


RESEARCH

Open Access



# Big data of tree species distributions: how big and how good?

Josep M. Serra-Diaz<sup>1,4\*</sup> , Brian J. Enquist<sup>2</sup>, Brian Maitner<sup>2</sup>, Cory Merow<sup>3</sup> and Jens-C. Svenning<sup>1,4</sup>

## Abstract

**Background:** Trees play crucial roles in the biosphere and societies worldwide, with a total of 60,065 tree species currently identified. Increasingly, a large amount of data on tree species occurrences is being generated worldwide: from inventories to pressed plants. While many of these data are currently available in big databases, several challenges hamper their use, notably geolocation problems and taxonomic uncertainty. Further, we lack a complete picture of the data coverage and quality assessment for open/public databases of tree occurrences.

**Methods:** We combined data from five major aggregators of occurrence data (e.g. Global Biodiversity Information Facility, Botanical Information and Ecological Network v.3, DRYFLOR, RAINBIO and Atlas of Living Australia) by creating a workflow to integrate, assess and control data quality of tree species occurrences for species distribution modeling. We further assessed the coverage – the extent of geographical data – of five economically important tree families (Arecaceae, Dipterocarpaceae, Fagaceae, Myrtaceae, Pinaceae).

**Results:** Globally, we identified 49,206 tree species (84.69% of total tree species pool) with occurrence records. The total number of occurrence records was 36.69 M, among which 6.40 M could be considered high quality records for species distribution modeling. The results show that Europe, North America and Australia have a considerable spatial coverage of tree occurrence data. Conversely, key biodiverse regions such as South-East Asia and central Africa and parts of the Amazon are still characterized by geographical open-public data gaps. Such gaps are also found even for economically important families of trees, although their overall ranges are covered. Only 15,140 species (26.05%) had at least 20 records of high quality.

**Conclusions:** Our geographical coverage analysis shows that a wealth of easily accessible data exist on tree species occurrences worldwide, but regional gaps and coordinate errors are abundant. Thus, assessment of tree distributions will need accurate occurrence quality control protocols and key collaborations and data aggregation, especially from national forest inventory programs, to improve the current publicly available data.

**Keywords:** Tree distributions, Big data, Quality control and assessment, Occurrence data

## Background

Monitoring and understanding the distribution of tree species in the world is a major research agenda in plant ecology, especially under ongoing rapid global change (Enquist et al. 2016; Franklin et al. 2016). Estimating tree species distributions is ultimately needed in order to provide better understanding of tree diversity, a key driver

of forest functioning (Paquette and Messier 2011; Pichancourt et al. 2014) and forest ecosystem service provisioning (Gamfeldt et al. 2013; Thompson et al. 2014). Species distributions constitute, in addition, basic information required in systematic conservation planning and forecast range dynamics under future climate change (Franklin 2010; Serra-Diaz et al. 2012; Guisan et al. 2013; Zhang et al. 2017). The extent to which populations will respond to climate change is thought to depend upon variation in geographic distribution, phenotypic variation, response to CO<sub>2</sub> fertilization, strength of selection, fecundity, and degree of biotic interactions (Franklin et al. 2016). For example, certain geographic locations such as

\* Correspondence: pep.serra.diaz@bios.au.dk

<sup>1</sup>Section for Ecoinformatics and Biodiversity, Department of Bioscience, Aarhus University, Ny Munkegade 114, DK-8000 Aarhus C, Denmark

<sup>4</sup>Center for Biodiversity Dynamics in a Changing World (BIOCHANGE), Department of Bioscience, Aarhus University, Ny Munkegade 114, DK-8000 Aarhus, Denmark

Full list of author information is available at the end of the article

the Amazon basin, western United States, boreal forests, southern Europe, and Australia appear to be more susceptible to forest loss due to region specific heightened changes in temperature and precipitation (Allen et al. 2010; Choat et al. 2012). Further, widespread species with large populations and high fecundity are more likely to persist and adapt, but species with small populations, fragmented ranges, or low fecundity, may be more at risk for population collapse and may be candidates for facilitated migration. We currently lack a general assessment of global tree distributions, which limits our understanding of forest function, may dangerously over-simplify climate change projections, and hampers effective conservation planning under the ongoing planetary anthropogenic change.

In recent years biodiversity informatics research has rapidly advanced (Graham et al. 2004; Botkin et al. 2007) and forest ecology has swiftly benefited from it (Ash et al. 2017; Zhang et al. 2017). Currently, regional forest monitoring plans are being implemented worldwide, scientific collaboration networks are being developed and large-scale data infrastructures and tools for big-data biodiversity research and ecological studies have been developed (Franklin et al. 2017). While the future of forest diversity research and monitoring is promising, urgent efforts are still needed to properly document worldwide tree distributions and diversity in order to assess the relationship between species occurrence and climate and set up conservation plans. First, data are sparse or absent in various regions (Sousa-Baena et al. 2014; Meyer et al. 2016). Second, public data availability – a key aspect of reproducible science (Hampton et al. 2015) – is often limited as forests may represent key national resources with huge economic impact. In addition, some national forest inventories have not been fully integrated in large-scale biodiversity infrastructures or registries (e.g. Global Biodiversity Infrastructure Facility, Global Index of Vegetation Databases, Botanical Information and Ecological Network). Therefore, data on tree species distributions, even when available, may be highly heterogeneous. The result is that most insights on the role of biodiversity in forest ecosystems come from certain well-studied regions (Ruiz-Benito et al. 2014) or concentrate on certain biomes for which research collaboration networks are in place (ter Steege et al. 2006; Sullivan et al. 2017).

Despite such concerns related to geographic coverage and sampling effort, another key challenge is the quality of the data being used (Boyle et al. 2013; Enquist et al. 2016; Franklin et al. 2017). For instance, Wiser (2016) found that a common issue in vegetation plot data bases is incorrect or missing geo-coordinates. Other challenges generally found in large aggregations of occurrence records are related to taxonomic miss-identification and taxonomic shifts (Thessen and Patterson 2011), for which tools have been specifically designed to harmonize naming

conventions (Boyle et al. 2013). A formal protocol for quality assessment and quality control is still a challenge (Franklin et al. 2017), although key recommendations are in place (Costello et al. 2014; Enquist et al. 2016; Anderson et al. 2016).

In this study, we tackle the challenge of big-data assessment for occurrence data for all 60,065 tree species identified in the world (Beech et al. 2017). Specifically, our aim is to determine the current (1) geographical coverage and (2) quality of big data of tree species occurrences publicly available in major biodiversity repositories. We aggregated four major vegetation biodiversity occurrence data sources and developed a workflow to categorize occurrence records according to their quality for the purpose of macroecological species distribution analysis – a major tool to explore the exposure of species to climate change (Dawson et al. 2011; Serra-Diaz et al. 2014).

## Methods

### Species selection and occurrence records

We selected species from the world tree species checklist (GlobalTreeSearch; GTS; Beech et al. 2017) for use in our analysis. In this list, a species is considered to have a tree growth habit when it is “*a woody plant with usually a single stem growing to a height of at least two metres, or if multi-stemmed, then at least one vertical stem five centimetres in diameter at breast height*” – a definition from the Global Tree Specialist Group of the International Union for Conservation of Nature. The tree species checklist is available through an online searching engine, which provide tree species list by country ([http://www.bgci.org/global\\_tree\\_search.php](http://www.bgci.org/global_tree_search.php), accessed June 2017). To avoid potential taxonomic issues, we used the Taxonomic Name Resolution Service (TNRS) online tool (Boyle et al. 2013). This tool provides a method for species name standardization, resolving issues related to taxonomic semantic heterogeneity. From the initial species checklist, we selected species that TNRS rendered an ‘accepted name’ taxonomic status. In order to maximize the inclusion of species, we also included species which name was classified as ‘*no opinion*’ in the TNRS. The final set includes 58,100 species.

