Fungal Ecology 53 (2021) 101097

Contents lists available at ScienceDirect

Fungal Ecology

journal homepage: www.elsevier.com/locate/funeco

Enhancing repository fungal data for biogeographic analyses

Tianxiao Hao [a,*], Jane Elith [a], Gurutzeta Guillera-Arroita [a], José J. Lahoz-Monfort [a], Tom W. May [b]

[a] The University of Melbourne, Melbourne, Victoria, 3010, Australia
[b] Royal Botanic Gardens Victoria, Melbourne, Victoria, 3004, Australia

ARTICLE INFO

Corresponding editor: Daniel P Bebber

Keywords:
Australia
Biogeography
Biogeographic regions
Citizen science
Data cleaning
Data quality
Fungal biogeography
Fungarium
GBIF

ABSTRACT

Open-access occurrence data are useful for studying spatial patterns of fungi, but often have quality issues. These include errors in taxonomy and geo-coordinates, and incomplete coverage across areas and taxonomic groups. We identify 15 quality issues that can lead to incorrect biogeographic inference, and develop a reproducible pipeline that flags and removes problematic entries. This pipeline tests accuracy of geographic records and names. Then, if information on non-native status is unavailable or unreliable, it detects non-native species via a predictive model. Finally, it identifies spatial and environmental outliers and removes them when biologically improbable. We test the pipeline by cleaning data for Australian fungi, with 251,642 records retained after cleaning the initial 1,034,601 records. Exploratory analysis showed that the cleaned data is useful for analyses such as biogeographic regionalisation, but recording gaps and lack of saturation in collection effort also caution that more surveys are needed to improve collection completeness.

1. Introduction

Open-access repositories of geo-referenced biological data, such as the Global Biodiversity Information Facility (GBIF), are a central resource for the study of biogeography across taxa. For fungi, repository data are especially important because they are often the only source of data applicable for biogeographic studies (Hao et al., 2020; Wüest et al., 2020). Repository data for fungi consist primarily of citizen science observations and digitised information on reference specimens. In some regions, large amounts of these data are now available, and demonstrably useful in large-scale studies of fungal biogeography (e.g., Andrew et al., 2017, 2019; Gange et al., 2019). Although promising new opportunities, repository data also contain undesirable features that compromise their use in biogeography (Thessen and Patterson, 2011). A major issue is that, despite automated quality-control protocols implemented in most repositories, errors in both geo-coordinates and in taxonomy are common (Thessen and Patterson, 2011; Serra-Diaz et al., 2018; Hao et al., 2020). In addition, common characteristics of repository data, such as uneven collection effort across space, can create artefactual patterns, further complicating the use of such data in biogeographical studies (Phillips et al., 2009; Troia and McManamay, 2016; Hao et al., 2020). These problems underline the need for precursory data treatment in studies using repository data (Anderson et al., 2016; Zurell et al., 2020). This step is particularly important in fungal studies (Andrew et al., 2019), because of the likely high prevalence of issues including ongoing taxonomic changes, high proportion of citizen science observations, and the frequent lack of taxonomic curation (due to small number of active experts working on and curating records of fungi compared to plants and animals).

As repository data are recognised as crucial but challenging to use in the study of fungal biogeography, it is important to ask how to explore such data so that their problematic features can be identified and dealt with where possible. While data cleaning is practiced and reported by some researchers using repository fungal data (e.g., Andrew et al., 2017), there is no comprehensive and accessible knowledge base concerning the types of problematic features expected in fungal data, and how to identify and address them. For example, it is unclear what types of automated error-detection tests are applicable to repository data, and how to best integrate taxonomic expert curation in the cleaning process. It is also unclear how collection effort changes over time and space, and how this might affect the completeness of recorded biological patterns. To address these knowledge gaps, in this paper, we (1) develop an

* Corresponding author. Quantitative and Applied Ecology (QAEco) Group, School of BioSciences, Building 2, The University of Melbourne, Victoria, 3010, Australia.
E-mail address: tianxiaoh@student.unimelb.edu.au (T. Hao).

https://doi.org/10.1016/j.funeco.2021.101097
Received 12 January 2021; Received in revised form 8 July 2021; Accepted 15 July 2021
Available online 4 August 2021
1754-5048/© 2021 Elsevier Ltd and British Mycological Society. All rights reserved.

iterative and reproducible data enhancement pipeline, combining script-based error detection, taxonomic matching with external datasets, and validation with taxonomic expertise, and (2) demonstrate a range of exploratory analyses suitable for summarising the main patterns in repository data, in terms of taxonomy, space, time, environments, collection effort, and bioregional patterns. We do so through a case study on enhancing and exploring Australian fungi data. Australia acts as a useful model for wider analyses due to the large variations in climate and ecosystems across the region and because there is an advanced biodiversity infrastructure already in place. We also make available the resulting cleaned dataset of Australian fungi, readily useable for future biogeographic analyses.

2. Materials and methods

2.1. Creating a fungal occurrence dataset for Australia

We report our methods in the present tense to emphasise that we are describing a process that can be done repeatedly at any time, though our results describe the dataset downloaded on 15/10/2019. Fungal occurrence records are gathered from five major repositories containing data on fungi occurring in Australia, namely: Fungimap (https://fungimap.org.au/; data obtained directly from Fungimap Inc.), and from the websites of: MycoPortal (https://mycoportal.org/portal/index.php), the Atlas of Living Australia (ALA; https://www.ala.org.au/; DOI of our download at Atlas of Living Australia occurrence download), the Global Biodiversity Information Facility (GBIF; https://www.gbif.org/; DOI of our download at GBIF.org 2019 Occurence download), and iNaturalist (https://www.inaturalist.org/home). The number of records obtained from each database is presented in Fig. 1; in total there are 1,034,601 records. GBIF and ALA are not fungi-specific databases, so we search for all records that (1) belong to the kingdom of Fungi, (2) are found in Australia, and (3) whose 'basis of record' are 'preserved specimen' or 'human observation'. We select these two categories, defined in the Darwin Core (Wieczorek et al., 2012), as "basis of record" because they are the most abundant and share the same collection process via mostly opportunistic human visitation. Records based on 'environmental DNA' are also common, but they have different features (e.g., usually not opportunistically collected) and harbour different detection and identification issues (e.g., falsely detecting presence from sample contamination; Guillera-Arroita et al., 2017), so we exclude them here. Other 'basis of record' types (e.g., genomic DNA, material samples, living specimen) are also excluded because they are rare and often georeferenced to research facilities. These 'basis of record' choices mean that most species considered in our dataset are lichens and macrofungi, the latter defined as fungi producing visible sporing bodies (i.e. sporophores or fruit bodies; we prefer the term "sporing bodies" over "fruit bodies" as being a technically correct and easily understandable term to replace the plant-centric terminology inherent in "fruit" body). However, a small amount of pathogenic microfungi are also captured, as they can be observed/collected through infected hosts (e.g., the well-known amphibian chytrid fungus *Batrachochytrium dendrobatidis*). In the ALA specifically, we also exclude records from Fungimap since those are separately obtained. iNaturalist is a citizen science observation-specific database, so we search for all 'research grade' Fungi records in Australia. MycoPortal is a fungi-specific data aggregator but not Australia-specific, so we search for all records that are (1) in Australia, and (2) not from the ALA or iNaturalist. Fungimap is both fungi-specific and Australia-specific, so all data are included.

We created a script in the R statistical language version 3.5.1 (R Core Team, 2017), to combine all records into one dataset with the following fields: source repository, original ID, scientific name, state, locality description, coordinate uncertainty in metres, date, longitude and latitude, basis of record, collector/observer, and substrate and habitat. Correspondence between our fields and those in the source datasets are documented in the R script in Hao et al., (2021).

2.2. Data treatment pipeline

Initial careful exploration of the dataset by co-authors TH and TM reveals 15 potential quality issues relevant to individual records or entire taxa that can lead to erroneous biogeographic signals, detailed in Table 1. We pass all records through a pipeline consisting of discrete steps, to detect and flag records or taxon names that potentially harbour such issues. In the following paragraphs, we describe in detail steps in the pipeline, and in Fig. 1, we detail the amount of data retained after each step of the pipeline.

