 'Min-Max scalling'. Based on their relative frequency, values are scaled from 0 — very common climate types — to 1 — very 'rare' or climatically unique. We also applied a Kolmogorov-Smirnov test to verify whether the distribution of well-sampled cells is biased to a certain climate type.

Land-use change coverage. We assessed the adequacy of moss data for representing changes in moss assemblages driven by recent land-use modifications in the Iberian Peninsula, following the method used for climatic coverage. We characterised recent land-use variations using information from *Corine Land Cover Changes* (Corine Land Cover seamless vector database- CLC v. 20; European Environmental Agency 2018) in different periods (1990-2000, 2000-2006, 2006-2012 and 2012-2018), available for Spain and Portugal. We reclassified the original CLC classes into simplest categories, according to the importance of each land-use type for bryophyte natural history (Suppl. material 11, Reclassification 1). We quantified the number of land use changes and their occupied area using the previous climatic grid of 30' resolution cells from 1990 to 2018. We also assessed the 'anthropised change ratio' of the cells, based on a reclassification into artificial surfaces ('Anthropic') and natural surfaces (Suppl. material 11, Reclassification 2), as follows: 'Anthropised only' (Natural to Artificial); 'Mostly anthropised' (Natural to Artificial > Artificial to Natural); 'Equally changed' (Natural to Artificial = Artificial to Natural) and 'Naturalised' (Natural to Artificial < Artificial to Natural).

Results

Overall assessment of 'IberBryo' database

Version 1.0 of IberBryo database (Suppl. material 1; Ronquillo and Hortal 2020) includes 82,582 records after pre-processing validations, out of the 103,473 occurrences initially retrieved (Fig. 1). Only 57.80% (47,730) of these processed records include year information from 1783 to 2018 and, therefore, they comprise the bulk of IberBryo v1.1. We could retrieve 14.91% (14,549) of GBIF records mainly through the assignment of coordinates according to the locality description, while we had to delete 19.15% (18,696) of them due to geospatial errors (Fig. 1, see also Suppl. material 2). By countries, Spain contributes with most occurrences with year information (84.83%), followed by Portugal (14.75%) and Andorra (0.41%). There is only one record attributed to Gibraltar.

The taxonomic validation led to the deletion of 1,717 occurrences because of taxonomic issues (Fig. 1). Scientific names were unified in IberBryo v1.0 into 869 different species (including infraspecific taxa) from 57 families (857 out of 893 Spanish taxa, 369 out of 522 Portuguese taxa and 207 out of 274 Andorran taxa — totals extracted from Ros et al. 2013). Most of the species recorded in IberBryo are associated with rock and soil

substrates (see Suppl. material 3). The altitudinal range covered by records of IberBryo v1.1 is biased towards high altitude places compared to the study area [Two-sample Kolmogorov-Smirnov test D = 0.272, p < 0.001] (Fig. 2).

The historical pattern of moss surveys shows a steady increase in number of records and new species gathered through time. Due to the evaluation based on IberBryo v1.1 (only records with collecting date), the observed number of species (excluding infraspecific taxa) acummulated until 2018 was reduced to 745. The highest survey rates take place after 2000, and the accumulated number of observed species increased especially in the period 1970-1999 (Fig. 3). Records and number of species accumulated per year are strongly correlated through the whole time series (rho = 0.73, p < 0.001). Four distinct periods of collection — seemingly related to the political and overall academic situation of the Iberian countries — can be identified depending on changes in survey trends along the studied period: before 1935 (rho = 0.663, p < 0.001), 1936-1969 (rho = 0.256, p = 0.13), 1970-1999 (rho = 0.518, p = 0.003) and 2000-2018 (rho = -0.858, p < 0.001) (see Fig. 3).

Spatial coverage and survey completeness

The higher numbers of moss records, observed species richness and inventory completeness are mainly located in mountainous areas of the north and eastern Spanish coasts between 1970 and 1999 and in northern Portugal, central Spain and the mountainous area of Sierra Nevada between 2000 and 2018 (Fig. 5). Cells with very limited surveys or no information at all are located mainly in the inner plateaus and south-western Iberia, particularly after the year 2000 (Fig. 5). In the highly-surveyed period between 1970 and 2018, 4.98% of Iberian 30' resolution cells (14 out of 281 cells) meet the criteria needed to be considered well-surveyed (Fig. 4), while only 0.36% (9 out of 2441 cells) do so at 5' resolution (Suppl. material 6). An additional 8.9% of the 30' cells and 1.04% of the 5' cells were moderately-surveyed (25 and 26 cells, respectively). Considering the IberBryo v1.0 database, we find high levels in number of records, observed richness and completeness in north-eastern and north inner plateau of Spain with no information for the collecting date (Fig. 6). In addition, some of these cells present extremely high levels of survey completeness at 30' resolution (Suppl. material 8), highlighting the potential value of these data if records' information were complete.

Environmental coverage

The PCA identified the two main gradients that characterise the climate of the Iberian Peninsula: one axis mainly related to seasonality — that separates the Mediterranean from Atlantic zones; and another axis related to temperature and (to a less extent) precipitation variations — that describes a gradient from cold (northern-mountainous) to warm-dry zones (central-south-eastern Iberia) (Suppl. material 9). We used these two axes to define an environmental space of 51 climate types at 30' resolution (Fig. 7A), which captures 78% of the climate variability (Suppl. material 10). Well-surveyed cells cover 10 of these climate types (19.61%) (Fig. 7B), representing 21.75% of all climatic

variability in the Iberian Peninsula (Schoener's D = 0.218, p value = 0.002) (Suppl. material 9B). The coverage of the climatic variability occupied by well-surveyed moss cells is not biased in both axes when compared to the whole Iberian Peninsula: PC1 Two-sample Kolmogorov-Smirnov test D = 0.273, p = 0.272 and PC2 Two-sample Kolmogorov-Smirnov test D = 0.292, p = 0.203 (see also Suppl. material 9D). Well-surveyed cells also occupy more frequently 'rare climatic conditions' (Fig. 7C), but they show no differences compared to the distribution of climatic rarity in the Iberian Peninsula (Two-sample Kolmogorov-Smirnov test D = 0.176, p = 0.957) — which also present high levels of 'rare climatic conditions' (Fig. 7D). However, well-sampled areas provide a biased description of land-use changes across Iberia, as they are mostly placed in areas that have been changing towards higher proportions of artificial surfaces in the last decades, lacking data for cells that have followed naturalisation processes (Fig. 8C). Interestingly, the well-surveyed cells of northern Portugal are placed in the Iberian region with the highest rates of land use transformation (Fig. 8A).