We collected tree species occurrence data from five major sources of widely used, easy to access and publicly available data of species occurrences: the Global Biodiversity Information Facility (GBIF; <http://www.gbif.org>), the public domain Botanical Information and Ecological Network v.3 (BIEN; <http://bien.nceas.ucsb.edu/bien/>), Latin American Seasonally Dry Tropical Forest Floristic Network (DRYFLOR; <http://www.dryflor.info/>; Banda et al. 2016), RAINBIO database (<http://rainbio.cesab.org/>; Gilles et al. 2016) and the Atlas of Living Australia (ALA; <http://www.ala.org.au/>). These databases were selected because they aggregate many different sources of species

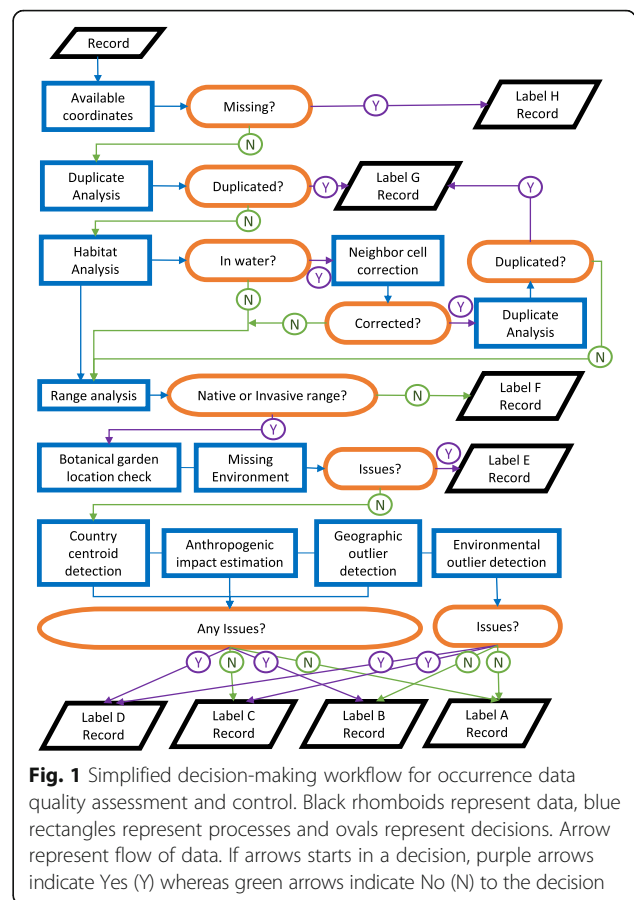
occurrence records, cover a wide range of ecosystems, and constitute the main sources of information used in biodiversity studies. GBIF was accessed using the *rgbif* package (Chamberlain 2017), BIEN data was accessed using the *BIEN* package (Maitner 2017), and ALA data was accessed using the *ALAAR* package (Raymond et al. 2017) developed in R v3.4.1(R Core Team 2017). DRYFLOR and RAINBIO were accessed and downloaded online.

**Quality assessment of occurrence records**

We used a workflow to assess the quality, quantity and coverage of occurrence data for each species (Table 1; Fig. 1). The quality control workflow follows a hierarchical procedure and labels each occurrence record with

**Table 1** Occurrence categories identified by our data cleaning workflow. A list of potential actions (non-comprehensive) are outlined to emphasize the potential of the workflow to improve current datasets

Label	Label name	Label information (I) and potential actions to be developed (A)
H	Missing coordinates	(I) No detected coordinates. (A) Trace back record and assign coordinates.
G	Duplicated records	(I) If record is duplicated within the environmental grid cell, it may give information of sampling effort.
F	Unknown range	(I) Not known range. (A) Double check country-level databases and invasive registries of countries where the record occurs. If present, update database. (A) Re-check common coordinate errors (Yesson et al. 2007).
E	Missing environmental information or unlikely environment (botanic garden)	(A) Check suitability of spatial layers. (A) Confirm botanic garden location. (A) Re-check common coordinate errors (Yesson et al. 2007).
D	Geographic coordinate issues and environment issues	(A) Re-check common coordinate errors (Yesson et al. 2007). (A) Check values in environmental layers.
C	Geographic coordinate issues	(A) Re-check common coordinate errors.
B	Environmental space issues	(A) Check values in environmental layers.
A	No issues detected	(A) Send a 'thank you' email to the database custodian.
AA	High precision	(A) Send a 'thank you' email to the database custodian.
AAA	High precision and low environment uncertainty	(A) Send a 'thank you' email to the database custodian.



**Fig. 1** Simplified decision-making workflow for occurrence data quality assessment and control. Black rhomboids represent data, blue rectangles represent processes and ovals represent decisions. Arrow represent flow of data. If arrows starts in a decision, purple arrows indicate Yes (Y) whereas green arrows indicate No (N) to the decision

an alphabetic code from A to H (Fig. 1). These labels broadly categorize positional as well as biological characteristics of the record and are fit for the purpose of big data macro-ecological analysis. That is, some labels reflect the precision of the occurrence data with respect to the resolution of the environmental spatial layers used for analysis. Therefore, a label for a given record may vary if input environmental data used for the workflow is different. In this study, we used worldclim v. 1.4, a collection of climate environmental layers at 0.5 arcsec spatial resolution (Hijmans et al. 2005).

The workflow starts with a table of species presence records in latitude-longitude WGS84 geographic coordinate system. This table results from the integration of the sources listed above with the following fields: 'x' (longitude), 'y' (latitude), 'elevation', 'country', and 'locality'. In step 1, we identify those species for which we have no spatial data, and we categorize them as quality label H. In step 2, we identified duplicate records. This step is important because data aggregators may use the same sources of information and data integration and subsequent analysis may lead to pseudo-replication or overestimation of sampling intensity. Two types of duplicates were identified and flagged: records with identical coordinates –true duplicates—and those

records that fall in the same grid cell of the environmental spatial layers – geographical duplicates. These records were assigned the quality label G. Subsequently, in step 3, the workflow performs an environmental congruence analysis: it checks that occurrence records are not located in the sea or lakes. Due to accuracy of presence records or the environmental data used, some tree species occurrences located in shorelines may wrongly be assigned to such environments. To avoid such spurious effects we use the function *nearestcell* in package *bioGeo* in R (Robertson et al. 2016), which checks whether neighboring environmental cells have congruent (e.g. terrestrial) environments. If that is the case, the function assigns new coordinates to the record and creates a new field with the names 'x.orginal' and 'y.orginal' to keep track of the transformation. Records with transformed coordinates are evaluated again for duplicates, as outlined previously in step 2.

In step 4, we identified potential geolocation issues based on known species ranges. We used the GTS database (Beech et al. 2017) to identify countries where the target species is considered native. Additionally, we identified countries where the species is considered introduced and naturalized through a database merging the global invasive species database (GISD; Invasive Species Specialist Group ISSG 2017) and the global register of introduced and invasive species (GRIIS; Invasive Species Specialist Group ISSG 2017). The latter provide a conservative assessment of the naturalized range, as areas where the species only has sparse naturalized occurrences may not be registered. If the location of the record did not fall in the combined list of countries in which a species is known to occur, the record was assigned a quality label E. Country spatial layer was obtained from the Global Administrative database v.2.8 (GADM 2015).

After these four steps, several independent tests were performed with the remaining records – i.e. those without assigned quality label. These tests consist of:

- 1) Missing environment checks: Identification of locations for which the spatial environmental layer has no information, which precludes the use of such records for environmental modeling.
- 2) Potential botanical garden locations: Identification of records in botanical gardens, for which local conditions may widely differ from environmental conditions in environmental spatial layers at large scales. Locality data, if available, is used to check whether the location contains words that could be identified as a botanical garden location: 'botanic', 'botanische', 'botanico', 'jardin', 'garden', 'botanical'.
- 3) Centroid detection: During data digitization, a well-known error is the automatic assignation of a specimen to the centroid country or its capital. We used the World Factbook (CIA 2014) for information of countries and capital centroids. The data can freely be downloaded from packages *speciesgeocodeR* (Zizka 2015). The workflow identifies records that fall within the environmental grid cell of the country or capital centroid, or an adjacent 8-cell neighbor.
- 4) Hyper-anthropogenic environment: This analysis identifies records in highly urbanized landscapes, for which global environmental layers may not suitably portray the conditions on site, and where occurrence records may reflect planted specimens. We overlay the record location with the human influence index v2 spatial layer (Wildlife Conservation Society - WCS and Center for International Earth Science Information Network - CIESIN - Columbia University 2005). The human influence index is a value that integrates several spatial layers (land use, population, etc.) to estimate the level of anthropogenic impact. We determined a high human influence record those records with an index greater than 45. This threshold was set because it enables the characterization of large, highly-dense metropolitan areas.
- 5) Geographical outlier: We identify potential errors in coordinates using an alpha-hull methodology. This technique has been used to determine species distribution and respective range sizes under data bias (Hui et al. 2011; Capinha and Pateiro-López 2014). Using this methodology, a convex-hull is drawn in the coordinate plane using record locations; and an alpha parameter is used to split the polygon in places where few records are present. Alpha-shapes have been used to identify potential outliers in species ranges. We used an alpha parameter of 2 for this analysis – chosen based on preliminary analysis in ALA. If there is not enough information to compute alpha hull, then the geographic outlier analysis is not performed.
- 6) Environmental outlier: We identified potential errors in presence records by identifying records located in environments considered an outlier in the environmental space. These locations may significantly affect environmental analysis using species distributions (Soley-Guardia et al. 2014). The environmental outlier detection was performed by running a reverse-jackknife method on six climatic variables: Annual mean temperature, maximum temperature of the warmest month, minimum temperature of the coldest month, annual precipitation, precipitation of the wettest month, precipitation of the driest month. These variables were selected to reflect the averages as well as the climatic boundaries of the species range. The method consists of

identifying outlier samples (e.g. occurrences) based on a critical threshold calculated based on mean, standard deviation and range of the whole set of samples (e.g. all occurrence points). This method has been applied to detect outliers (Chapman 2005). Potential outliers were identified if more than 20% of the variables were estimated as outliers (minimum of two variables in our case). If there is not enough information to compute this analysis, then the environmental outlier analysis is not performed.