2.2.1. Pipeline step: geographic information

Since all records are geo-coordinated in decimal degrees, we first remove records with missing or invalid (i.e. < -180 or >180 for longitude and < -90 or >90 for latitude) geo-coordinate information,

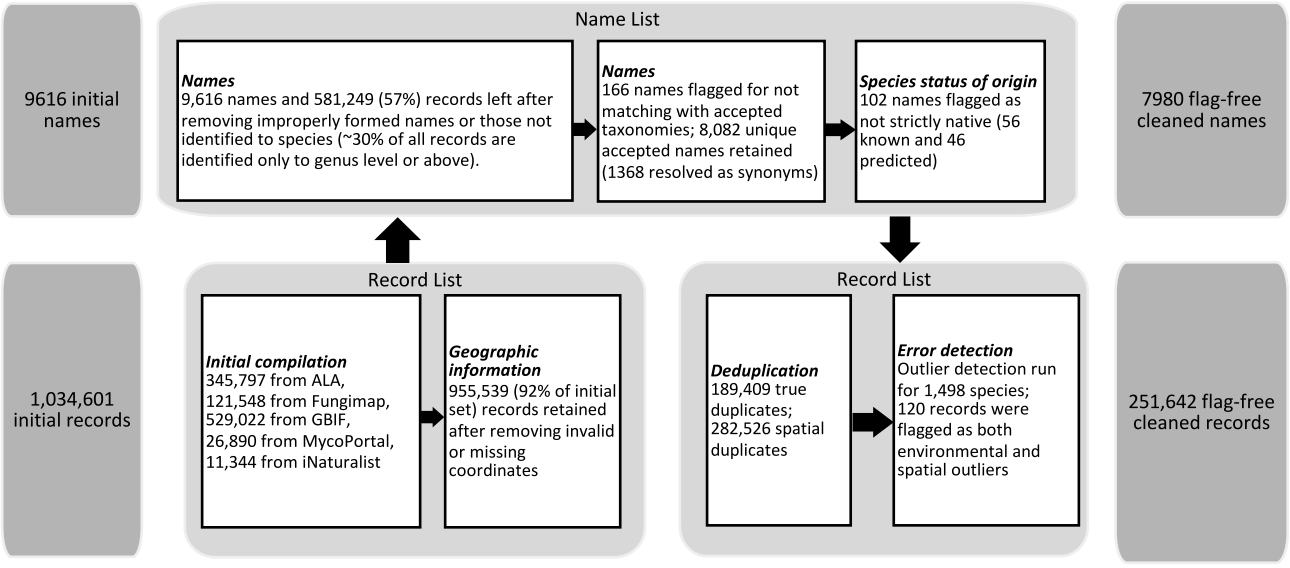

Fig. 1. Overview of the data treatment pipeline (including number of records at each step) beginning with the Record List for which *Geographic information* is required, from which a Name List is generated and cleaned according to *Names* and *Species status of origin* and used to refine the records, followed by *Deduplication* and *Error detection* for the remaining records.

Table 1
Potential quality issues in repository data and respective steps in our data cleaning pipeline designed to address them. As an example of how they may result in erroneous patterns in biogeographic analyses, we tabulate the consequences of such issues if the data were used to cluster areas into bioregions based on species dissimilarity in 100 by 100 km spatial blocks; for details on this regionalisation method see (Laffan et al., 2010; González-Orozco et al., 2014) and section **Biogeographic regionalisation**.

#	Step in pipeline for addressing issue	Quality Issue	Consequence for biogeographic regionalisation
1	*Geographic information*	Invalid or missing geo-coordinate	Record cannot be used in geographic analyses.
2	*Geographic information*	Geo-coordinate outside study region	Record cannot be used in geographic analyses.
3	*Names*	Same taxonomic name used for different species, each with distinct distribution patterns	If distributed across multiple underlying bioregions, these records will erroneously inflate species similarity between bioregions, making it harder to distinguish them.
4	*Names*	Same taxonomic name used for different species with similar distribution patterns	These species may co-occur in the same bioregion – while this may confuse sub-regional clustering, it does not confound distinct signals of the bioregion. If these species are all cosmopolitan, then this issue is similar to issue #8.
5	*Names*	Different taxonomic names (e.g. unmerged synonyms) of the same species used in geographically distinct areas	Falsely increases dissimilarity between bioregions. However, this issue can only result from geographically distinct naming conventions.
6	*Names*	Different taxonomic names of the same species used in same areas	The appearance of the same species under different names falsely inflates diversity, but if the species is indeed restricted to one bioregion, then the unlinked synonyms are also restricted to the same bioregion, not affecting the discoverability of bioregions.
7	*Species status of origin*	Species is introduced	Such species likely do not conform with native bioregion boundaries (e.g. they may be restricted by distribution of exotic host plants), and likely have not reached dispersal equilibrium, thus their records can create noise in biogeographic signals, if mistaken as native species.
8	*Species status of origin*	Species is cosmopolitan (i.e. distributed across bioregions), and unclear whether native	Such species are likely recorded across bioregions, thus they are unlikely to affect dissimilarity between regions.
9	*Species status of origin*	Species is native, but may occur in ruderal, human modified locations (e.g. disturbed soil, or watering) where it would not persist under natural circumstances	'Ruderal' records, if not removed, inflate the estimation of species natural range. This may create an incorrect bioregional signal if multiple species exhibit the same pattern (e.g. growing in urban gardens outside of their range due to watering, but cannot survive in surrounding native areas). However, this can be mitigated by removing all records from cleared/ modified locations.
10	*Deduplication*	Record is duplicated, as a result of appearing in multiple source databases (termed 'true duplicate')	In clustering-based bioregionalisation, species composition in blocks is calculated from species presence or absence (1s or 0s), not abundance, so duplication does not affect the outcome.
11	*Deduplication*	Record shares exact same geo-coordinate with other records of the same species, but is different in dates.	Analogous to issue #10.
12	*Outlier detection*	Geo-coordinate of a record is erroneous and placed outside the true species range	Record appears in wrong location, creating incorrect signals in species composition of blocks. If such error is repeated across a number of species (e.g. if all records from a fungarium are geo-coordinated at the fungarium's location), then false bioregion patterns may emerge.
13	*Outlier detection*	Geo-coordinate is erroneous (including cases where the coordinate is imprecise) but placed within the true range	Record still appears in wrong location, but contributes to an 'expected' species composition profile of the bioregion, so such errors should not affect bioregion patterns.
14	*Outlier detection*	Species in record is misidentified as another species with different distribution	Analogous to issue #12.
15	*Outlier detection*	Species in record is misidentified as another species with similar distribution	Analogous to issue #13.

and correct records with inverted latitudes (i.e., where southern hemisphere is not represented as a negative latitudinal value). We then reproject all longitude and latitude coordinates to GDA94 Australian Albers equal area projection – this is necessary for binning data into equal-sized spatial blocks in later analyses. Then, using a coarse raster representation of Australia at 100 km resolution, we remove records whose coordinates fall outside the raster (i.e., far away from Australia). To further correct for records that may incorrectly appear just off-coast due to coordinate imprecision, we use a 1 km resolution Australia raster to remap all oceanic coastal records to the nearest adjacent terrestrial cell (using the 'biogeo' package in R; Robertson, 2016). We also remove a minor number of sensitive species records whose coordinates as published in the repositories are known to be manipulated to hide locations.