Discussion

Our analysis of the publicly-available data on Iberian mosses evidences the large extent of the shortfalls of the distributional information for this group. Besides, our study proves the crucial importance of data (and metadata) quality for evaluating the Wallacean shortfall for mosses, in the same way as has been established previously for other groups (Hortal et al. 2007, Yang et al. 2014, Meyer et al. 2016, Stropp et al. 2016). The diversity patterns of European mosses have been scarcely studied, at least when compared to flowering plants (Mutke and Barthlott 2005, Mutke and Geffert 2010, Geffert et al. 2013, Berdugo et al. 2018). The results above exposed evidence that the knowledge of such a common group with a long history of surveys in the Iberian Peninsula is, overall, insufficient. These surveys provide poor coverage of the distribution of moss diversity in this highly-heterogeneous region, which makes the assessment of its assemblage responses to climatic and land-use variations a challenging task. In fact, our results reveal that surveys are biased towards the location of the most important bryophyte researchers' groups and mountainous areas. Issues on data quality, particularly the absence of information on collecting date, enlarge the existing biases even further. Despite these limitations, well-surveyed places are distributed throughout the whole study area. Indeed, they provide a fair (though limited) cover of about one fifth of the climate types of the Iberian Peninsula, which may allow using these data to model species and community responses to climate and assess the effects of climate change. Other aspects of global change are however less represented, because moss information is biased towards anthropised areas and some of the well-surveyed cells are located nearby an area that has suffered frequent land-use changes in the last decades.

The different biases, identified in moss biodiversity information, could compromise the reliability of eventual macroecological analysis carried out with the publicly-available data. Indeed, the main geographical pattern of observed species richness of Iberian mosses can be easily attributed to the recorders' home range (*sensu* Dennis and Thomas

2000; also known as taxonomist survey bias, Sastre and Lobo 2009). This is a common bias that has also been previously described for other groups (e.g. Lobo and Martin-Piera 2002, Oliveira et al. 2016, Girardello et al. 2019). In the case of Iberian mosses, well-surveyed areas and those with high density of records are placed near the bryologists' homes and working places, especially in northern Portugal and eastern Spain, along with some exceptions determined by the particular location of PhD works. The spatial pattern of surveys also follows the relatively-common bias towards mountains, which results in a distribution of records shifted towards high altitudes within IberBryo. Many Iberian moss survey hotspots are located in classical mountainous survey places, such as the Cantabrian and Sierra Nevada mountain ranges (see Suppl. material 6). Such preference of recorders for mountainous areas and natural reserves has been previously described for other taxa and may be related to the lower human impacts in these areas, their higher diversity due to their typically steeper environmental and habitat gradients or their general attraction for naturalists and the general public (see, for example, Lobo and Martin-Piera 2002, Yang et al. 2014, Meyer et al. 2015, Girardello et al. 2019). In contrast, we found noticeable gaps of information in the south-west of Iberia, the Ebro river basin and the inner plateaus, which should be considered for future moss surveys.

It is remarkable how much the absence of basic information aggravates the general limitations of our database. This evidences the necessity of gathering good quality data, as well as documenting metadata information properly. By an in-depth process of record verification and data-cleaning, we were able to improve the first versions of IberBryo, increasing the amount of data useful for the analysis. Despite these improvements, we found an important problem in the records' metadata. The absence of information on the collecting dates, that affected ca. 42% of the occurrences and prevented us from detecting duplicate records, limited significantly our assessment of inventory completeness (see Hortal et al. 2007). This problem especially affected a particular area of our study, Catalonia and Andorra and, to a much less extent, the northern inner plateau. Thus, we had to exclude one of the most surveyed zones of Iberia from all analyses with the temporal component, preventing any global change analysis that requires information on a key aspect, such as date (see Suppl. material 2). We also found inconsistent dates in some records of the Medina et al. (2015) catalogue during the data curation process. Some years of survey were incorrectly added to the catalogue, based on oral communication with B. Albertos and we needed to search for the original sampling years in the field notes. This implies that a revision of the vouchers and/or field notebooks by the recorders is fundamental to check the actual quality of the available information. These practices could also mobilise a massive amount of data and significantly increase the coverage of high-quality information provided by IberBryo.

Publicly-available Iberian moss records presented other common problems of biodiversity data related to georeferencing (Yesson et al. 2007, Yang et al. 2013, Maldonado et al. 2015, Meyer et al. 2016, Stropp et al. 2016). The absence of geographical coordinates affected ca. 30% of the occurrences in IberBryo v0.1 and the lack or inaccuracy of locality information led to discarding a substantial part. Fortunately, we were able to recover nearly half of these records through geocoding. This process is

not often considered in this kind of studies because it is thought to imply an unaffordable effort, but in our case, the improvement obtained was worth the time invested. We also detected taxonomic issues in the GBIF records, although to a lesser extent. These were related to the necessity of taxonomic standardisation of the data and the update of synonymies to currently-accepted names and misidentifications of wrong locations (as, for example, some American species are attributed to our study area). It is also important to mention the absence of substrate and/or habitat type information in many records, which implies the need to acquire it from external references. This prevents the assessment of eventual changes in substrate due to climatic variations, responses to land-use changes or any other ecological effect. The overall knowledge on the ecological responses of moss species would be highly beneficial if this information were added as part of the metadata of their records. The generality of these issues altogether evidences how simple and costless practices of collectors, such as digitising metadata information, could improve the public knowledge of a whole group of organisms, such as bryophytes.

The spatial coverage of Iberian moss surveys through time shows two distinct periods. On the one hand, records follow a patchy distributed pattern until 1970. The surveys showed a remarkable stop in the acquisition of new records between 1935 and 1969 – a setback attributable to the Spanish Civil War and the dictatorships suffered during this period in Spain and Portugal that has been previously described in other groups of organisms (Hortal et al. 2008). However, the overall surveys of the Iberian Peninsula identified many different species – ca. 450 – relatively early (before 1935), which is more than half of the total of species included in the current checklist of IberBryo. The second period shows a clear intensification in moss surveys since 1970, which increased their spatial extent to cover almost the entire Iberian Peninsula. Particularly, after the year 2000, our results show that surveys are concentrated in specific areas where bryophyte research has been more intense (see above), with a limitation in the extent of coverage in several regions that had been moderately well surveyed in the past. This pattern is common in many distributional information, where some well-surveyed areas remain biodiversity hotspots despite lacking recent surveys (see Meyer et al. 2015, Meyer et al. 2016, Stropp et al. 2016) and new surveys come from ecological studies concentrated in particular areas (see below), without following a geographically-stratified sampling design adequate for macroecological studies (Brown 1995, Funk et al. 2005, Hortal and Lobo 2005). However, the quality and usefulness of biodiversity information decays with time due to the unavoidable effects of taxonomic, land use and climatic changes, amongst others (Ladle and Hortal 2013, Tessarolo et al. 2017). This calls for establishing a series of re-surveys of classical localities, which would allow updating the information on these areas, as well as assessing eventual changes in the composition of bryophyte floras.

Interestingly, our findings on spatial coverage at two different cell resolutions allowed us to show that local surveys of mosses are not reflected at regional scale, so well-surveyed areas coincide only partially amongst resolutions (see Suppl. materials 4, 5). Actually, the correlation between survey effort and observed species richness is comparatively lower (0.68) in the 2000-2018 period at finer resolution (Suppl. material 7). Besides, the number of well-surveyed cells does not increase at the coarser spatial resolution,

reflecting that heterogeneous and incomplete local inventories could generate reliable regional species inventories under some circumstances. This result is opposite to Lobo et al. (2018) and La Sorte and Somveille (2020), who observed a close similarity of well-sampled areas at different resolutions for other groups. This is likely due to the effect of surveys orientated towards the ecological study of moss communities, where replicates of the same location and/or substrate are desirable (see, for example, Rams 2007, Cezón and Muñoz 2013, Medina et al. 2015, Hespanhol 2017). This is opposite to the typical floristic surveys of former decades, where interest was focused on inventorying as many species and localities as possible. This kind of information on survey trends is lost in higher scales and does not generate well-surveyed areas. In this sense, the assessment of collecting effort at different resolutions can be a good tool to understand the overall quality of the surveys (Oliveira et al. 2016).