The six independent tests outlined above allow classification into five quality categories (A to E; highest to lowest). Label E is assigned to records for which there was not environmental information (analysis 1) or the occurrence was likely to be in a botanical garden (analysis 2). Label D is assigned to records with at least one issue in the environmental (analysis 3–5) and at least one issue in the geographical space (analysis 6), thus compromising potential macroecological analysis. Label B indicates records with issues only in geographical space (analysis 3–5), whereas label C was assigned to records with potential issues in the environmental space (analysis 6). Label A was assigned for records with no apparent issues.

Finally, we further differentiated several categories among the highest quality records (Label A). Specifically, we add two more high-quality categories: AA and AAA. These labels determine the geographic precision of the data and may be especially useful in climate-distribution analysis. Precision was determined by the location of the record with respect to the gridded environmental data. Low precision records are identified when presence record was located exactly at the top-left corner or center of the spatial resolution of a grid cell (Robertson et al. 2016). AA label is assigned to already classified A records with good environmental precision. Subsequently, AAA label was assigned to already classified AA records which had suitable altitudinal precision (100 m). This is important in locations with topographic heterogeneity, where climate may differ greatly within a coarse-resolution grid cell of a climate spatial layer. We compared the recorded altitude of the presence to the elevation derived from digital elevation model used to obtain the environmental variables (Hijmans et al. 2005). If mismatch was less than 250 m, we reassigned AA record to an AAA record.

**Results**

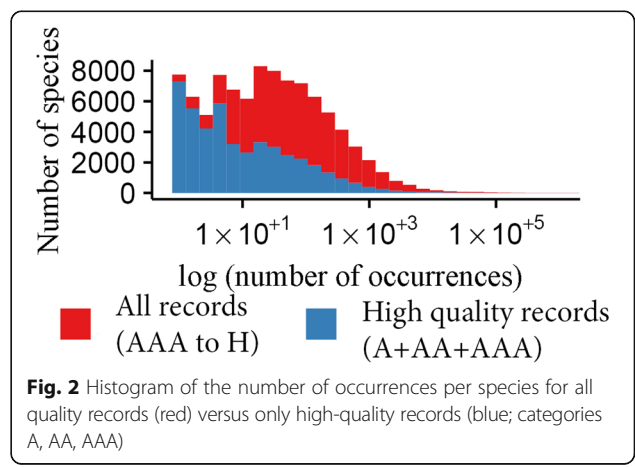
The data integration performed in this study gathered a total number of 36.69 M occurrence records of tree species, corresponding to a total 1.28 M unique locations in the world (Table 2). Our quality assessment indicated that 6.40 M were found in the highest quality category

**Table 2** Number of occurrence records per quality category in the dataset

Category	No. occurrences	Percentage
AAA	968,884	2.64
AA	5,059,644	13.79
A	377,063	1.03
B	1,212,822	3.31
C	6,880	0.02
D	17,580	0.05
E	32,698	0.09
F	859,433	2.34
G	23,177,564	63.17
H	4,979,574	13.57
Total	36,692,142	

(AAA + AA + A) (Table 2). Among these three high quality categories, AA was the most frequent indicating a good match between the environmental data used (worldclim 30 arcsec) and the occurrence records (Table 2). The occurrence quality with the most records correspond to label G (duplicates) with 23.18 M records (63.17%) and the label with the fewest records corresponds to label C (environmental outliers; 6880 occurrences; 0.02%; Table 2). The number of records without coordinates represent 13.57% of the total records (4.98 M), almost equaling the amount of high quality records (Table 2).

The number of species analyzed were reduced from 60,065 species to 58,100 species after the taxonomic name filtering and cleaning process (see Methods). Among the species analyzed, 49,206 (84.69%) had at least one occurrence record, and 45,797 (78.82%) had at least one record of high quality. Only 15,140 species (26.05%) had at least 20 records of high quality (AAA to A), which could be regarded as a minimum to perform an SDM. In general, most of the species had low occurrence record numbers (Fig. 2). The number of species with less than 10



records only constitutes 14.29% of the species analyzed (8,307 species). Past the threshold of 10 records per species, the number of species decreases sharply with the number of occurrence per species (Fig. 2). That is, fewer and fewer species exist with higher number of occurrences. The quality of the records varied largely among species. Crucially, a low number of species (1,109) were characterized by more than 90% of the records without coordinates (label H), and only 65 of these had all records in this category.

The patterns of the different quality labels change between the percentages over the total number of occurrences, and the percentages of quality at the species level (Table 2 vs. Fig. 3, respectively). At the species level, the majority of tree species had a high percentage of duplicate records (label G; Fig. 3) followed by records without coordinates (label H; Fig. 3). This was somewhat expected given the combination of large aggregated databases that may share the same data sources. Across species, these account for 38.4% and 31.3% respectively (Table 3), but the variation across species is large in these categories (Fig. 3). High-quality records show a remarkable percentage of the records per species. Together (AAA + AA + A), high-quality records represent the 18% of the records per species (median values; Fig. 3 and Table 3). Among these categories, AA is the most prominent category showing that these occurrences tend to have good precision. However, the low values of AAA compared to AA show that most of the elevation accuracy data is not present in the records. Label B records – potential geographical outliers – represent around 10% of the records per species, almost half of the high-quality records.

The species with highest number of records is *Pinus taeda* (Pinaceae) with 1.48 M records – 38,938 considered high quality for SDMs. In general, the species with

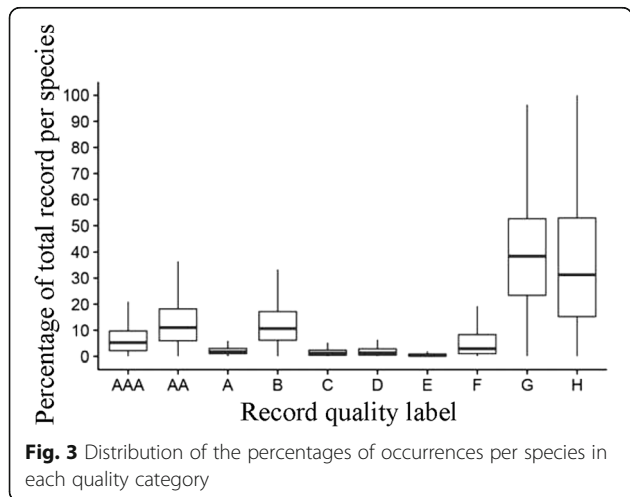
**Table 3** Species' (median) percentage of occurrence in each quality category, shown for the whole dataset and for five economically important tree families