2.2.2. Pipeline step: names

We first remove all records with improperly formed names (e.g. missing genus or species name, or name containing numerical characters) or those with taxon rank higher than species. We also remove some closely related species that cannot be confidently separated in the field on current knowledge (e.g. those in the genus *Trametes* formerly referred to *Pycnoporus*). Then, we tabulate all unique names left in the dataset, creating a master 'name list'. To check the validity of names and update any synonyms in the name list to their up-to-date accepted name, we match all names in the name list against two databases of names: the fungi and lichen components of the Australian National Species List (NSL), and the Catalogue of Life (CoL) accessed via GBIF. The NSL and CoL provide an accepted name for all names along with higher taxonomy. These databases allow us to check whether the names in our name

list are accepted, or if they are a synonym of an accepted name, in which case they are updated with the corresponding accepted name from the matched database. Where NSL disagreed with CoL, CoL names are kept because CoL contains more recently recognised/described species. For names at infraspecific rank (such as variety) that match neither database at first, we attempt rematch after removing infraspecific epithets. Names still matching neither of the databases are checked by co-author TM, who identifies and corrects names that are spelling errors or synonyms not yet trapped in NSL or CoL. The remaining non-matching names, which consist largely of manuscript names, non-fungi, or improperly formed names, are flagged for removal. In the name list, we also add higher taxonomy (i.e. phylum, class, order, and family; from either NSL and CoL, according to the source of the name) and guild, trophic mode, and growth form information. Information on guild, trophic mode, and growth form is obtained by matching names and their higher taxonomies against the FUNGuild dataset (Nguyen et al., 2016) – this is important for e.g. distinguishing lichenised and non-lichenised fungi.

2.2.3. Pipeline step: species status of origin

Species that are not native to Australia or native but displaying ruderal distributions may confuse bioregional signals and should thus be removed (see Table 1 for explanations, also see Pyšek et al., 2004). Because there is no checklist specifying which fungi are native to Australia that can be employed to match our species against, we can only assess origin status in our dataset using expert knowledge of TM, which can practically only be done for a small number of species due the size of the name list. This issue of unknown species origin is common for fungi across the globe, and the determination of species origins is likely a common need for users of repository fungi data. While the establishment of species origin databases is the necessary long-term solution, here we also explore if patterns observable in data can allow rapid sorting of species origins using a statistical learning model (Hastie et al., 2009). The basis of this idea is that species origin status correlates with species habitat preference (namely that native species are found more often in native habitat and exotic species more often in modified habitat), and this relationship is likely reflected in the observed frequency of species occurrence in different habitats. Therefore, we can train a model to recognise the relationship between species origin and occurrence frequency in different habitats, and use the model to predict species with unknown origins. In Appendix A, we describe in detail our implementation of a multi-class random forest model (Cutler et al., 2007; implemented through the 'randomForest' package in R; Liaw and Wiener, 2002) for this prediction task. In summary, for each species we tabulate the percentage of records occurring in relevant land use and vegetation classes, then use those as predictor variables in a model where the response is the known native/exotic status of the species. We train and cross-validate the model using data for 483 species. When cross-validating model predictions against known origins, the model achieves good accuracy at predicting exotic species, with ~0.8 sensitivity and perfect specificity (Hastie et al., 2009). Using this model, we predict the origin status of all species whose status is unknown, and flag for removal those that are (1) predicted to be exotic, (2) not lichens, and (3) having more than 50 records. We do not flag lichens because there are very few documented records of exotic lichens in Australia; we also do not flag species with <50 records because it is difficult to establish their habitat preference with just a few records.

2.2.4. Pipeline step: deduplication

We flag records that are duplicates of another record, either as 'true duplicates' (same species, same coordinates, same date), or as 'spatial duplicates' (same species, same coordinates, but different date). For this step, we use the original decimal degree coordinates rather than the reprojected coordinates. This step is important since some of our data sources are aggregators that combine other data sources, thus duplication is common.

2.2.5. Pipeline step: outlier detection

Some errors (e.g., misplaced coordinates, misidentification) are difficult to detect automatically, but manually checking every record is also impractical. Here, it is useful to prioritise records more likely erroneous, e.g., spatial and environmental outliers for each species. Outliers are worth detecting because: (a) outlierness can imply biological improbability, and (b) erroneous but inlying records will likely not contribute greatly to biogeographic signals (Table 1). To detect spatial outliers, we calculate the Local Outlier Factor (LOF; Breunig et al., 2000) for each record using the 'DDoutlier' R-package (Madsen, 2018), then flag any record whose LOF is more than five standard deviation higher than the species-mean LOF. Rather than comparing the coordinates of a record to the central mass of species records, LOF measures how isolated a record is compared to its closest neighbours in terms of Euclidean distance. This neighbourhood-based approach is desirable for our application, because records without close neighbours are most worth checking – these are most biologically improbable but if proven valid they could signal undiscovered populations. To detect environmental outliers, we use the reverse jackknife method (using the 'biogeo' R-package; Robertson, 2016) to flag outlying records in terms of four climatic variables: annual averages and seasonality of temperature and Moisture Index. These variables and reasons for using them are described in detail in Appendix B. We use the reverse jackknife method because it is consistent with GBIF and ALA quality checks, which also report this approach to be successful at detecting erroneous outliers (Chapman, 2005). We perform spatial and environmental outlier detection for species with more than 30 records (1498 species), since we cannot establish outlierness for rarer species. We initially also attempt to validate the outliers through a supporting specimen or photograph, but such information is not readily available for most records.

2.3. Using the pipeline to enhance data

We use the pipeline to treat the fungal occurrence dataset, with the following steps. We first build a cleaned name list by excluding: species not matched with name databases (n = 166), species known to be exotic (n = 56), or flagged with exotic prediction (n = 46); this results in 7980 issue-free names. We next build a cleaned dataset by first excluding any record not matching with the cleaned name list (n = 27,874), followed by spatial duplicates (n = 270,952), records occurring at least 100 m away from native vegetation (n = 30,661; inferred by the native vegetation layer described in Appendix A), and records flagged as both temporal and spatial outliers (n = 120). The resulting cleaned dataset contains 251,642 unique records – while this is only 24% of the original 1,034,601 records downloaded, most of the removed records are identified at genus level or above, or are duplicated copies from different databases.

2.4. Exploring the cleaned dataset

We next analyse the cleaned dataset to reveal trends and patterns. We first focus on the breakdown of records and unique species over trophic guilds, basis of record, and time. Then, by binning records into 100 by 100 km spatial blocks (875 in total over Australia), we create sampling 'blocks' and summarise the number of species and records over blocks. We also construct a species accumulation curve (Gotelli and Colwell, 2001) over all blocks (via the 'vegan' package in R; Oksanen et al., 2019) and, for the block with most records, a species accumulation curve over time. We next explore record coverage over nine environmental variables relating to climate and soil properties, detailed in Appendix B. We first extract environmental conditions from all 100 km blocks and ordinate them in an environmental hyperspace generated by a principal component analysis (PCA). This allows us to examine the environmental overlap between blocks with records and blocks without, and discover if some unique environments are not yet recorded. To further map potential unrecorded environments at a finer resolution, we

next build a Multivariate Environmental Similarity Surface (MESS). MESS calculates the similarity between each cell in the environmental rasters and a reference set of environmental conditions at locations with records; this is useful for highlighting raster cells outside the environmental range of the reference dataset (Elith et al., 2010). We build MESS with a 5-km resolution raster stack of the aforementioned nine variables. Next, we explore collector-related patterns in the dataset. We first assess the relationship between collection locations and accessibility to cities, as the latter is well-known to affect the behaviour of opportunistic collectors (Mair and Ruete, 2016). Accessibility of collection locations to cities is extracted from a 1 km raster of estimated travel time to the closest city, produced by Weiss et al., (2018). We also extract 'baseline' travel times from 30,000 random points sampled across Australia, which we compare collection locations against. Finally, we also assess the number of records made by individual collectors, and map the spatial footprint of the most prolific collectors.