Despite all the gaps and biases identified by our study, we find that Iberian climatic gradients — including the rarest climates — are fairly represented by the limited number of well-surveyed 30' cells, which just represent 5% of Iberia. That said, it is clear that it is highly desirable to enlarge the climatic coverage to improve the reliability of any species distribution model or similar approaches that are conducted with these data to assess the effects of climate change, invasions or other aspects of global change (Oliveira et al. 2016). The fact that well-surveyed cells are biased towards anthropised areas would not allow assessing macroecological effects of land-use intensification with fairness. This is despite the opportunity provided by the high density of recent surveys in northern Portugal, where well-surveyed areas are located in an area of intense land-use changes, mainly due to the wood-fires of the last decade. These novel results call for investigating whether these type of biases are general for other regions and biological groups. Additionally, updated information on comparable areas that have not suffered such transformations would be needed to provide a fair evaluation of the effects of this recent land transformation on moss communities, allowing us to assess the impact of global change on this group of organisms.

Final remarks and future insights

We show that the publicly-available information on Iberian mosses presents significant biases, related to the Wallacean shortfall, but also to basic knowledge on their ecology. This calls for a careful use of this information for biogeographical, ecological modelling and macroecological analysis. It could be argued that the over-representation of certain areas or environments caused by the spatial biases in the data is a relatively minor problem, if overall coverage of climatic and land-use gradients were good. However, opposite to the most intensely-sampled areas, we find noticeable spatial gaps in the information, particularly in the south-west of Iberia and the inner plateaus. The lack of information from these regions compromises any assessment of the processes behind species diversity patterns, as well as the implementation of conservation biogeography approaches (Reddy and Dávalos 2003, Lomolino 2004, Whittaker et al. 2005, Hortal et al. 2007, Hortal et al. 2008). Furthermore, the development of ecological, evolutionary and biogeographical research on Iberian mosses currently requires more surveys with an

adequate spatial design and planning (see Hortal and Lobo 2005, Medina et al. 2013). This would maximise their effectiveness, as exemplified by the results of one performed on Iberian epiphytic mosses (Medina et al. 2015). We, therefore, encourage Iberian bryologists to base their future surveys on the information of data gaps provided by the analysis of IberBryo. They could design their surveys using spatially-explicit tools that account for maximising the coverage of the steep environmental and global change that currently characterises the highly dynamic Iberian landscapes. Finally, the limitations associated with incomplete data and metadata could be easily sorted out with improved protocols for data gathering and processing. Further, we are aware that substantial herbarium information may still be waiting for digitalisation and it is not yet accessible through online databases. Beyond reducing the existing biases, enlarging current collections with records from places with poor knowledge outside of the traditionally-surveyed and attractive places will allow us to evaluate the effects of global change on moss communities, leading to both advance knowledge on the ecology and biogeography of Iberian mosses and making informed recommendations for their conservation.

Acknowledgements

We thank Priscila Lemes and Joaquín Calatayud for their constructive comments on the development of methods. CR was funded by the Comunidad de Madrid and the European Social Fund co-financed through the Youth Employment Operational Program and the Youth Employment Initiative (YEI) grant PEJ-2017-AI/AMB/6655. This work is part of the project UNITED Unifying niches, interactions and distributions: A common theoretical framework for geographic range dynamics and local coexistence (CGL2016-78070-P, funded by AEI/FEDER, UE).

Author contributions

CR and JH designed research, with FAM and NGM. CR gathered and processed all data, with VM and NGM. TS-S and BV provided novel R scripts. CR and FAM analysed the data, with aid from TS-S, BV, VM, NGM and JH. CR wrote the paper, with NGM and JH. All authors discussed results and approved the last version of the manuscript.

Conflicts of interest

References

- Albertos B (2001) Estudio biogeográfico de los briófitos epífitos del noroccidente peninsular. Tesis Doctoral. Universidad Autónoma de Madrid, Madrid.
- Alonso M, Jiménez JA, Nylinder S, Hedenäs L, Cano MJ (2016) Disentangling generic limits in *Chionoloma*, *Oxystegus*, *Pachyneuropsis* and *Pseudosymblepharis* (Bryophyta:

Pottiaceae): An inquiry into their phylogenetic relationships. Taxon 65 (1): 3-18. https://doi.org/10.12705/651.1
- Alonso M, Jiménez JA, Cano MJ (2018) New synonyms and typifications in *Chionoloma tenuirostre* (Pottiaceae, Bryophyta). Phytotaxa 373 (2): 147 -154. https://doi.org/10.11646/phytotaxa.373.2.5
- Amano T, Lamming JL, Sutherland W (2016) Spatial Gaps in Global Biodiversity Information and the Role of Citizen Science. BioScience 66 (5): 393-400. https://doi.org/10.1093/biosci/biw022
- Aranda S, Gabriel R, Borges PV, de Azevedo EB, Lobo J (2011) Designing a survey protocol to overcome the Wallacean shortfall: a working guide using bryophyte distribution data on Terceira Island (Azores). The Bryologist 114 (3): 611. https://doi.org/10.1639/0007-2745-114.3.611
- Aranda S, Hespanhol H, Homem N, Borges PV, Lobo J, Gabriel R (2015) The iterative process of plant species inventorying for obtaining reliable biodiversity patterns: The evaluation of samping performance. Botanical Journal of the Linnean Society 177 (4): 491-503. https://doi.org/10.1111/boj.12259
- Beck J, Ballesteros-Mejia L, Nagel P, Kitching I (2013) Online solutions and the 'Wallacean shortfall': what does GBIF contribute to our knowledge of species' ranges? Diversity and Distributions 19 (8): 1043-1050. https://doi.org/10.1111/ddi.12083
- Beck J, Böller M, Erhardt A, Schwanghart W (2014) Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. Ecological Informatics 19: 10-15. https://doi.org/10.1016/j.ecoinf.2013.11.002
- Berdugo M, Quant J, Wason J, Dovciak M (2018) Latitudinal patterns and environmental drivers of moss layer cover in extratropical forests. Global Ecology and Biogeography 27 (10): 1213-1224. https://doi.org/10.1111/geb.12778
- Bisby FA (2000) The Quiet Revolution: Biodiversity Informatics and the Internet. Science 289 (5488): 2309-2312. https://doi.org/10.1126/science.289.5488.2309
- Boakes E, McGowan PK, Fuller R, Chang-qing D, Clark N, O'Connor K, Mace G (2010) Distorted views of biodiversity: Spatial and temporal bias in species occurrence data. PLOS Biology 8 (6). https://doi.org/10.1371/journal.pbio.1000385
- Broennimann O, Fitzpatrick M, Pearman P, Petitpierre B, Pellissier L, Yoccoz N, Thuiller W, Fortin M, Randin C, Zimmermann N, Graham C, Guisan A (2012) Measuring ecological niche overlap from occurrence and spatial environmental data: Measuring niche overlap. Global Ecology and Biogeography 21 (4): 481-497. https://doi.org/10.1111/j.1466-8238.2011.00698.x
- Brown JH (1995) Macroecology. University of Chicago Press, Chicago, 284 pp.
- Brugués M, Cros RM, Guerra J (2007) Flora Briofítica Ibérica. Vol. I. Sphagnales, Andreaeales, Polytrichales, Tetraphidales, Buxbaumiales, Diphysciales. Universidad de Murcia, Sociedad Española de Briología, Murcia.
- Brugués M, Guerra J (2015) Flora Briofítica Ibérica. Vol. II. Archidiales, Dicranales, Fissidentales, Seligeriales, Grimmiales. Universidad de Murcia, Sociedad Española de Briología, Murcia.
- Calabrese B (2019) Data cleaning. In: Ranganathan S, Gribskov M, Nakai K, Schönbach C (Eds) Encyclopedia of bioinformatics and computational biology. Elsevier [ISBN 978-0-12-811432-2]. https://doi.org/10.1016/B978-0-12-809633-8.20458-5