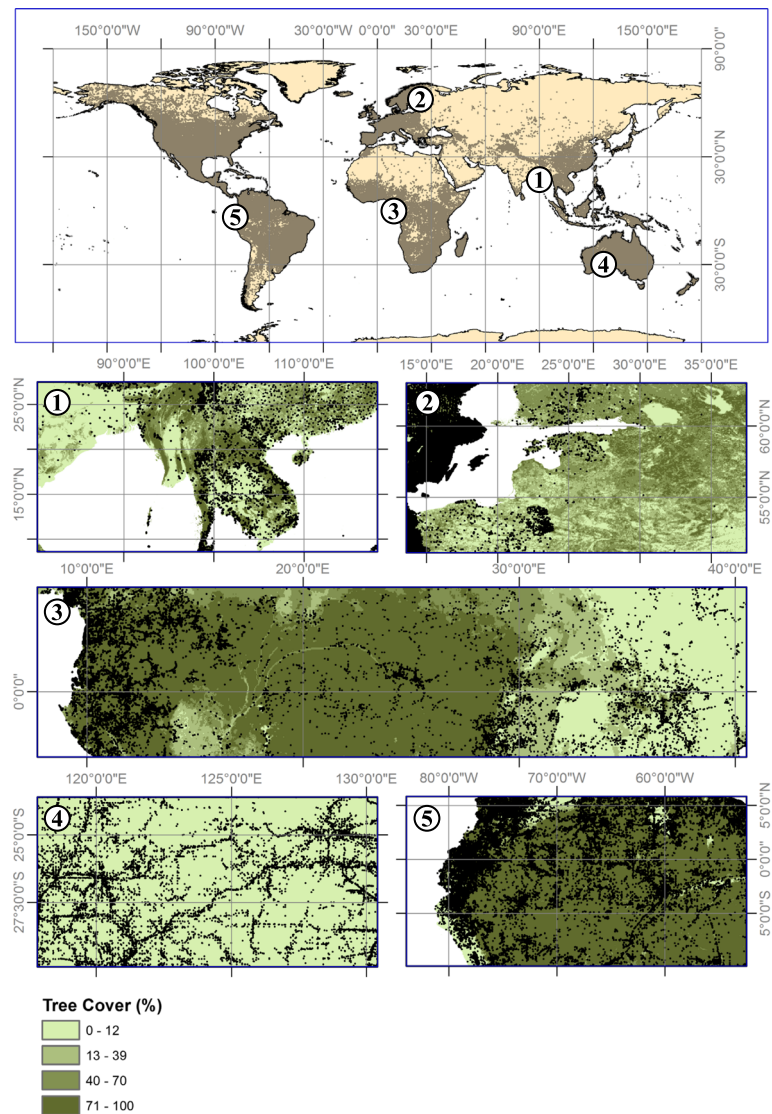
Family	A-type	B	C	D	E	F	G	H
All	18.0	10.6	1.2	1.3	0.3	2.9	38.4	31.3
Pinaceae	11.6 ↓	3.3 ↓	0.4 ↓	0.4 ↓	0.0 ↓	2.0 ↓	44.9 ↑	25.2 ↓
Arecaceae	12.8 ↓	9.1 ↓	1.3 ↑	1.0 ↓	0.4 ↑	3.2 ↑	41.3 ↑	35.3 ↑
Myrtaceae	20.6 ↑	7.4 ↓	0.9 ↓	0.7 ↓	0.3 =	1.8 ↓	42.2 ↑	29.4 ↓
Dipterocarpaceae	7.3 ↓	4.6 ↓	0.6 ↓	1.1 ↓	0.8 ↑	2.2 ↓	7.4 ↓	88.9 ↑
Fagaceae	19.5 ↑	8.3 ↓	0.8 ↓	0.8 ↓	0.1 ↓	1.7 ↓	42.9 ↑	25.0 ↓

Arrows represent an increase or decrease with respect to the percentages for the whole dataset. Red arrows indicate an increase or reduction of more than 50% with respect to the reference of percentages of all dataset

the highest records were North American trees. The second and the third species in number of occurrences were *Acer rubrum* (Sapindaceae) and *Liquidambar styraciflua* (Altingiaceae), both accounting for more than 0.5 M occurrence records. European and North American species dominate the top-100 list. Other 'important' trees outside these geographical areas are orders of magnitude lower than the top three species. For instance, *Faramaea coffeoides* (Rubiaceae) in South America had a total of 51,698 occurrence records, but only 20 of those had enough quality for macroecological SDM (AAA-A). In New Caledonia, *Amborella trichopoda* (Amborellaceae) had a total of 51,348 occurrences, but only 67 records usable for macroecological SDM. In New Zealand, higher levels of occurrence records were found for some key tree species. *Griselinia littoralis* (Griselinaceae) – with 68,182 total occurrences, and *Coprosma foetidissima* (Rubiaceae) – with a total of 49,184 occurrences, maintained a large number of high quality occurrence records (AAA-A): 12,236 and 10,090 records, respectively.

The geographical coverage of the occurrences showed strong spatial patterns (Fig. 4). A huge area of the world has accessible survey records at relatively high intensity. Continents like North America, Western Europe and Australia show large geographical space of surveyed area (Fig. 4, shaded areas). Conversely, the Russian boreal forests are very under-represented by the databases assessed in this study. Regional inspection also showed different focal regions of sampling. In Southeast Asia clear differences in sampling intensity arise between high occurrence areas in forests of Vietnam, Laos, Cambodia and Thailand, and low levels in Northern Myanmar and Bangladesh (Fig. 4, Inset 1). Similarly, country-level differences also arose within Europe (Fig. 4, Inset 2), and a clear West-East gradient of occurrence density is apparent in central African tropical forests (Fig. 4, Inset 3). Australia had high levels of occurrence data, even in places of low tree cover (Fig. 4, Inset 4). In addition, we found that the Amazon forest samples were also





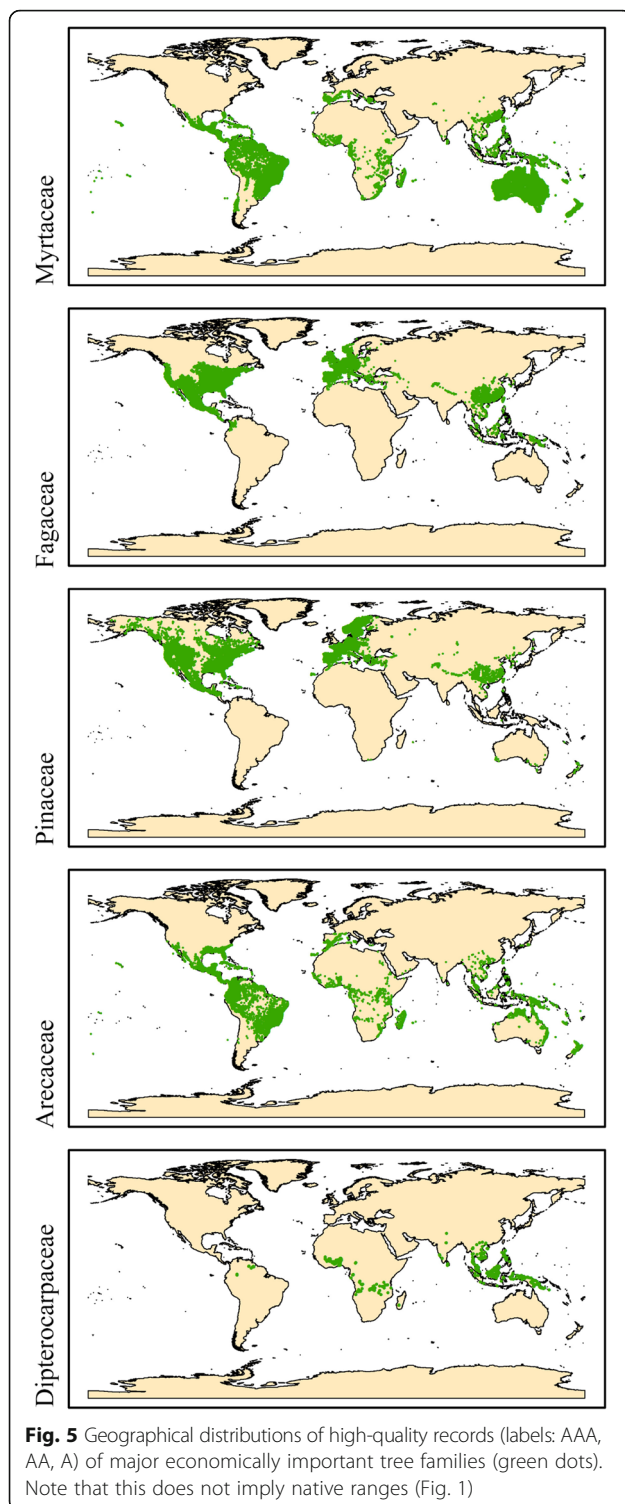
**Fig. 4** Geographical coverage of tree occurrence data. Shaded grey areas show areas of general sampling in the databases integrated. Numbers in map represent insets where regional occurrence data (black dots) is shown: (1) South and Southeast Asia, (2) Northeast Europe, (3) Central Africa, (4) South-Southwest Australia and (5) Pacific coast of South America and Amazon

highly clustered (Fig. 4, Inset 5), reflecting that research efforts in this region are still concentrated in its more accessible areas.