2.5. Biogeographic regionalisation

As a further exploration on how the cleaned data can be used to uncover biogeographic patterns, we use the data to infer biogeographic regions (defined as regions sharing similar species composition; Cox, 2001; Mackey et al., 2008) using a clustering method (González-Orozco et al., 2014). We choose this analysis as a demonstrative use of our data because, while fungal bioregions of Australia have been suggested (May, 2017), analytical approaches explicitly for mapping bioregions have not been applied previously. Using the software Biodiverse v.3.10 (Laffan et al., 2010), we first bin the data into 100 km blocks and construct a species-block matrix. Here, we exclude species occupying only one block and blocks containing only one species – these create noise interfering with the clustering method, and are not informative to the aim of finding contiguous groups of blocks with similar species composition. 543 blocks and 3796 species were retained after this filtering. Using the Biodiverse built-in workflow, we calculate pairwise block-by-block species composition dissimilarity using Simpson's beta diversity index, or S2 (Lennon et al., 2001), and then use the Weighted Pair Group Method with Arithmetic mean (WPGMA) clustering (Sokal, 1958) to group blocks accordingly. The results of clustering analyses are visualised both on a dendrogram and on a map with colour-coded clusters, using Biodiverse. We then examine whether we could hypothesise clusters as bioregions, based on their spatial contiguity, and their distinctness from other clusters on the dendrogram.

3. Results

3.1. Data description

The distribution of species in the cleaned dataset across trophic

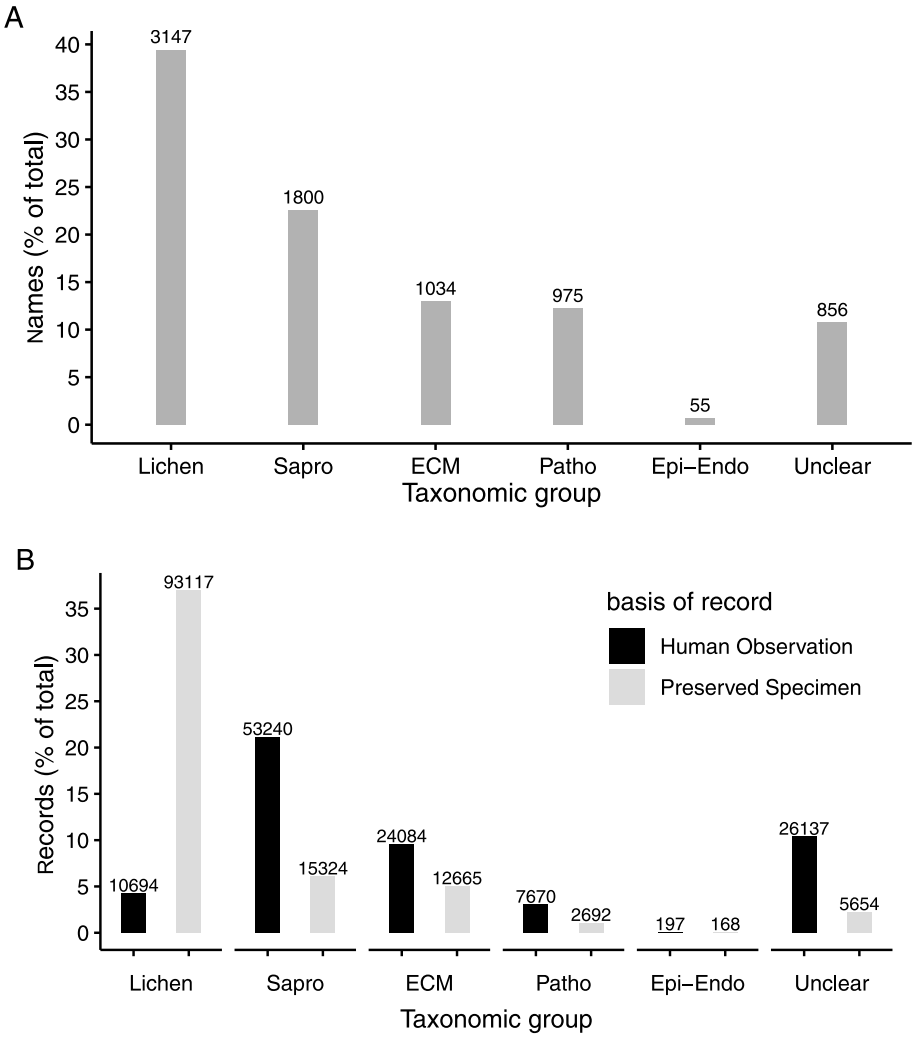

Fig. 2. Distributions of (A) taxonomic names and (B) records of Australian fungi across trophic guilds (Lichen = Lichenised; Sapro = Saprotrophic; ECM = Ectomycorrhizal; Patho = Pathogenic & Parasitic; Epi-Endo = Epiphytic & Endophytic; Unclear = not assigned to a single guild according to FUNGuild). Floating numbers indicate counts of names or records.

guilds shows that lichens are most common both in numbers of species and records (Fig. 2). Saprotrophic (feeding on decaying organic matter) and ectomycorrhizal (plant symbionts) fungi are also commonly recorded, followed by pathogens/parasites. Pathogens/parasites include both macrofungi such as *Tremella fuciformis, Cyttaria gunnii*, *Chondrostereum purpureum* and *Marasmius crinis-equi* (the former two are classified as parasites according to FUNGuild and the latter two as pathogens), and microfungi such as *Batrachochytrium dendrobatidis*. Endophytes include macrofungi such as *Xylaria castorea* and *Annulohypoxylon truncatum*. Species not designated to a single guild according to FUNGuild are designated as 'unclear'.

Lichens are unique in that their records are predominantly based on preserved specimens, whereas other fungi are commonly recorded by observations (Fig. 2B). These two collection modes also show different historical patterns, with specimens being more common in the past but observations rapidly taking over since the 1990s (Fig. 3A). Specimen collection-per-year has also decreased, although this could be explained to some extent by time-lag in digitising recent specimens. In terms of seasonality, non-lichen records are more common in the colder winter months, likely because these are the sporing seasons for many macrofungi, particularly in the well-recorded temperate south. In contrast, lichen records are mostly uniform across months (Fig. 3B), as most lichens have persistent sporing bodies that are observable throughout the year.

In terms of spatial distribution, records aggregate towards coastal areas of Australia, or coincide with road infrastructure inland (Fig. 4). The island state of Tasmania has particularly high abundance of records (Figs. 4 and 5). This could partially be explained by exceptionally prolific volunteer collectors known to be active in the area. However, it is also likely that Tasmania is a natural fungal diversity hotspot due to its wet and temperate climate and large tracts of intact remnant vegetation.

Binning records into 100 by 100 km blocks revealed that recording intensity is highly variable across space (Fig. 5). Records are most abundant near Melbourne (capital city of Victoria) or Hobart (capital city of Tasmania), and the block with most records (n = 21,841) coincides with Melbourne. In contrast, 73 blocks contain only a single record each, and 168 blocks, mostly in the dry interior or the tropical north, contain no records at all. Although we also expected concentration of records in national parks, such patterns are not visible on a continental scale.

In general, blocks containing more records also contain more species. The number of species increases with the number of records in a broadly logarithmic shape, but becomes more variable among blocks at higher record numbers (Fig. 6). The relationship does not appear to approach an asymptote suggesting that even the most visited blocks are far from saturation in species discovery.

The curve of species accumulation over all blocks also exhibits log-like behaviour and is not plateauing for high number of blocks (Fig. 7A), indicating that new species are being discovered as more blocks are visited. For the block with the most records (geographic location shown in Fig. 6), species accumulation over time also does not plateau (Fig. 7B), indicating that new species keep being discovered