- Callaghan D, Ashton P (2008) Bryophyte distribution and environment across an oceanic temperate landscape. Journal of Bryology 30 (1): 23-35. https://doi.org/10.1179/174328208X282148
- Casas C, Brugués M, Ros RM, Sérgio C, Barrón A, Filella I, Ruiz E, Perry AR (2006) Handbook of mosses of the Iberian Peninsula and the Balearic Islands: illustrated keys to genera and species. Institut d'Estudis Catalans, Barcelona.
- Cezón K, Muñoz J (2013) Catálogo de los musgos de Castilla-La Mancha (España). Boletín Sociedad Española de Briología 40-41: 15-41.
- Chapman AD (2005) Principles and methods of data cleaning – Primary species and species-occurrence data, version 1.0. Report for the Global Biodiversity Information Facility, Copenhagen.
- Clench H (1979) How to make regional fists of butterflies: Some thoughts. Journal of the Lepidopterists' Society 33: 216-231.
- Dennis RLH, Thomas CD (2000) Bias in butterfly distribution maps: The influence of hot spots and recorder's home range. Journal of Insect Conservation 4: 73-77. https://doi.org/10.1023/A:1009690919835
- Devictor V, Bensaude-Vincent B (2016) From ecological records to big data: the invention of global biodiversity. History and Philosophy of the Life Sciences 38 (4). https://doi.org/10.1007/s40656-016-0113-2
- Dierssen K (2001) Distribution, ecological amplitude and phytosociological characterization of European bryophytes. Bryophytorum Bibliotheca 56: 1-283.
- Engemann K, Enquist B, Sandel B, Boyle B, Jørgensen P, Morueta-Holme N, Peet R, Violle C, Svenning J (2015) Limited sampling hampers “big data” estimation of species richness in a tropical biodiversity hotspot. Ecology and Evolution 5 (3): 807-820. https://doi.org/10.1002/ece3.1405
- European Environmental Agency (2018) Corine Land Cover (CLC) 1992, 2000, 2006, 2012 and 2018 seamless vector data (Version 20). https://land.copernicus.eu/pan-european/corine-land-cover. Accessed on: 2018-9-01.
- Fick S, Hijmans R (2017) WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas: New climate surfaces for global land areas. *International Journal of Climatology* 37 (12): 4302-4315. https://doi.org/10.1002/joc.5086
- Franklin J, Serra-Diaz J, Syphard A, Regan H (2017) Big data for forecasting the impacts of global change on plant communities: Big data for forecasting vegetation dynamics. Global Ecology and Biogeography 26 (1): 6-17. https://doi.org/10.1111/geb.12501
- Funk VA, Richardson KS, Ferrier S (2005) Survey-gap analysis in expeditionary research: where do we go from here? Biological Journal of the Linnean Society 85 (4): 549-567. https://doi.org/10.1111/j.1095-8312.2005.00520.x
- Gandomi A, Haider M (2015) Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management 35 (2): 137-144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007
- García-Roselló E, Guisande C, Manjarrés-Hernández A, González-Dacosta J, Heine J, Pelayo-Villamil P, González-Vilas L, Vari R, Vaamonde A, Granado-Lorencio C, Lobo J (2015) Can we derive macroecological patterns from primary Global Biodiversity Information Facility data?: Macroecological patterns and GBIF data. Global Ecology and Biogeography 24 (3): 335-347. https://doi.org/10.1111/geb.12260

- Garilleti R, Albertos B (2012) ABrA: Atlas y libro rojo de los briófitos amenazados de España. Organismo Autónomo Parques Nacionales, Madrid, 288 pp. [ISBN 978-84-8014-836-8]
- GBIF (2018a) GBIF.org (8th August 2018) GBIF Occurrence Download. URL: https://doi.org/10.15468/dl.eujakg
- GBIF (2018b) GBIF.org (9th October 2018) GBIF Occurrence Download. URL: https://doi.org/10.15468/dl.ogvrsc
- Geffert JL, Frahm J, Barthlott W, Mutke J (2013) Global moss diversity: spatial and taxonomic patterns of species richness. Journal of Bryology 35 (1): 1-11. https://doi.org/10.1179/1743282012Y.0000000038
- Girardello M, Chapman A, Dennis R, Kaila L, Borges PV, Santangeli A (2019) Gaps in butterfly inventory data: A global analysis. Biological Conservation 236: 289-295. https://doi.org/10.1016/j.biocon.2019.05.053
- Global Administrative Areas (2018) GADM database of Global Administrative Areas, version 3.4. https://gadm.org/data.html. Accessed on: 2018-9-01.
- Goffinet B, Shaw AJ, Cox CJ, Wickett NJ, Boles S (2004) Phylogenetic inferences in the Orthotrichoideae (*Orthotrichaceae*: *Bryophyta*) based on variation in four loci from all genomes. Monographs in Systematic Botany from the Missouri Botanical Garden 98: 270-289.
- Guerra J, Cano MJ, Ros RM (2006) Flora Briofítica Ibérica. Vol. III. Pottiales, Encalyptales. Universidad de Murcia, Sociedad Española de Briología, Murcia.
- Guerra J, Brugués M, Cano MJ, Cros RM (2010) Flora Briofítica Ibérica. Vol. IV. Funariales, Splachnales, Schistostegales, Bryales, Timmiales. Universidad de Murcia, Sociedad Española de Briología, Murcia.
- Guerra J, Cano MJ, Brugués M (2014) Flora Briofítica Ibérica. Vol. V. Orthotrichales, Hedwigiales, Leucodontales, Hookeriales. Universidad de Murcia, Sociedad Española de Briología, Murcia.
- Guerra J, Cano MJ, Brugués M (2018) Flora Briofítica Ibérica. Vol. VI. Hypnales. Universidad de Murcia, Sociedad Española de Briología, Murcia.
- Gueta T, Carmel Y (2016) Quantifying the value of user-level data cleaning for big data: A case study using mammal distribution models. Ecological Informatics 34: 139-145. https://doi.org/10.1016/j.ecoinf.2016.06.001
- Guisande C, Heine J, González-DaCosta J, García-Roselló E (2014) RWizard Software. Universidad de Vigo. Vigo, Spain. URL: http://www.ipez.es/RWizard/
- Hampton SE, Strasser CA, Tewksbury JJ, Gram WK, Budden AE, Batcheller AL, Duke CS, Porter JH (2013) Big data and the future of ecology. Frontiers in Ecology and the Environment 11 (3): 156-162. https://doi.org/10.1890/120103
- Hespanhol H, Cezón K, Felicísimo Á, Muñoz J, Mateo R (2015) How to describe species richness patterns for bryophyte conservation? Ecology and Evolution 5 (23): 5443-5455. https://doi.org/10.1002/ece3.1796
- Hespanhol H (2017) Bryophyte collection of Porto Herbarium (PO). 2.1. Natural History and Science Museum of the University of Porto (MHNC-UP). Occurrence dataset https://doi.org/10.15468/j0ks2f.
- Hill MO, Bell N, Bruggeman-Nannenga MA, Brugués M, Cano MJ, Enroth J, Flatberg KI, Frahm JP, Gallego MT, Garilleti R, Guerra J, Hedenäs L, Holyoak DT, Hyvönen, Ignatov MS, Lara F, Mazimpaka V, Muñoz J, Söderström L (2006) An annotated checklist of the