We did not capture any general deviation between the percentage of quality records between all tree species and tree species in families of economic importance (Table 3), except in the case of Dipterocarpaceae. This family consistently showed a high number of records without coordinates, and fewer high-quality records, likely reflecting the general dearth of accessible records from Southeast Asia. Other economically important families showed higher levels of duplications, which is due to the high sampling intensity in some regions. However, we cannot rule out that it could be to the

relatively low resolution of the environmental grids used. Spatial distributions of such families were consistent with previously known distribution of such taxa (Fig. 5), but some exceptions arose. For instance, trees of the family Pinaceae are present in Australia and New Zealand (Fig. 5), which reflect introduced species in the continent from Europe and North America.

Geographical 'holes' for these families are similar to those found for the entire dataset (Fig. 4 vs Fig. 5), highlighting regions like Russia, Central Africa and to some extent the Amazon region as key data gaps. For instance, relatively lower quality percentages were found Dipterocarpaceae, a family distributed in the tropics.



## Discussion

Our study highlights the wealth of information – 6.4 M high quality records and 57,849 species – in public databases of tree species occurrence and discovers geographical gaps in the data that need to be filled in order

to improve our understanding of tree distributions and diversity on Earth.

Several aspects of the geographical coverage of tree distribution share similarities with the geographical coverage of plants in big-data biodiversity databases. For instance, around six orders of magnitude difference in occurrence data are found among species considered (Meyer et al. 2016). In addition, significantly high occurrence densities are biased towards North America, Europe and Australia (Meyer et al. 2016), reflected also in the native countries of the tree species with most records. Geographical coverage patterns of tree species in some areas mimic sampling intensity of seed plants (Stropp et al. 2016), with a west to east gradient of decreasing sampling intensity (Fig. 4). In the case of trees, important data gaps exist in key biodiverse areas, especially in the Central African tropics, with highly heterogeneous sampling (Fig. 4, Inset 3). These differences result from colonial history, but also from incomplete digitization of existing records (Figueiredo et al. 2009; Sosef 2016; Gilles et al. 2016; Sosef et al. 2017). In addition, important conflicts in such biodiverse regions make ongoing data collections difficult (Hanson 2011). On the other hand, the geographic gap in Russia may be due to the databases selected, urging for an integration of Asian boreal forest data in future studies. This picture of geographical gaps was also present even in the case of key economic families.

Filtering large amounts of records due to potential errors was similar between our study of tree species and a study of all plant species when applying ‘strict’ filtering procedures (e.g. ~40%, Meyer et al. 2016). Conversely, occurrence data for trees shows higher levels of taxonomic confidence, with 96% of species classified taxonomically with at least one record, than for all plants, where this number decreases to 66% (Meyer et al. 2016). That is likely because our initial list is based on an already curated list of tree species.

Our global screening of tree species occurrences showed that only 13% of the records did not contain any geographical coordinates; and that the selected databases hold 6.4 M records of high quality, representing 17.4% of the total data available data on tree occurrences. Although these numbers may seem low, we argue that these are actually high numbers for big data analysis of species distributions. Part of this ‘low’ percentage corresponds to the fact that our data integration and quality control workflow not only reflect the properties of the occurrence data, but also the resolution of the environmental data. That is, in our framework duplicates may be originated by duplicated museum records, but also by data aggregators and by the spatial resolution of the environmental layers. For instance, BIEN3 aggregates large botanical databases, including GBIF. In addition, because our workflow is aimed at the eventual production of



SDMs, two different locations may be considered duplicates if they occur within the same grid cell. It is thus not surprising to find high levels of quality G records in our analysis.

#### **From scrubbing to profiling and corrections**

The filters applied in our workflow reflect that, for trees, major barriers to classifying high-quality records were due to the geographical space analysis (label B) as well as the range of the species (label F). These two labels held 1,212,822 and 859,433 records, respectively.

Label B reflects issues related to the geographical coordinates. That is, records within the environmental space of the species, but found in either the country centroid, the capital centroid, highly urbanized areas, or areas outside a typical alpha-shape extent of occupancy. Whether to include these records in an SDM or not may require further scrutiny and will largely depend on the question being addressed. These records, even if not correct, are unlikely to strongly bias estimates of the environmental space, although that will depend on total species sample size and model complexity (Merow et al. 2014). For example, such records would have little to no effect on delimiting range boundaries (depending on the modeling algorithm used) although they may have stronger effects on continuous estimates of habitat suitability. Sample biases may inflate the number of records under label B, despite the record being correct. For example, for a given species, sampling in two different countries may be very different and records in countries with lower sample intensity may appear as a geographical outlier. Therefore, we suggest that records under such profile (label B) should be considered in species with low sample sizes when the record is identified to be outside the alpha hull. In such cases, the risk of inclusion may be lower than the benefit of being able to fit an SDM with higher number of occurrences.

Issues with coordinates have long been discussed in occurrences (Yesson et al. 2007; Anderson et al. 2016) and new tools have been recently developed to correct them. This is of utmost importance for rare species or for species in scarcely sampled locations. Robertson et al. (2016) developed an R package capable of implementing alternative geographic coordinates for a sample based on common errors encountered in GBIF. Errors like substitution between latitude and longitude or wrong hemisphere recording are some of these common errors. An automated and scalable form of this would qualitatively advance the use of such records, currently being filtered (or scrubbed) from analysis.

Great care should be taken when correcting or identifying such geographical issues. We use alpha-shapes because they have been used to capture extent of occurrences and may be preferred over other methods to

delineate species range boundaries (García-Roselló et al. 2015). However, the method may be misleading for rare species or may tend to identify geographical outliers for rare species, even though these may be widely distributed (Zizka et al. 2017).

Unknown status of a species in a country (native, invasive, naturalized) was another label (F) that led to profiling a large amount of occurrences (859,433). Generally, this filter is imposed to account for the fact that, in some cases, occurrence records may have coordinates in countries where the collection is located, rather than where the specimen was collected. We acknowledge that this filter may be overly conservative, however it remains useful as a globally comprehensive registry of native, naturalized and invaded country-level ranges are lacking. We integrated global registries and species checklists (GTS, GIISD, GRIIS), but these are unlikely to be complete, and are under constant updates and may not cover other important sources, like alien species used in plantations. In addition, while checklists are key political instruments for conservation, they may be of limited use for quality control, especially in the case of large countries or countries with complex geographies. For instance, archipelago-countries may be composed of species endemic to an island and not necessarily to others; or large countries covering large variation in climates (e.g. Brasil, USA, China) may not be able to capture erroneous records that are endemic to some biomes, but not others. Current biodiversity informatics is performing large efforts and key developments given the need for plant range status information to increase the quality of the data. For instance, developing curated lists of native ranges of species (<http://bien.nceas.ucsb.edu/bien/tools/nsr/>).

The workflow developed in this study is not only aimed at filtering data, but rather at emphasizing and flagging errors and potentially increasing discoverability – the degree to find information associated with the occurrence data (Table 2). For instance, a record with a profile F may help cross-check species-country checklists. We expect that profiling may enhance the quality of species occurrence data but may also be useful when implementing workflows for geo-correction.

#### **Collaboration, networks and mobilization in the forest big data era**

The data aggregators used here include many different sources of information and researchers that have participated in this large collaborative effort. Yet, our geographical coverage analysis shows that geographical data gaps are yet to be filled for tree species (Fig. 4). This was somewhat surprising because trees are among the best studied groups of organisms in the world. Issues related to funding for data production and mobilization (Bradley et al. 2014; Nowogrodzki 2016) and proper

acknowledgement of investigators still need to be fully resolved to ‘uncover’ data for species distributions (Costello et al. 2014; Franklin et al. 2017). We encourage scientists to acknowledge the effort of those scientists and data infrastructures that support high-quality data (Table 2), but we are aware that a full system recognition of data work recognition is still to be developed. Scientific networks have increased the availability of some regions and to some extent under-sampled forests, while allowing recognition (co-authorship) of the network participants. However, these tend to be restricted to forest types of biomes of the research focused group (e.g., Amazon tree Network, etc.).