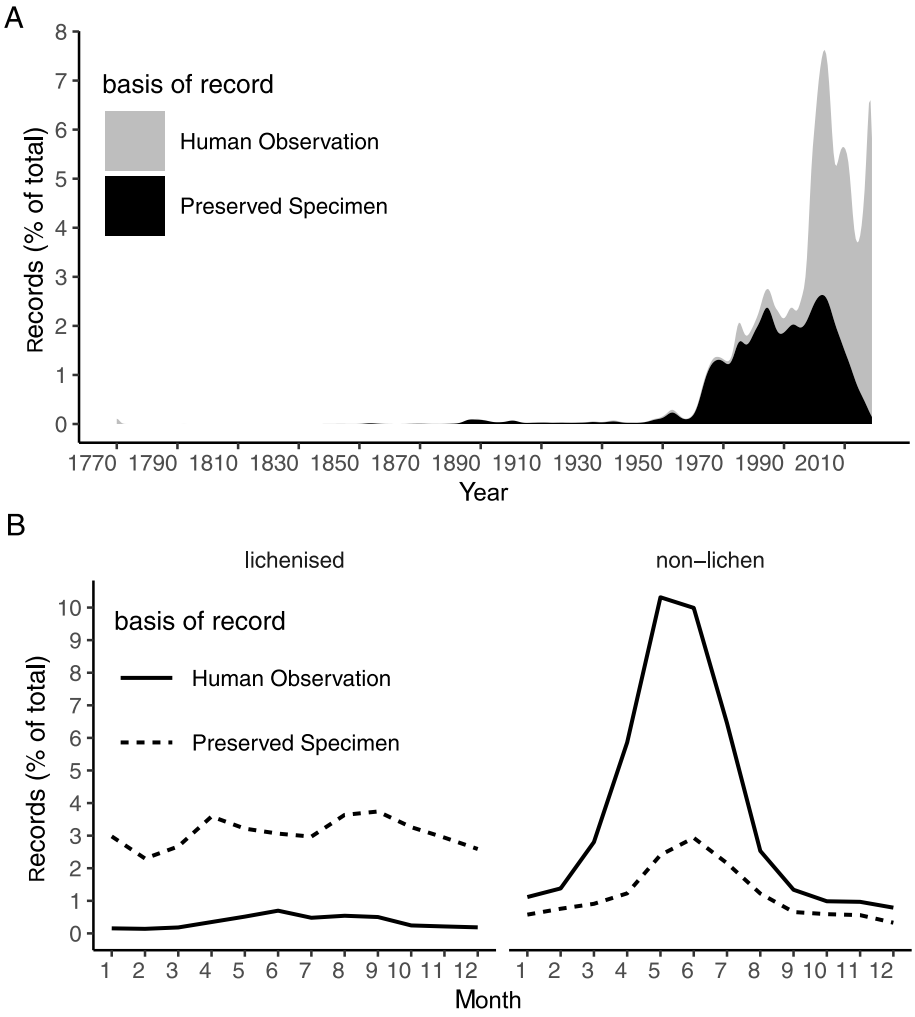

Fig. 3. (A) Percentage of records of Australian fungi over years, the density area is shaded according to proportional basis of record. (B) Percentage of records over months.

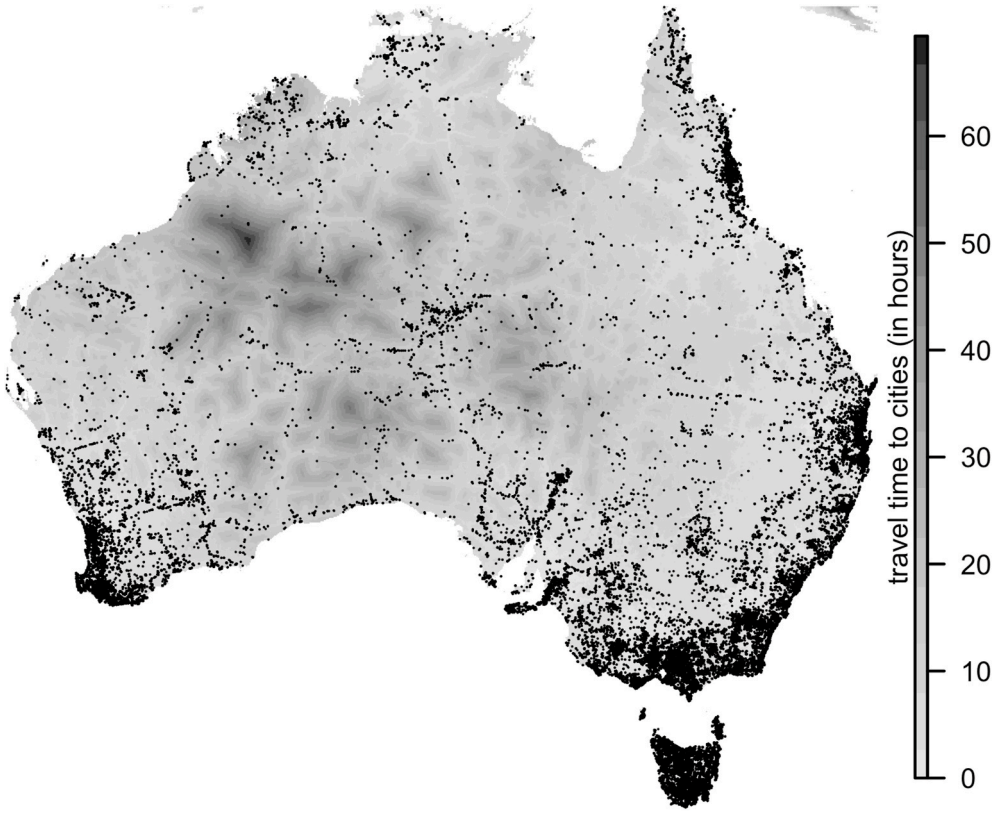

Fig. 4. Locations (black dots) of 251,642 records of Australian fungi in the cleaned dataset, overlaid on top of travel time to closest cities (in hrs) in Australia (scale shown on right).

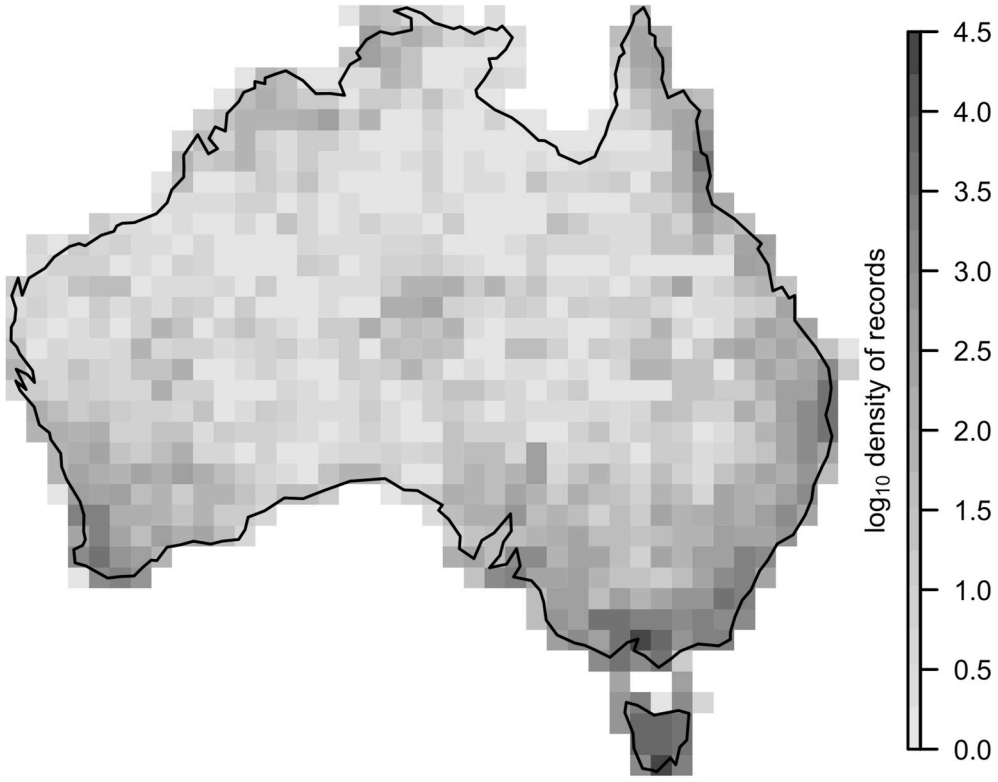

Fig. 5. Log_{10}-transformed count of records of fungi in the cleaned dataset, in 100 by 100 km blocks across Australia.