mosses of Europe and Macaronesia. Journal of Bryology 28 (3): 198-267. https://doi.org/10.1179/174328206X119998
- Hill MO (2007) BRYOATT: attributes of British and Irish mosses, liverworts and hornworts. Centre for Ecology and Hydrology, Huntingdon, Cambridgeshire.
- Hodgetts NG (2015) Checklist and country status of European bryophytes – towards a new Red List for Europe. Irish Wildlife Manuals, No. 84. National Parks and Wildlife Service, Department of Arts, Heritage and the Gaeltacht, Ireland.
- Hortal J, Lobo J (2005) An ED-based protocol for optimal sampling of biodiversity. Biodiversity and Conservation 14 (12): 2913-2947. https://doi.org/10.1007/s10531-004-0224-z
- Hortal J, Lobo J, Jiménez-Valverde A (2007) Limitations of biodiversity databases: Case study on seed-plant diversity in Tenerife, Canary Islands. Conservation Biology 21 (3): 853-863. https://doi.org/10.1111/j.1523-1739.2007.00686.x
- Hortal J, Jiménez-Valverde A, Gómez J, Lobo J, Baselga A (2008) Historical bias in biodiversity inventories affects the observed environmental niche of the species. *Oikos* 117 (6): 847-858. https://doi.org/10.1111/j.0030-1299.2008.16434.x
- Hortal J, de Bello F, Diniz-Filho J, Lewinsohn T, Lobo J, Ladle R (2015) Seven shortfalls that beset large-scale knowledge of biodiversity. Annual Review of Ecology, Evolution, and Systematics 46 (1): 523-549. https://doi.org/10.1146/annurev-ecolsys-112414-054400
- Isaac NB, Pocock MO (2015) Bias and information in biological records: Bias and information in biological records. Biological Journal of the Linnean Society 115 (3): 522-531. https://doi.org/10.1111/bij.12532
- Ladle R, Hortal J (2013) Mapping species distributions: living with uncertainty. Frontiers of Biogeography 5 (1). https://doi.org/10.21425/F5FBG12942
- Lara F, Garilleti R, Goffinet B, Draper I, Medina R, Vigalondo B, Mazimpaka V (2016) *Lewinskya*, a new genus to accommodate the phaneroporous and monoicous taxa of *Orthotrichum* (*Bryophyta*, *Orthotrichaceae*). Cryptogamie, Bryologie 37: 361-382. https://doi.org/10.7872/cryb/v37.iss4.2016.361
- La Sorte F, Somveille M (2020) Survey completeness of a global citizen-science database of bird occurrence. Ecography 43 (1): 34-43. https://doi.org/10.1111%2Fecog.04632
- Lobo J, Martin-Piera F (2002) Searching for a Predictive Model for Species Richness of Iberian Dung Beetle Based on Spatial and Environmental Variables. Conservation Biology 16 (1): 158-173. https://doi.org/10.1046/j.1523-1739.2002.00211.x
- Lobo J, Hortal J, Yela JL, Millán A, Sánchez-Fernández D, García-Roselló E, González-Dacosta J, Heine J, González-Vilas L, 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. https://doi.org/10.1016/j.ecolind.2018.03.077
- Lomolino MV (2004) Conservation biogeography. In: Lomolino MV, Heaney LR (Eds) Frontiers of biogeography: New directions in the geography of nature. Sunderland, MA: Sinauer, 293-296 pp.
- Magill RE (2010) Moss diversity: New look at old numbers. Phytotaxa 9 (1): 167. https://doi.org/10.11646/phytotaxa.9.1.9
- Maldonado C, Molina C, Zizka A, Persson C, Taylor C, Albán J, Chilquillo E, Rønsted N, Antonelli A (2015) Estimating species diversity and distribution in the era of Big Data: to what extent can we trust public databases?: Species diversity and distribution in the era

of Big Data. Global Ecology and Biogeography 24 (8): 973-984. https://doi.org/10.1111/geb.12326
- Mateo R, Broennimann O, Normand S, Petitpierre B, Araújo M, Svenning J, Baselga A, Fernández-González F, Gómez-Rubio V, Muñoz J, Suarez G, Luoto M, Guisan A, Vanderpoorten A (2016) The mossy north: an inverse latitudinal diversity gradient in European bryophytes. Scientific Reports 6 (1). https://doi.org/10.1038/srep25546
- Mazimpaka V, Lara F (2014) *Codonoblepharon*. In: Guerra J, Cano MJ, Brugués M (Eds) Flora Briofítica Ibérica. Vol. V. Universidad de Murcia, Sociedad Española de Briología, Murcia, 27-30 pp.
- Medina NG, Lara F, Mazimpaka V, Hortal J (2013) Designing bryophyte surveys for an optimal coverage of diversity gradients. Biodiversity and Conservation 22 (13-14): 3121-3139. https://doi.org/10.1007/s10531-013-0574-5
- Medina NG, Mazimpaka V, Hortal J, Lara F (2015) Epiphytic bryophytes of *Quercus* forests in Central and North inland Iberian Peninsula. Frontiers of Biogeography 7: 21-28.
- Medina NG, Ronquillo C (2020) Epiphytic mosses from the northwest Iberian quadrant (Spain). Spanish National Museum of Natural Sciences (CSIC). Occurrence dataset. GBIF.org. URL: https://doi.org/10.15470/rqv6jb
- Meyer C, Kreft H, Guralnick R, Jetz W (2015) Global priorities for an effective information basis of biodiversity distributions. Nature Communications 6 (1). https://doi.org/10.1038/ncomms9221
- Meyer C, Weigelt P, Kreft H (2016) Multidimensional biases, gaps and uncertainties in global plant occurrence information. Ecology Letters 19 (8): 992-1006. https://doi.org/10.1111/ele.12624
- Morueta-Holme N, Svenning J (2018) Geography of plants in the New World: Humboldt's relevance in the age of big data. Annals of the Missouri Botanical Garden 103 (3): 315-329. https://doi.org/10.3417/2018110
- Müller F (2014) *Bartramia aprica*— the correct name for the Mediterranean and Western North American species historically recognized as *"Bartramia stricta*". Herzogia 27 (1): 211-214. https://doi.org/10.13158/heia.27.1.2014.211
- Mutke J, Barthlott W (2005) Patterns of vascular plant diversity at continental to global scale. Biologiske Skrifter 55: 521-537.
- Mutke J, Geffert JL (2010) Keep on working: the uneven documentation of regional moss floras. Bryophyte Diversity and Evolution 31 (1): 7. https://doi.org/10.11646/bde.31.1.5
- Oliveira U, Paglia AP, Brescovit A, de Carvalho CB, Silva DP, Rezende D, Leite FSF, Batista JAN, Barbosa JPPP, Stehmann JR, Ascher J, de Vasconcelos MF, De Marco P, Löwenberg-Neto P, Dias PG, Ferro VG, Santos A (2016) The strong influence of collection bias on biodiversity knowledge shortfalls of Brazilian terrestrial biodiversity. Diversity and Distributions 22 (12): 1232-1244. https://doi.org/10.1111/ddi.12489
- Plášek V, Sawicki J, Ochyra R, Szczecińska M, Kulik T (2015) New taxonomical arrangement of the traditionally conceived genera *Orthotrichum* and *Ulota* (Orthotrichaceae, Bryophyta). Acta Musei Silesiae, Scientiae Naturales 64 (2): 169-174. https://doi.org/10.1515/cszma-2015-0024
- QGIS Development Team (2019) QGIS Geographic Information System. Open Source Geospatial Foundation Project . v 3.4 Madeira. URL: http://qgis.osgeo.org/
- Rams S (2007) Estudios briológicos sobre flora, vegetación, taxonomía y conservación en Sierra Nevada (Andalucía, S de España). Universidad de Murcia, Murcia. URL: http://hdl.handle.net/10201/33265