In the case of trees, national forest inventory programs offer good opportunities for expansion of tree occurrence data. Assessment of wood resources has historically been performed in many countries (Vidal et al. 2016) and may offer good temporal and geographical coverage for many forest species. In addition, they may lack some of the typical spatial biases present in museum records collected rather opportunistically (Pyke and Ehrlich 2010). Some forest inventories are readily incorporated in the big data aggregators. For instance, BIEN3 incorporates the United States of America Forest Inventory and Analysis (FIA) data, and the Spanish Forest Inventory dataset is incorporated in GBIF. However, a larger mobilization (e.g. incorporation into digital format) of forest inventory data into biodiversity databases is lacking, probably due to difference in the tradition of specimen collections versus assessment of wood resources. Initiatives like the Global Forest Biodiversity Initiative (<http://www.gfbinitiative.org/>) world forest plot data will be key for improving tree coverage.

It is important to note that mobilization of data may increase geographical biases, sometimes in surprising ways. For instance, Yang et al. (2016) found that vascular plants were relatively well sampled in mountains in China, but densely-populated areas were under-sampled. It is likely that forest inventories could increase geographical bias, towards countries with certain economic status. That said, new techniques to overcome such biases are in place for overcoming such unintended consequences of data mobilization (Phillips et al. 2009; Merow et al. 2016). Therefore, the issue of data sparsity may be more critical than the biases incurred when mobilizing data. Under this context, forest inventories and capacity building performed by the FAO-forestry in certain regions is an important effort that could significantly contribute to global biodiversity datasets.

## Conclusions

Big data for tree species occurrence is abundant and readily useable for macroecological analysis of species

distributions, despite geographical gaps which are still present in some regions. We identified geographical coordinate errors in many records and therefore suggest that future data analysis and filtering of occurrences should be coupled with a workflow for profiling and geo-correction. While data mobilization for published datasets is always beneficial, we expect that incorporation of forest national inventories to biodiversity databases will enhance the quality of world tree distribution assessments.

## Abbreviations

ALA: Atlas of Living Australia; BIEN: Botanical information and ecological network; DRYFLOR: Latin American Seasonally Dry Tropical Forest Floristic Network; GADM: Global administrative database; GBFI: Global biodiversity information facility; GISD: Global invasive species database; GRIS: Global register of introduced and invasive species; GTS: Global tree search; TNRS: Taxonomic name resolution service

## Acknowledgements

JMSD and JCS acknowledge support from the Danish Council for Independent Research | Natural Sciences (Grant 6108-00078B) to the TREECHANGE project. JCS also considers this work a contribution to his VILLUM Investigator project “Biodiversity Dynamics in a Changing World” funded by VILLUM FONDEN. This paper was presented in Beijing as part of the program of the Global Forest Biodiversity Initiative Symposium. JMSD acknowledges support from Beijing Forestry University to attend the meeting. BE, BM, BM, and CM acknowledge funding from NSF – DBI-1565046.

## Funding

Danish Council for Independent Research | Natural Sciences (grant 6108-00078B). VILLUM FONDEN – VILLUM investigator “Biodiversity Dynamics in a Changing World”. NSF – DBI-1565046.

## Availability of data and materials

Data is available in public repositories. Global Biodiversity Information Facility (GBIF; available at <http://www.gbif.org>), public domain Botanical Information and Ecological Network v.3 (BIEN, available at <http://bien.nceas.ucsb.edu/bien/>), Latin American Seasonally Dry Tropical Forest Floristic Network (DRYFLOR; available at <http://www.dryflor.info/>), RAINBIO database (available at <http://rainbio.cesab.org/>), Atlas of Living Australia (ALA; available at <http://www.ala.org.au/>).

## Authors’ contributions

JMSD and JCS designed the methodology for the workflow. JMSD implemented the workflow and drafted the initial manuscript. JCS, BJE, BM, CM provided subsequent input to the initial draft. All authors read and approved the final manuscript.

## Authors’ information

JMSD is a postdoctoral researcher at the Ecoinformatics and Biodiversity group in Aarhus University. He is interested in cross-scaling dynamics in tree species distributions, and global change effects in forests worldwide.

## Ethics approval and consent to participate

Not applicable.

## Consent for publication

Not applicable.

## Competing interests

The authors declare that they have no competing interests.

## Author details

<sup>1</sup>Section for Ecoinformatics and Biodiversity, Department of Bioscience, Aarhus University, Ny Munkegade 114, DK-8000 Aarhus C, Denmark.

<sup>2</sup>Department of Ecology and Evolutionary Biology, University of Arizona, Tucson AZ 85721, USA. <sup>3</sup>Department of Ecology and Evolutionary Biology,

Yale University, New Haven CT 06520, USA. <sup>4</sup>Center for Biodiversity Dynamics in a Changing World (BIOCHANGE), Department of Bioscience, Aarhus University, Ny Munkegade 114, DK-8000 Aarhus, Denmark.