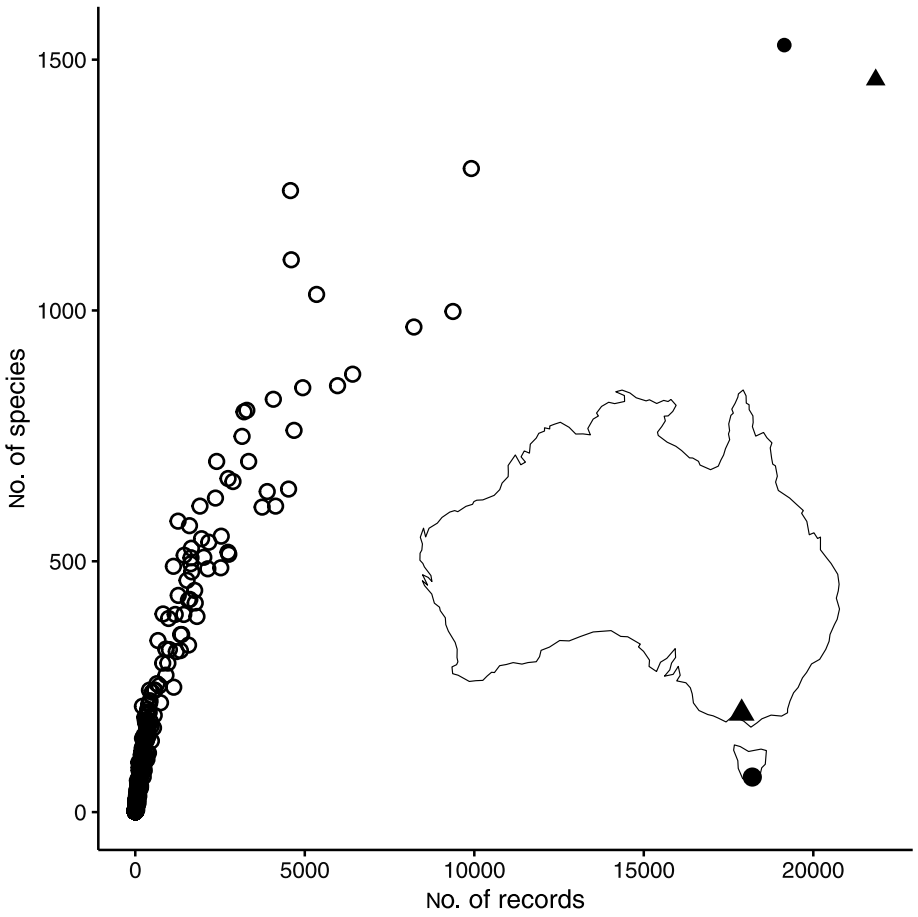

Fig. 6. Number of species versus number of records for each 100 by 100 km block in the cleaned data. The block with most records (filled triangle) and the block with most species (filled circle) coincide with Melbourne and Hobart respectively – their locations in Australia are shown.

even in well-recorded blocks. While these trends could relate to rare species being discovered only with increased collection effort or new species being described over time, they also suggest an overall lack of saturation in collection effort across and within blocks, and that no blocks should be treated as entirely known (all species catalogued).

We next explore collection coverage across environments. Although we initially suspect that some environmental conditions such as the dry deserts of interior Australia may be under-recorded, PCA ordination of blocks with no records vs blocks with at least one record in environmental hyperspace reveal that blocks with no records largely overlap environmentally with blocks with records, in terms of the environmental variables considered (Fig. 8A; for details on environmental variables see Appendix B). The Multivariate Environmental Similarity Surface analysis also finds that most of Australia is within the environmental range of existing records (Fig. 8B). The negative areas on the MESS map highlight locations dissimilar to existing record locations, including: small pockets in central Australia (drier), western Tasmania (wetter throughout the year, more organic content in soil), and scattered discrete pockets in inland Australia with different soil attributes (e.g. the very small strip in mid-eastern Australia which has high soil Phosphorus).

As suggested by our earlier observations on Fig. 4, collection pattern strongly correlates with travel time to cities. About half of all collection locations are within 2-h travel from cities, whereas less than 5% of randomly sampled locations in Australia are within the same distance. Records based on human observations also exhibit somewhat stronger affinity to cities than those based on preserved specimen, but both share generally similar patterns (Fig. 9).

Finally, we report on individual collector patterns. A total of 7483 unique collector names or ID numbers are discovered (4% of records have no collector information). Collectors contribute 34 records on average, but a small number of prolific collectors are highly influential, with the ten most prolific collectors contributing about half of all records – in Fig. 10 we map the collection footprints of some of these prolific collectors. Prolific collectors record most often around one area (likely a place of permanent or long-term residence), but many also record across their state or region (e.g., records for Pamela Catcheside & David Catcheside significantly diminished beyond the eastern border of South Australia; and for Gintaras Kantvilas and Genevieve Gates & David Ratkowsky records are predominantly from Tasmania, see Fig. 10). Prolific recorders may also be active in places far from home, often around major cities or along roads (Fig. 10).

3.2. *Biogeographic regionalisation*

The cleaned data can be divided into clusters that represent areas sharing similar species composition. In Fig. 11 (displaying the 9 highest level clusters based on dendrogram node length), we can observe two large contiguous clusters: southern coastal (orange) and central arid (blue). These are biologically sensible candidates of bioregions because they share similar shapes and boundaries with bioregions for other taxa, and because they correlate with climatic differences likely explaining species composition differences (Ebach, 2012). Smaller contiguous clusters are also observable, particularly along the east coast, and in the dry tropical northwest of Australia. These are promising clues to bioregions, but the lack of collection effort in these areas limit our ability to clearly map bioregional geometry, or to compare them to existing bioregions of other taxa. While preliminary, our bioregionalisation exercise reveals that cleaned fungal repository data are useful for uncovering

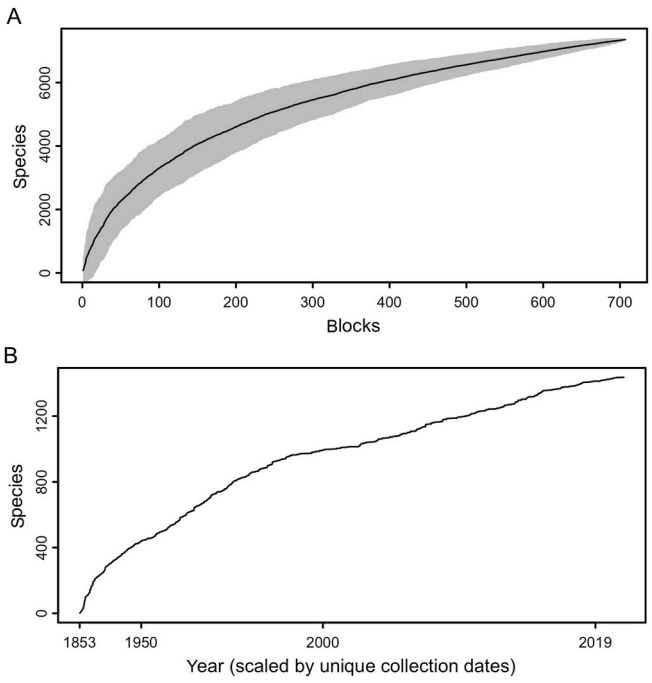

Fig. 7. (A) Species accumulation curve over all 707 blocks with records. The grey band represents variation from randomising the order at which blocks are added to the calculation 100 times, and the black line indicates the central trend. (B) Species accumulation curve for the block with the most records. Species accumulation is calculated by adding chronologically-ordered collection dates. The x-axis is scaled according to unique collection dates by years in which collecting occurred, reflecting that collection happened more frequently in recent years.

biogeographic patterns, but variation in collection effort means that the biogeographic information of some areas remains unclear or unknown.