- R Development Core Team (2019) R: a language and environment for statistical computing. v 3.6. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.Rproject.org
- Reddy S, Dávalos L (2003) Geographical sampling bias and its implications for conservation priorities in Africa. Journal of Biogeography 30 (11): 1719-1727. https://doi.org/10.1046/j.1365-2699.2003.00946.x
- Rocchini D, Hortal J, Lengyel S, Lobo J, Jiménez-Valverde A, Ricotta C, Bacaro G, Chiarucci A (2011) Accounting for uncertainty when mapping species distributions: The need for maps of ignorance. Progress in Physical Geography: Earth and Environment 35 (2): 211-226. https://doi.org/10.1177/0309133311399491
- Ronquillo C, Hortal J (2020) IberBryo - iberian mosses occurrences dataset. Version 1.0. DIGITAL-CSIC. Release date: 2020-3-18. URL: http://hdl.handle.net/10261/204405
- Ros R, Mazimpaka V, Abou-Salama U, Aleffi M, Blockeel T, Brugués M, Cros RM, Dia MG, Dirkse G, Draper I, El-Saadawi W, Erdağ A, Ganeva A, Gabriel R, González-Mancebo J, Granger C, Herrnstadt I, Hugonnot V, Khalil K, Kürschner H, Losada-Lima A, Luís L, Mifsud S, Privitera M, Puglisi M, Sabovljević M, Sérgio C, Shabbara H, Sim-Sim M, Sotiaux A, Tacchi R, Vanderpoorten A, Werner O (2013) Mosses of the Mediterranean, an annotated checklist. Cryptogamie, Bryologie 34 (2): 99. https://doi.org/10.7872/cryb.v34.iss2.2013.99
- RStudio Team (2019) RStudio: Integrated Development Environment for R. RStudio, PBC, Boston, MA. v 1.2. URL: http://www.rstudio.com/.
- Saarenmaa H, Nielsen ES (2002) Towards a global biological information infrastructure. Challenges, opportunities, synergies, and the role of entomology. 70. European Environment Agency, Copenhagen, 72 pp.
- Sastre P, Lobo J (2009) Taxonomist survey biases and the unveiling of biodiversity patterns. Biological Conservation 142 (2): 462-467. https://doi.org/10.1016/j.biocon.2008.11.002
- Schoener A (1974) Colonization curves for Planar Marine Islands. Ecology 55 (4): 818-827. https://doi.org/10.2307/1934417
- Soberón J, Llorente J (1993) The use of species accumulation functions for the prediction of species richness. Conservation Biology 7 (3): 480-488. https://doi.org/10.1046/j.1523-1739.1993.07030480.x
- Soberón J, Peterson T (2004) Biodiversity informatics: managing and applying primary biodiversity data. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences 359 (1444): 689-698. https://doi.org/10.1098/rstb.2003.1439
- Sotiaux A, Vanderpoorten A (2017) A checklist of the bryophytes of Andorra. Journal of Bryology 39 (4): 353-367. https://doi.org/10.1080/03736687.2017.1346744
- Sousa-Baena MS, Garcia LC, Peterson AT (2014) Completeness of digital accessible knowledge of the plants of Brazil and priorities for survey and inventory. Diversity and Distributions 20 (4): 369-381. https://doi.org/10.1111/ddi.12136
- Stropp J, Ladle R, Malhado A, Hortal J, Gaffuri J, Temperley W, Olav Skøien J, Mayaux P (2016) Mapping ignorance: 300 years of collecting flowering plants in Africa: 300 Years of collecting flowering plants in Africa. Global Ecology and Biogeography 25 (9): 1085-1096. https://doi.org/10.1111/geb.12468
- Tessarolo G, Ladle R, Rangel T, Hortal J (2017) Temporal degradation of data limits biodiversity research. Ecology and Evolution 7 (17): 6863-6870. https://doi.org/10.1002/ece3.3259

- Troia M, McManamay R (2016) Filling in the GAPS: evaluating completeness and coverage of open-access biodiversity databases in the United States. Ecology and Evolution 6 (14): 4654-4669. https://doi.org/10.1002/ece3.2225
- U.S. Geological Survey (2010) Global Multi-resolution Terrain Elevation Data 2010 (GMTED2010). https://topotools.cr.usgs.gov/gmted_viewer/viewer.htm. Accessed on: 2018-10-09.
- Wetzel F, Bingham H, Groom Q, Haase P, Kõljalg U, Kuhlmann M, Martin C, Penev L, Robertson T, Saarenmaa H, Schmeller D, Stoll S, Tonkin J, Häuser C (2018) Unlocking biodiversity data: Prioritization and filling the gaps in biodiversity observation data in Europe. Biological Conservation 221: 78-85. https://doi.org/10.1016/j.biocon.2017.12.024
- Whittaker R, Araújo M, Jepson P, Ladle R, Watson JM, Willis K (2005) Conservation Biogeography: assessment and prospect. Diversity and Distributions 11 (1): 3-23. https://doi.org/10.1111/j.1366-9516.2005.00143.x
- Wieczorek J, Guo Q, Hijmans R (2004) The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. International Journal of Geographical Information Science 18 (8): 745-767. https://doi.org/10.1080/13658810412331280211
- Wüest R, Zimmermann N, Zurell D, Alexander J, Fritz S, Hof C, Kreft H, Normand S, Cabral JS, Szekely E, Thuiller W, Wikelski M, Karger DN (2019) Macroecology in the age of Big Data – Where to go from here? Journal of Biogeography 47 (1): 1-12. https://doi.org/10.1111/jbi.13633
- 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 (8): 1415-1426. https://doi.org/10.1111/jbi.12108
- Yang W, Ma K, Kreft H (2014) Environmental and socio-economic factors shaping the geography of floristic collections in China: Geography of floristic collections in China. Global Ecology and Biogeography 23 (11): 1284-1292. https://doi.org/10.1111/geb.12225
- Yesson C, Brewer P, Sutton T, Caithness N, Pahwa J, Burgess M, Gray WA, White R, Jones A, Bisby F, Culham A (2007) How global Is the Global Biodiversity Information Facility? PLoS ONE 2 (11). https://doi.org/10.1371/journal.pone.0001124