Received: 25 August 2017 Accepted: 11 December 2017

Published online: 15 January 2018

## References

- Allen CD, Macalady AK, Chenchouni H, Bachelet D, McDowell N, Vennetier M, Kitzberger T, Rigling A, Breshears DD, Hogg EH, Gonzalez P, Fensham R, Zhang Z, Castro J, Demidova N, Lim JH, Allard G, Running SW, Semerci A, Cobb N (2010) A global overview of drought and heat-induced tree mortality reveals emerging climate change risks for forests. *For Ecol Manag* 259:660–684
- Anderson RP, Araújo M, Guisan A, Lobo JM, Martínez-Meyer E, Peterson AT, Soberón J (2016) Final report of the task group on GBIF data fitness for use in distribution modelling [https://seval.unil.ch/resource/seval:BiB\\_768D188CEA5B.P001/REF](https://seval.unil.ch/resource/seval:BiB_768D188CEA5B.P001/REF). Accessed 17 June 2017
- Ash JD, Givnish TJ, Waller DM (2017) Tracking lags in historical plant species' shifts in relation to regional climate change. *Glob Change Biol* 23:1305–1315
- Banda K, Delgado-Salinas A, Dexter KG, Linares-Palomino R, Oliveira A, Prado D, Pullan M, Quintana C, Riina R, Rodriguez JM, Weintritt J, Acevedo-Rodriguez P, Adarve J, Alvarez E, Aranguren A, Arteaga JC, Aymard G, Castano A, Ceballos-Mago N, Cogollo A, Cuadros H, Delgado F, Devia W, Duenas H, Fajardo L, Fernandez A, Fernandez MA, Franklin J, Freid EH, Galetti LA, Gonto R, Gonzalez-M R, Graveson R, Helmer EH, Idarraga A, Lopez R, Marcano-Vega H, Martinez OG, Maturu HM, McDonald M, McLaren K, Melo O, Mijares F, Mogni V, Molina D, Moreno ND, Nassar JM, Neves DM, Oakley LJ, Oatham M, Olvera-Luna AR, Pezzini FF, Dominguez OJR, Rios ME, Rivera O, Rodriguez N, Rojas A, Sarkinen T, Sanchez R, Smith M, Vargas C, Villanueva B, Pennington RT (2016) Plant diversity patterns in neotropical dry forests and their conservation implications. *Science* 353:1383–1387
- Beech E, Rivers M, Oldfield S, Smith PP (2017) GlobalTreeSearch: the first complete global database of tree species and country distributions. *J Sustain For* 36:454–489. <https://doi.org/10.1080/10549811.2017.1310049>
- Botkin DB, Saxe H, Araujo MB, Betts R, Bradshaw RHW, Cedhagen T, Chesson P, Dawson TP, Etterson JR, Faith DP, Ferrier S, Guisan A, Hansen AS, Hilbert DW, Loehle C, Margules C, New M, Sobel MJ, Stockwell DRB (2007) Forecasting the effects of global warming on biodiversity. *AIBS Bull* 57:227–236
- Boyle B, Hopkins N, Lu Z, Garay JAR, Mozzerin D, Rees T, Matasci N, Narro ML, Piel WH, Mckay SJ, Lowry S, Freeland C, Peet RK, Enquist BJ (2013) The taxonomic name resolution service: an online tool for automated standardization of plant names. *BMC Bioinformatics* 14:16
- Bradley RD, Bradley LC, Garner HJ, Baker RJ (2014) Assessing the value of natural history collections and addressing issues regarding long-term growth and care. *Bioscience* 64:1150–1158. <https://doi.org/10.1093/biosci/biu166>
- Capinha C, Pateiro-López B (2014) Predicting species distributions in new areas or time periods with alpha-shapes. *Ecol Inform* 24:231–237. <https://doi.org/10.1016/j.ecoinf.2014.06.001>
- Chamberlain S (2017) rgbif: Interface to the Global "Biodiversity" Information Facility "API". R package version 0.9.8. <https://CRAN.R-project.org/package=rgbif>. Accessed 17 June 2017
- Chapman AD (2005) Principles and methods of data cleaning: primary species and species-occurrence data. Global Biodiversity Information Facility, Copenhagen <http://www.gbif.org/document/80528>. Accessed 17 June 2017.
- Choat B, Jansen S, Brodribb TJ, Cochard H, Delzon S, Bhaskar R, Bucci SJ, Feild TS, Gleason SM, Hacke UG, Jacobsen AL, Lens F, Maherali H, Martinez-Vilalta J, Mayr S, Mencuccini M, Mitchell PJ, Nardini A, Pittermann J, Pratt RB, Sperry JS, Westoby M, Wright IJ, Zanne AE (2012) Global convergence in the vulnerability of forests to drought. *Nature* 491:752
- Core Team (2017) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria
- Costello MJ, Appeltans W, Bailly N, Berendsohn WG, de Jong Y, Edwards M, Froese R, Huettmann F, Los W, Mees J, Segers H, Bisby FA (2014) Strategies for the sustainability of online open-access biodiversity databases. *Biol Conserv* 173:155–165. <https://doi.org/10.1016/j.biocon.2013.07.042>
- Dawson TP, Jackson ST, House JI, Prentice IC, Mace GM (2011) Beyond predictions: biodiversity conservation in a changing climate. *Science* 332:53–58
- Enquist BJ, Condit R, Peet RK, Schildhauer M, Thiers BM (2016) Cyberinfrastructure for an integrated botanical information network to investigate the ecological impacts of global climate change on plant biodiversity. *PeerJ Preprints* 4:e2615v2 <https://doi.org/10.7287/peerj.preprints.2615v2>. Accessed 17 June 2017
- Figueiredo E, Smith GF, César J (2009) The flora of Angola: first record of diversity and endemism. *Taxon* 58:233–236
- Franklin J (2010) Mapping species distributions: spatial inference and prediction. Cambridge University Press, Cambridge
- Franklin J, Serra-Diaz JM, Syphard AD, Regan HM (2016) Global change and terrestrial plant community dynamics. *Proc Natl Acad Sci* 113:3725–3734. <https://doi.org/10.1073/pnas.1519911113>
- Franklin J, Serra-Diaz JM, Syphard AD, Regan HM (2017) Big data for forecasting the impacts of global change on plant communities: big data for forecasting vegetation dynamics. *Glob Ecol Biogeogr* 26:6–17. <https://doi.org/10.1111/geb.12501>
- Gamfeldt L, Snäll T, Bagchi R, Jonsson M, Gustafsson L, Kjellander P, Ruiz-Jaen MC, Froberg M, Stendahl J, Philipson CD, Mikusinski G, Andersson E, Westerlund B, Andren H, Moberg F, Moen J, Bengtsson J (2013) Higher levels of multiple ecosystem services are found in forests with more tree species. *Nat Commun* 4:1340. <https://doi.org/10.1038/ncomms2328>
- García-Roselló E, Guisande C, Manjarrés-Hernández A, Gonzalez-Dacosta J, Heine J, Pelayo-Villamil P, Gonzalez-Vilas L, Vari RP, Vaamonde A, Granada-Lorenzo C, Lobo JM (2015) Can we derive macroecological patterns from primary global biodiversity information facility data?: macroecological patterns and GBIF data. *Glob Ecol Biogeogr* 24:335–347. <https://doi.org/10.1111/geb.12260>
- Gilles D, Zaiss R, Blach-Overgaard A, Catarino L, Damen T, Deblauwe V, Dessein S, Dransfield J, Droissart V, Duarte MC, Engledow H, Fadeur G, Figueira R, Gereau RE, Hardy OJ, Harris DJ, de Heij J, Janssens S, Klomberg Y, Ley AC, Mackinder BA, Meerts P, van de Poel JL, Sonke B, Sosef MSM, Stevart T, Stoffelen P, Svenning JC, Sepulchre P, van der Burgt X, Wieringa JJ, Couvreur TLP (2016) RAINBIO: a mega-database of tropical African vascular plants distributions. *PhytoKeys* 74:1–18. <https://doi.org/10.3897/phytokeys.74.9723>
- Graham C, Ferrier S, Huettman F, Moritz C, Peterson AT (2004) New developments in museum-based informatics and applications in biodiversity analysis. *Trends Ecol Evol* 19:497–503. <https://doi.org/10.1016/j.tree.2004.07.006>
- Guisan A, Tingley B, Baumgartner JB, Naujokaitis-Lewis I, Sutcliffe PR, Tulloch AIT, Regan TJ, Brotons L, McDonald-Madden E, Mantyka-Pringle C, Martin TG, Rhodes JR, Maggini R, Setterfield SA, Elith J, Schwartz MW, Wintle BA, Broennimann O, Austin M, Ferrier S, Kearney MR, Possingham HP, Buckley YM (2013) Predicting species distributions for conservation decisions. *Ecol Lett* 16:1424–1435. <https://doi.org/10.1111/ele.12189>
- Hampton SE, Anderson SS, Bagby SC, Gries C, Han X, Hart EM, Jones MB, Lenhardt WC, Macdonald A, Michener WK, Mudge J, Pourmokhtarian A, Schildhauer MP, Woo KH, Zimmerman N (2015) The Tao of open science for ecology. *Ecosphere* 6:art120. <https://doi.org/10.1890/ES14-00402.1>
- Hanson T (2011) War and biodiversity conservation: the role of warfare ecology. In: Machlis G, Hanson T, Špirić Z, McKendry J (eds) Warfare ecology. NATO science for peace and security series C: environmental security. Springer, Dordrecht
- Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A (2005) Very high resolution interpolated climate surfaces for global land areas. *Int J Climatol* 25:1965–1978. <https://doi.org/10.1002/joc.1276>
- Hui C, Richardson DM, Robertson MP, Wilson JR, Yates CJ (2011) Macroecology meets invasion ecology: linking the native distributions of Australian acacias to invasiveness. *Divers Distrib* 17:872–883. <https://doi.org/10.1111/j.1472-4642.2011.00804.x>
- Invasive Specialist Group ISSG (2017). Global Invasive Species database. Accessed at <http://www.iucngisd.org/gisd/>.
- Invasive Species Specialist Group ISSG 2017. Global Register of Introduced and Invasive Species. Version 2017.1. Accessed at <http://www.griis.org/>.
- Maitner BS, Boyle B, Casler N, Condit R, Donoghue J, Durán SM, D Guaderrama, et al. (2017) The bien r package: A tool to access the Botanical Information and Ecology Network (BIEN) database. *Methods Ecol Evol*. In Press. Accessed at <https://cran.r-project.org/web/packages/BIEN/index.html>. Accessed 17 June 2017
- Merow C, Smith MJ, Edwards TC, Guisan A, McMahon SM, Normand S, Thuiller W, Wuest RO, Zimmermann NE, Elith J (2014) What do we gain from simplicity versus complexity in species distribution models? *Ecography* 37:1267–1281. <https://doi.org/10.1111/ecog.00845>
- Merow C, Allen JM, Aiello-Lammens M, Silander JA (2016) Improving niche and range estimates with Maxent and point process models by integrating spatially explicit information: Minxent. *Glob Ecol Biogeogr* 25:1022–1036. <https://doi.org/10.1111/geb.12453>