4. Discussion

Our data cleaning results show that repository data for fungi indeed contain many quality issues, but these issues can be ameliorated by the use of a data curation pipeline and involvement of taxonomic expertise. Many issues in our data, such as unlinked synonyms, and unknown status of origin, relate to somewhat limited taxonomic curation in repository datasets and the lack of region-specific species checklists, thus our results provide support for the importance of these endeavours. Moreover, we find that automating filtering with databases providing accepted taxonomies and nomenclatures do not resolve all taxonomic issues, and active expert curation of taxonomy can help recover/retain taxonomic names rejected by the automatic process, and spot issues not detected by automatic filtering, such as closely related species that are unidentifiable in the field on current knowledge. This highlights that, when cleaning data for taxa with frequent taxonomic changes or many newly described species, one cannot solely rely on automated pipelines targeting geographic issues (e.g. Serra-Diaz et al., 2018; Zizka et al., 2019) – expert-curated taxonomic filtering is also essential. Finally, we find that predictive models are promising for detecting non-native species, when such information is limited in a species dataset. This is a new methodology, and may be of broad interest. It should be explored further, with rigorous tests of model behaviour and performance. It would be enhanced by involvement of more experts and testing on data from other regions, to both provide more data for training the model, and to create some independent testing datasets.

Our cleaned dataset reveals contrasting patterns between specimen-based and observation-based records. Consistent with studies on other taxa (Speed et al., 2018), specimen-based records are older, whereas observation-based records are more recent and growing rapidly. This is

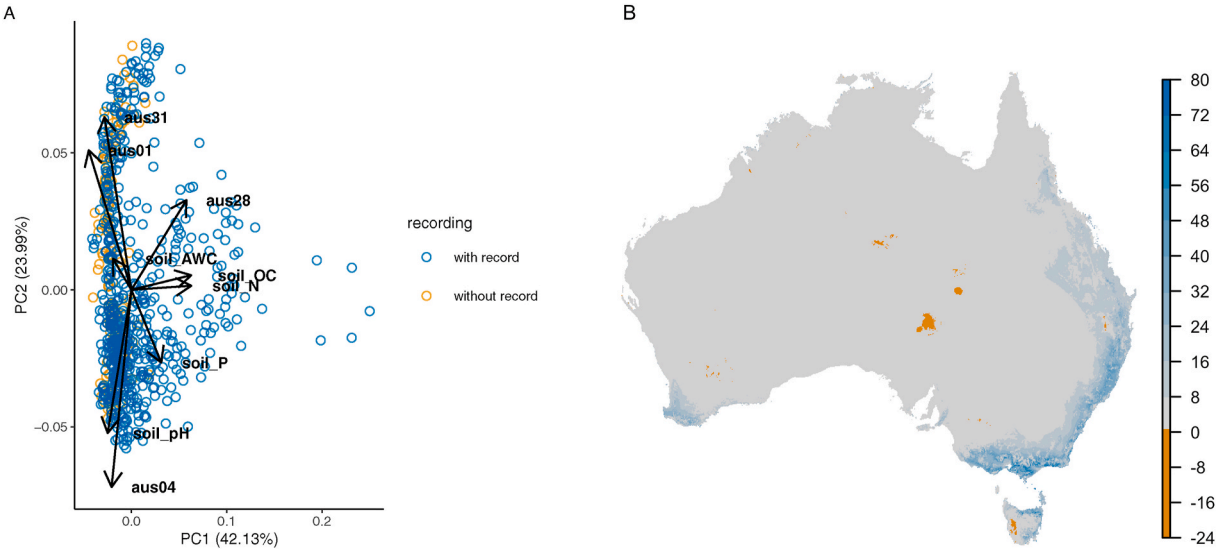

Fig. 8. (A) Ordination of blocks with no records vs blocks with at least one record over the first two axes of Principal Component Analysis (PCA) transformed environmental space; loadings of environmental variables for the first two axes used are shown, for explanations of these variables see Appendix B. (B) Multivariate Environmental Similarity Surface (MESS) for the same environmental variables. The MESS map compares the environment of a given location against the reference environmental range of all the locations with records. Positive MESS values mean the location is inside the reference environmental range, and negative values mean the location is outside the reference range of at least one environmental variable. The range of values correspond to the degree of environmental similarity/dissimilarity to the reference environmental range. Areas with negative values are of interest, since they represent environments unencountered by existing locations with records. The scale of MESS values is shown on the right, with orange values highlighting regions of negative MESS value. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

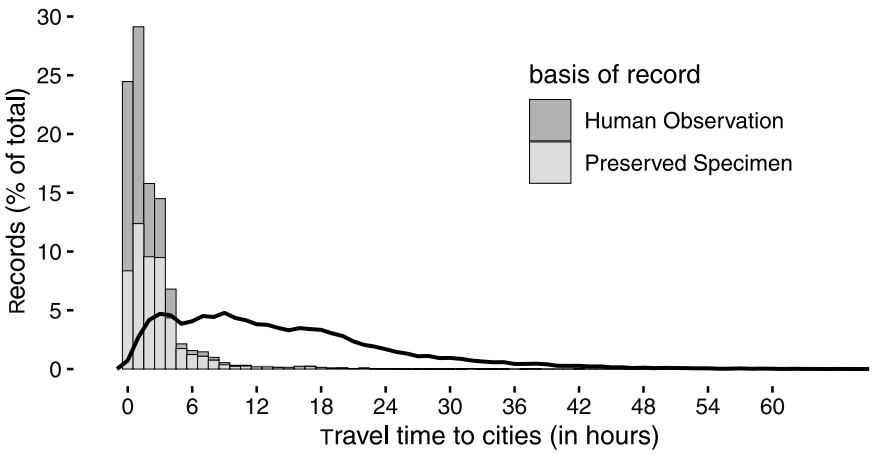

Fig. 9. Histogram of locations with records over travel time to cities in Australia. Each bar corresponds to 1-h travel time and is shaded according to proportional basis of record. The black curve represents background travel time to cities from random locations in Australia.

likely explained by technological advances allowing citizen science observers to submit observations digitally with increasing ease. The two collection modes also correlate strongly with taxonomy (lichens – specimens, non-lichens – observations), likely explained by the difficulty of preserving macrofungal sporing body specimens, the difficulty of field identification for many lichens (well-defined keys exist for lichens, but they often involve microscopic and chemical characters and are more amenable to be identified by experts), and stronger citizen science interest in macrofungi compared to lichens. The lichen/non-lichen dichotomy also impacts the seasonality of specimen versus observation records (Fig. 3B) , because most observation records are made during ephemeral macrofungal sporing seasons. Finally, specimen collectors also tend to be professionally employed (typically lichenologists), whereas observers include many more citizen scientists. These contrasting patterns suggest that professional specimen collection for macrofungi is historically rare due to the challenging nature of the task, but emerging citizen science contributions substantially increase data availability, and provide important support for research on under-collected taxa such as macrofungi (Andrew et al., 2017, 2018; Wüest et al., 2020).

Through our data exploration, we also identify the locations of collection gaps in terms of space and environments. Spatial collection gaps are observed throughout the dry interior and tropical north (Figs. 4 and 5), likely explained by remoteness and the lack of road access making them difficult to visit for opportunistic collectors. In contrast with the patchy spatial collection coverage in central and northern Australia, environmental collection coverage in these areas is more complete, with only small patches of unrecorded environments (Fig. 8). This is an important sign that, although many locations are unvisited, they are environmentally similar to visited locations. This suggests that species occurrence in these unrecorded locations can be interpolated with tools such as species distribution models (Elith and Leathwick, 2009). Nevertheless, such endeavours still require adequate amount of data to accurately model species-environment relationships (Guisan et al., 2017), so it is still important to collect more data in the under-recorded central and northern Australia. Interestingly, western Tasmania also contains unique unrecorded environments (Fig. 8), despite high collection effort in other parts of Tasmania. This shows that collection gaps are not necessarily far from areas with records, and that exploring data coverage in environmental dimensions is important for revealing details overlooked by spatial explorations. Finally, an important caveat is that our observations on environmental coverage are based on a set of environmental variables believed to be generally relevant for fungi (e.g. a similar set of variables were used to model fungal productivity in Morera et al., 2021), but different collection gaps likely exist across different environmental gradients. Data users interested in exploring different environmental patterns should also test the coverage of their chosen environmental variables, using tools such as MESS maps.