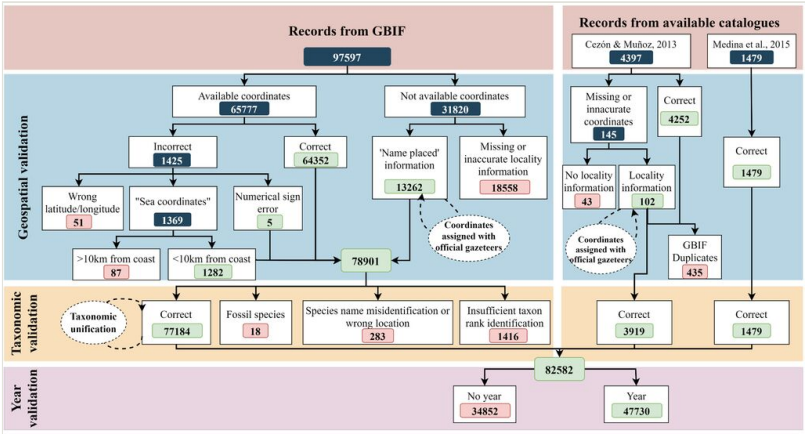

Figure 1.

Pre-processing steps in the generation of IberBryo database and numbers of records managed in each one. Green numbers correspond to validated records and red numbers to deleted ones.

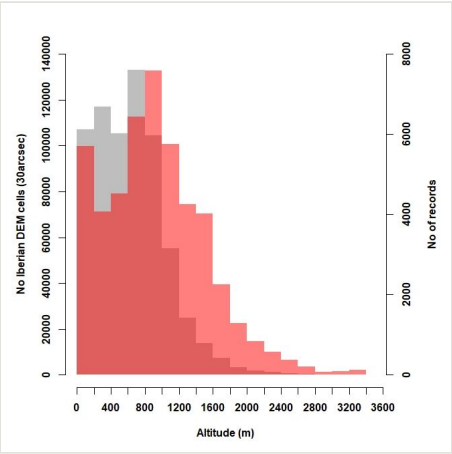

Figure 2.

Altitudinal coverage of moss surveys, as the comparison between the altitudinal distributions of IberBryo v1.1 records (red bars) and the whole surface of the Iberian Peninsula (grey bars).

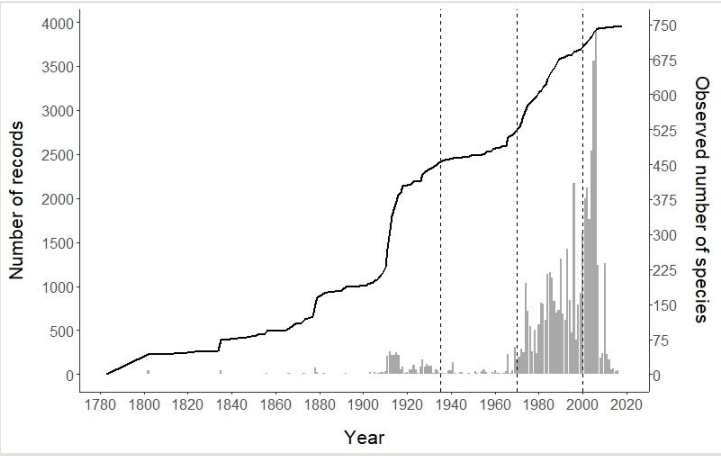

Figure 3.

Historical progression of moss surveys in the Iberian Peninsula. Number of moss records gathered each year (grey bars) and accumulated number of species recorded in IberBryo (black line). Vertical dashed lines define different periods of historical data collection.

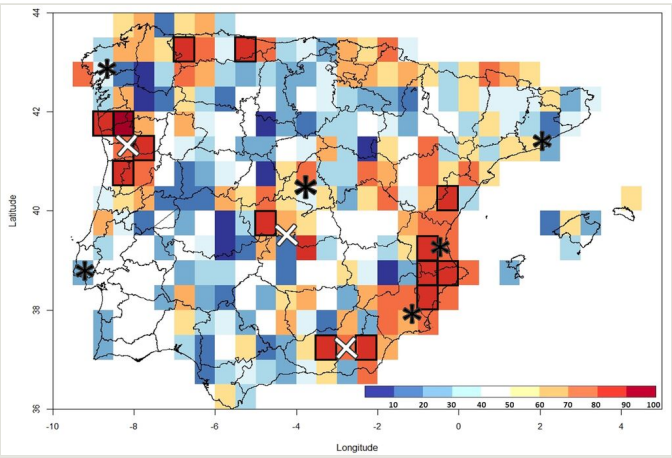

Figure 4.

Geographic distribution of inventory completeness in the 1970-2018 period at 30' resolution, according to the IberBryo v1.1 database. Values close to red represent higher percentages of completeness. Black squares correspond to well-surveyed cells (completeness ≥ 80% and number of records ≥ 10), white X-crosses to PhD theses – from left to right: Helena Hespanhol (NW Portugal), Katia Cezón (Castilla-La Mancha) and Susana Rams (Sierra Nevada) and black asterisks to major Iberian bryologist groups. These main research centres on bryophytes correspond to: Universidad Autónoma de Barcelona, Universidad Autónoma de Madrid, Universidad Complutense de Madrid, Universidade de Lisboa, Universidad de Murcia, Universidad Rey Juan Carlos, Universidade de Santiago de Compostela, Universitat de València, Museo Nacional de Ciencias Naturales (MNCN-CSIC) and Real Jardín Botánico (RJB-CSIC).

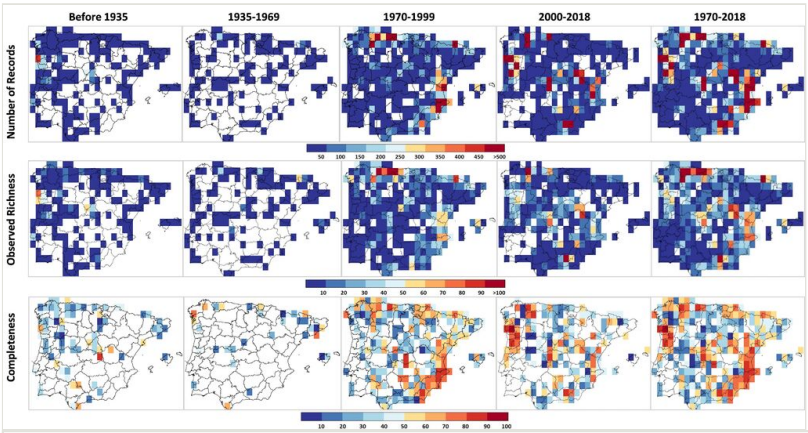

Figure 5.

Geographical coverage of moss surveys along time in the Iberian Peninsula. Maps show the distribution of records numbers, observed richness and inventory completeness of Iberian mosses in each period at 30' resolution in IberBryo v1.1.