- Meyer C, Weigelt P, Kreft H (2016) Multidimensional biases, gaps and uncertainties in global plant occurrence information. *Ecol Lett* 19:992–1006. <https://doi.org/10.1111/ele.12624>
- Nowogrodzki A (2016) Biological specimen troves threatened by funding pause. *Nature* 531:561
- Paquette A, Messier C (2011) The effect of biodiversity on tree productivity: from temperate to boreal forests: the effect of biodiversity on the productivity. *Glob Ecol Biogeogr* 20:170–180. <https://doi.org/10.1111/j.1466-8238.2010.00592.x>
- Phillips SJ, Dudík M, Elith J, Graham CH, Lehmann A, Leathwick J, Ferrier S (2009) Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecol Appl* 19:181–197
- Pichancourt J-B, Firm J, Chadès I, Martin TG (2014) Growing biodiverse carbon-rich forests. *Glob Change Biol* 20:382–393. <https://doi.org/10.1111/gcb.12345>
- Pyke GH, Ehrlich PR (2010) Biological collections and ecological/environmental research: a review, some observations and a look to the future. *Biol Rev* 85: 247–266. <https://doi.org/10.1111/j.1469-185X.2009.00098.x>
- Raymond B, VanDerWal J, Belbin L (2017) ALA4R: Atlas of Living Australia (ALA) Data and Resources in R. <https://www.forge.net/ALA4R/>. Accessed 17 June 2017
- Robertson MP, Visser V, Hui C (2016) Biogeo: an R package for assessing and improving data quality of occurrence record datasets. *Ecography* 39:394–401. <https://doi.org/10.1111/ecog.02118>
- Ruiz-Benito P, Gómez-Aparicio L, Paquette A, Messier C, Kattge J, Zavala MA (2014) Diversity increases carbon storage and tree productivity in Spanish forests: diversity effects on forest carbon storage and productivity. *Glob Ecol Biogeogr* 23:311–322. <https://doi.org/10.1111/geb.12126>
- Serra-Diaz JM, Ninyerola M, Lloret F (2012) Coexistence of *Abies alba* (mill.) – *Fagus sylvatica* (L.) and climate change impact in the Iberian peninsula: a climatic-niche perspective approach. *Flora - Morphol Distrib Funct Ecol Plants* 207:10–18. <https://doi.org/10.1016/j.flora.2011.10.002>
- Serra-Diaz JM, Franklin J, Ninyerola M, Davis FW, Syphard AD, Regan HM, Ikegami M (2014) Bioclimatic velocity: the pace of species exposure to climate change. *Divers Distrib* 20:169–180. <https://doi.org/10.1111/ddi.12131>
- Soley-Guardia M, Radosavljevic A, Rivera JL, Anderson RP (2014) The effect of spatially marginal localities in modelling species niches and distributions. *J Biogeogr* 41:1390–1401. <https://doi.org/10.1111/jbi.12297>
- Sofer MSM (2016) Producing the *Flore D'Afrique Centrale*, past, present and future. *Taxon* 65:937–939. <https://doi.org/10.12705/654.54>
- Sofer MSM, Dauby G, Blach-Overgaard A, van der Burgt X, Catarino L, Damen T, Deblauwe V, Desein S, Dransfield J, Droissart V, Duarte MC, Engledow H, Fadeur G, Figueira R, Gereau RE, Hardy OJ, Harris DJ, de Heij J, Janssens S, Klomberg Y, Ley AC, Mackinder BA, Meerts P, de Poel JLV, Sonke B, Stevart T, Stoffelen P, Svenning JC, Sepulchre P, Zais R, Wieringa JJ, Couvreur TLP (2017) Exploring the floristic diversity of tropical Africa. *BMC Biol*. <https://doi.org/10.1186/s12915-017-0356-8>
- Sousa-Baena MS, Garcia LC, Peterson AT (2014) Completeness of digital accessible knowledge of the plants of Brazil and priorities for survey and inventory. *Divers Distrib* 20:369–381
- ter Steege H, Pitman NCA, Phillips OL, Chave J, Sabatier D, Duque A, Molino JF, Prevost MF, Spichiger R, Castellanos H, von Hildebrand P, Vasquez R (2006) Continental-scale patterns of canopy tree composition and function across Amazonia. *Nature* 443:444–447. <https://doi.org/10.1038/nature05134>
- Stropp J, Ladle RJ, Malhado M, AC HJ, Gaffuri J, Temperley WH, Skoien JO, Mayaux P (2016) Mapping ignorance: 300 years of collecting flowering plants in Africa: 300 years of collecting flowering plants in Africa. *Glob Ecol Biogeogr* 25:1085–1096. <https://doi.org/10.1111/geb.12468>
- Sullivan MJ, Talbot J, Lewis SL, Phillips OL, Qie L, Begne SK, Chave J, Cuni-Sanchez A, Hubau W, Lopez-Gonzalez G, Miles L, Monteagudo-Mendoza A, Sonke B, Sunderland T, Ter Steege H, White LJT, Affum-Baffoe K, Aiba S, de Almeida EC, de Oliveira EA, Alvarez-Loayza P, Davila EA, Andrade A, Aragao LEOC, Ashton P, Aymard GA, Baker TR, Balinga M, Banin LF, Baraloto C, Bastin JF, Berry N, Bogaert J, Bonal D, Bongers F, Brienen R, Camargo JLC, Ceron C, Moscoso VC, Chezeaux E, Clark CJ, Pacheco AC, Comiskey JA, Valverde FC, Coronado ENH, Dargie G, Davies SJ, De Canniere C, Djuiouko MN, Doucet JL, Erwin TL, Espejo JS, Ewango CEN, Fauset S, Feldpausch TR, Herrera R, Gilpin M, Gloor E, Hall JS, Harris DJ, Hart TB, Kartawinata K, Kho LK, Kitayama K, Laurance SGW, Laurance WF, Leal ME, Lovejoy T, Lovett JC, Lukas FM, Makana JR, Malhi Y, Maracahipes L, Marimon BS, Marimon B, Marshall AR, Morandi PS, Mukendi JT, Mukinzi J, Nilus R, Vargas PN, Camacho NCP, Pardo G, Pena-Claros M, Petronelli P, Pickavance GC, Poulsen AD, Poulsen JR, Primack RB, Priyadi H, Quesada CA, Reitsma J, Rejou-Mechain M, Restrepo Z, Rutishauser E, Abu Salim K, Salomao RP, Samsodin I, Sheil D, Sierra R, Silveira M, Slik JWF, Steel L, Taedoung H, Tan S, Terborgh JW, Thomas SC, Toledo M, Umunay PM, Gamarra LV, Vieira ICG, Vos VA, Wang O, Willcock S, Zemagho L (2017) Diversity and carbon storage across the tropical forest biome. *Sci Rep* 7:39102. <https://doi.org/10.1038/srep39102>
- Thessen A, Patterson D (2011) Data issues in the life sciences. *ZooKeys* 150:15–51. <https://doi.org/10.3897/zookeys.150.1766>
- Thompson ID, Okabe K, Parrotta JA, Brockerhoff E, Jactel H, Forrester DI, Taki H (2014) Biodiversity and ecosystem services: lessons from nature to improve management of planted forests for REDD-plus. *Biodivers Conserv* 23:2613–2635. <https://doi.org/10.1007/s10531-014-0736-0>
- Vidal C, Alberdi I, Hernández L, Redmond JJ (2016) National Forest Inventories: assessment of wood availability and use. Springer, Switzerland
- Wildlife Conservation Society - WCS, Center for International Earth Science Information Network - CIESIN - Columbia University (2005) Last of the Wild Project, Version 2, 2005 (LWP-2): Global Human Influence Index (HII) Dataset (Geographic)
- Wiser SK (2016) Achievements and challenges in the integration, reuse and synthesis of vegetation plot data. *J Veg Sci* 27:868–879. <https://doi.org/10.1111/jvs.12419>
- Yang X, Huang Z, Venable DL, Wang L, Zhang KL, Baskin JM, Baskin CC, Cornelissen JHC (2016) Linking performance trait stability with species distribution: the case of *Artemisia* and its close relatives in northern China. *J Veg Sci* 27:123–132. <https://doi.org/10.1111/jvs.12334>
- Yesson C, Brewer PW, Sutton T, Caithness N, Pahwa JS, Burgess M, Gray WA, White RJ, Jones AC, Bisby FA, Culham A (2007) How global is the global biodiversity information facility? *PLoS One* 2:e1124. <https://doi.org/10.1371/journal.pone.0001124>
- Zhang J, Nielsen SE, Chen Y, Georges D, Qin YC, Wang SS, Svenning JC, Thuiller W (2017) Extinction risk of north American seed plants elevated by climate and land-use change. *J Appl Ecol* 54:303–312. <https://doi.org/10.1111/1365-2664.12701>
- Zizka A (2015) speciesgeocodeR: prepare species distributions for the use in Phylogenetic analyses. <https://rdrr.io/cran/speciesgeocodeR/>. Accessed 17 June 2017
- Zizka A, Steege H ter, Pessoa M do CR, Antonelli A (2017) Finding needles in the haystack: where to look for rare species in the American tropics. *Ecography*. <https://doi.org/10.1111/ecog.02192>

**Submit your manuscript to a SpringerOpen® journal and benefit from:**

- Convenient online submission
- Rigorous peer review
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

Submit your next manuscript at ► [springeropen.com](http://springeropen.com)