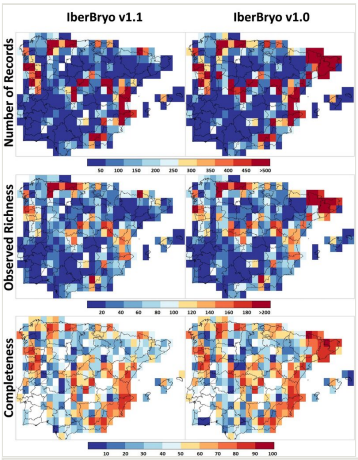

Figure 6.

Geographical coverage of moss surveys as number of records, observed richness and inventory completeness included in IberBryo v1.1 database (with information on collecting date at year level; 1783-2018) and in IberBryo v1.0 database (including records without information on collecting date) at 30' resolution.

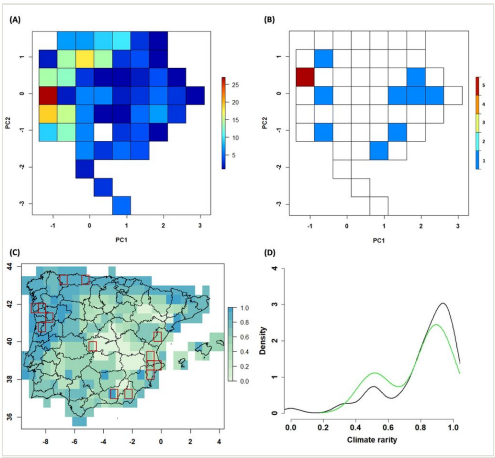

Figure 7.

Climatic coverage of Iberian moss surveys. (A) Frequency of climate types in the Iberian environmental space (values indicate the number of 30' cells of each climate type). (B) Frequency of climate types covered by well-surveyed cells (values indicate the number of 30' cells of each climate type). (C) Geographic distribution of climatic rarity index in the study area (rarest climate types = 1), red squares indicate the location of well-surveyed moss cells. (D) Density comparison of the climatic rarity covered by Iberian cells (black line) and well-surveyed moss cells (green line).

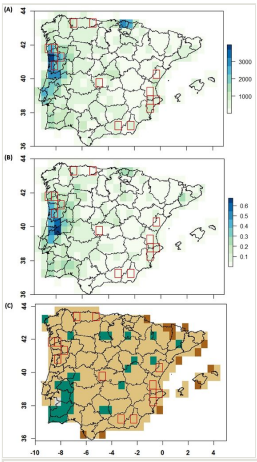

Figure 8.

(A) Geographical distribution of frequency in land-use changes in 1990-2018 at 30' resolution cells. (B) Proportion of land-use changed area in 1990-2018 at 30' resolution cells. (C) Geographical distribution of 'anthropised change ratio' as artificial surfaces [A] or natural surfaces [N] changes. Dark brown cells 'Anthropised only' N to A; Light brown cells 'Mostly anthropised' N to A > A to N; Grey cells 'Equally changed' N to A = A to N; Green cells 'Naturalised' N to A < A to N. Red squares indicate the location of well-surveyed moss cells.

Supplementary materials

Suppl. material 1: IberBryo Database v1.0

Authors: C. Ronquillo, V. Mazimpaka & J. Hortal
Data type: Occurrences
Brief description: IberBryo database (.txt format; UTF-8 encoding)
Also available in: Ronquillo, Cristina; Hortal, Joaquín; 2020; "IberBryo - iberian mosses occurrences dataset"; DIGITAL-CSIC; Version 1.0; http://dx.doi.org/10.20350/digitalCSIC/12494 (This excel version includes fields' descriptions).
Download file (3.41 MB)

Suppl. material 2: Distribution Maps of Iberian Moss Occurrences

Authors: C. Ronquillo
Data type: Map
Brief description: (A) IberBryo v1.1 occurrences (47,730), (B) Preprocessed occurrences without collecting date (34,852) (C) Occurrences from GBIF before data-cleaning and validation process (33,382).
Download file (2.41 MB)

Suppl. material 3: Checklist of species included in Iberbryo v1.0 and their frequency in each class of substrate.

Authors: C. Ronquillo & N. G. Medina
Data type: Table
Brief description: Frequency of used substrate [1] Rare substrate [2] Occasional substrate [3] Normal substrate.
Download file (5.40 MB)

Suppl. material 4: Spatial coverage at 5' resolution. Plates show the number of records in different periods, for the complete time series (IberBryo v1.1) and including records without information on the collecting date (IberBryo v1.0).

Authors: C. Ronquillo & J. Hortal
Data type: Map
Download file (2.41 MB)

Suppl. material 5: Spatial coverage of IberBryo v1.1 at 5' resolution. Plates show the observed richness in different periods, for the complete time series (IberBryo v1.1) and including records without information on the collecting date (IberBryo v1.0).

Authors: C. Ronquillo & J. Hortal
Data type: Map
Download file (2.85 MB)

Suppl. material 6: Spatial coverage of IberBryo v1.1 at 5' resolution. Plates show the inventory completeness in different periods, for the complete time series (IberBryo v1.1) and including records without information on the collecting date (IberBryo v1.0).

Authors: C. Ronquillo & J. Hortal
Data type: Map
Download file (2.38 MB)

Suppl. material 7: Correlations between records and observed richness per cell.

Authors: C. Ronquillo
Data type: Table
Download file (12.60 kb)

Suppl. material 8: Grid cells classified as 'survey hotspots' at 30' resolution.

Authors: C. Ronquillo
Data type: Table
Download file (12.77 kb)

Suppl. material 9: Climatic coverage PCA analysis

Authors: C. Ronquillo, F. Alves-Martins & J. Hortal
Data type: Figure
Brief description: (A) Distribution of Worldclim 2.0 biovariables at 30' resolution along the space described by the two climatic axes identified by a PCA. (B) Distribution of Schoener's D of climatic variability in our study area (grey bars). The dashed red line indicates the Schoener's D overlap value of well-sampled mosses sites. (C) Geographical distribution of PCA axes scores in the Iberian Peninsula. Colour gradients represent the values of each cell in the corresponding axis, ranging from the most negative (white) to the most positive (green) scores (see the corresponding scale bars). (D) Comparison between the density of PCA scores of the Iberian Peninsula (black line) and the well-surveyed bryophyte cells (red line) for each PCA axis.
Download file (1.30 MB)

Suppl. material 10: Results of the PCA of climatic variables based on WorldClim 2.0 biovariables at 30' resolution.

Authors: C. Ronquillo, F. Alves-Martins & J. Hortal
Data type: Table
Download file (14.79 kb)

Suppl. material 11: Reclassifications of land-use categories of CORINE classes used in this work.

Authors: C. Ronquillo, F. Alves-Martins & J. Hortal
Data type: Table

Brief description: Reclassification 1 corresponds to aggregated classes of CORINE according to the importance of bryophyte natural history. Reclassification 2 corresponds to whether each type of land-use is (arguably) of artificial or natural origin.
Download file (14.20 kb)

Suppl. material 12: IberBryo Database Protocol

Authors: C. Ronquillo
Data type: Text
Brief description: Detailed process of IberBryo creation
Download file (249.28 kb)

Suppl. material 13: Coverage analysis R scripts

Authors: C. Ronquillo, F. Alves-Martins, T. Sobral-Souza, B. Vilela-Silva
Data type: Scripts
Brief description: The folder contains 3 R scripts used in this work.: 'Climatic coverage analysis' , 'Land use coverage analysis' and 'Temporal and Spatial coverage analysis'
Download file (10.95 kb)