The widespread spatial recording gaps and the lack of species saturation across all blocks mean that more collection is needed both for discovering new species even in places with many records, and for exploring under-recorded regions. As citizen science observations are now the most important contributors of new opportunistic data (Fig. 3A), we could benefit substantially from improving citizen science collection protocols to maximise the usefulness of future data. For example, methods are available for directing citizen scientists to areas of highest interest, using a marginal value framework (Callaghan et al., 2019a, 2019b). Such efforts of coordinating target locations with citizen scientists are especially applicable to our Australian dataset, because there are many spatial gaps to fill, and because prolific naturalists and scientists are known to be willing to travel far and wide to make observations and collections (Fig. 10). In addition, the quality of citizen science data can be enhanced by improving identification skills of involved volunteers, through e.g., organising identification workshops, publishing guides and lists of target species, and providing logistical and material support. In particular, lists of target species can focus on species identifiable in the field, and warn observers about morphologically indistinguishable species whose identification require microscopy or sequencing. Finally, specimen collection can also aid citizen science by providing voucher material for validation, particularly for outlying observations far from confirmed specimens – such outlying records are suspicious, but they could also signal previously unknown populations.

In addition to citizen science, professional collection is still required, particularly for locations rarely visited by opportunistic volunteer collectors, and species difficult to detect or identify by volunteers. For such efforts, environmental DNA (eDNA) sampling of soil is a relevant method – in Australia, a large amount of eDNA data is already available through the BASE project (Bissett et al., 2016). Unlike observation and specimen data, eDNA recording has the ability to record the presence and absence of a large number of unique sequences from each sample, and can detect species undetectable from sporing bodies (not sporing, sporing very rarely, or sporing bodies hard to detect, i.e., truffles). However, eDNA-specific issues, including false presences from dead genetic material, merged or split species in processed operational taxonomic units, and lack of sequenced reference specimens to identify sequences against, could lead to erroneous biogeographic inferences, so the suitability of such data for studying biogeography needs to be thoroughly explored (Hao et al., 2020).

Although links between fungal assemblages and bioregions have been explored (Cassis and Laffan, 2017), our biogeographic regionalisation appears to provide the first analytical evidence for the

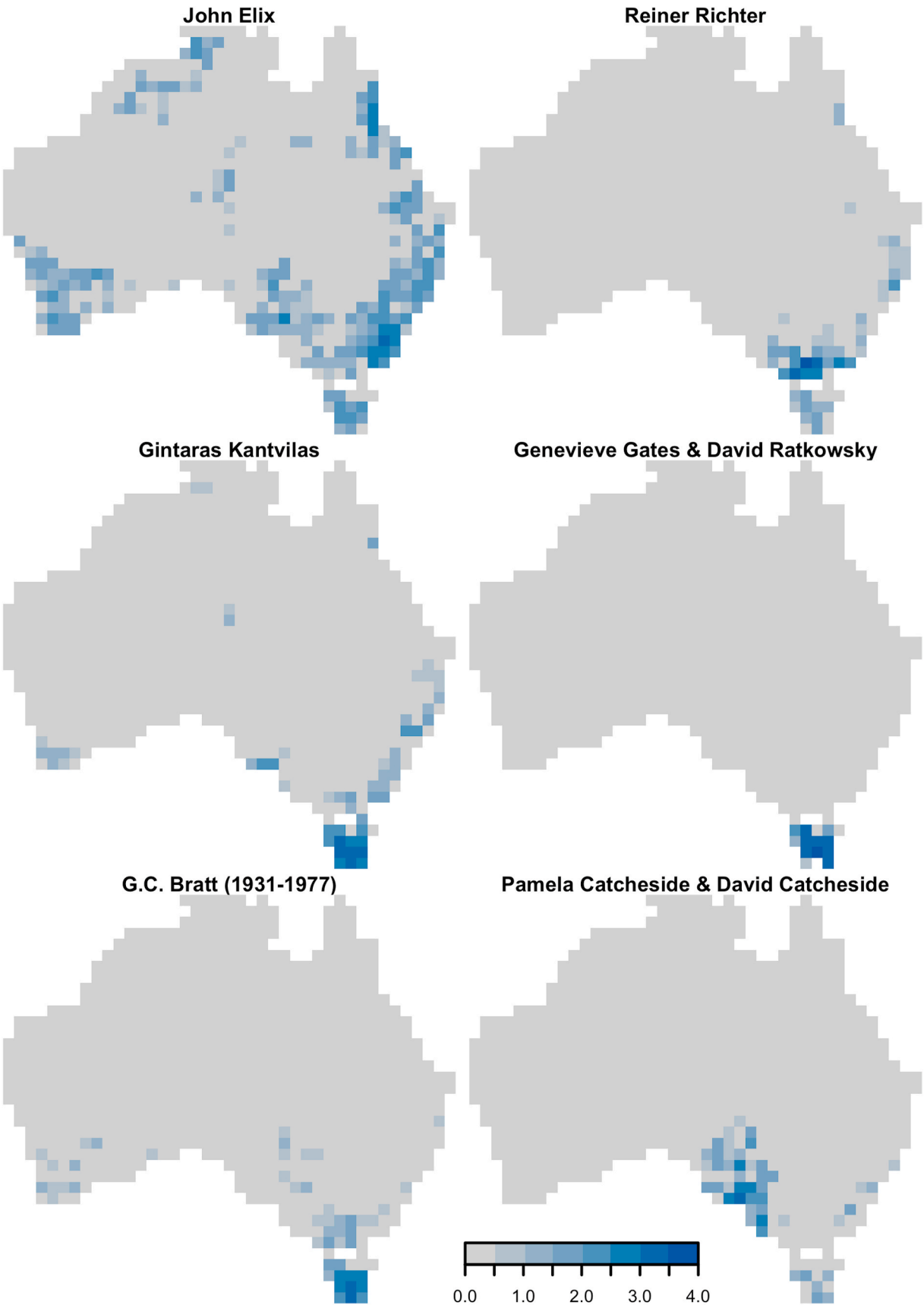

Fig. 10. Log_{10}-transformed record density in 100 by 100 km blocks for a selection of prolific collectors.

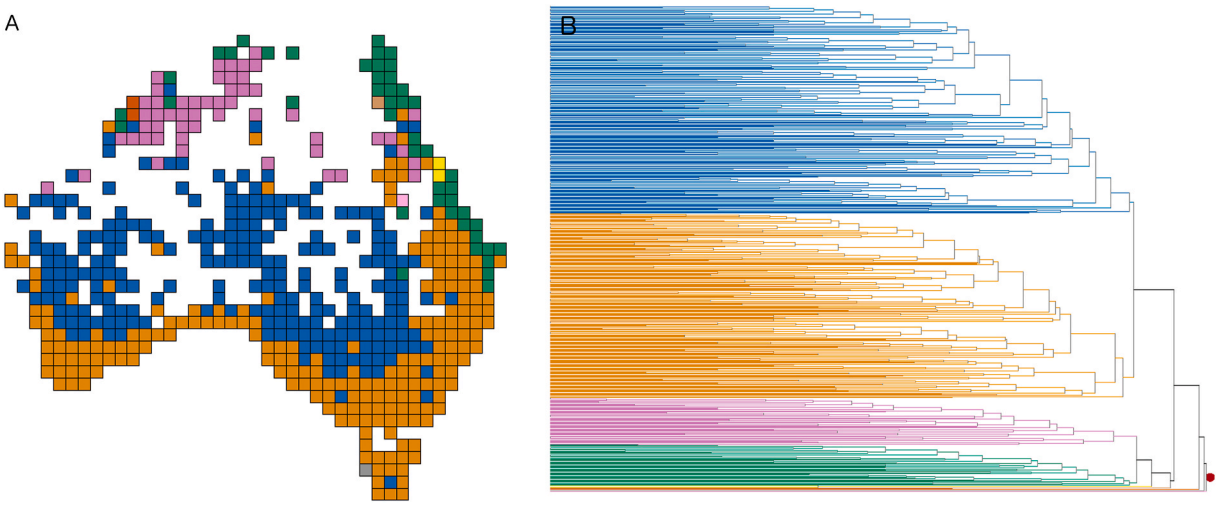

Fig. 11. Biogeographic regionalisation for Australian fungi based on cleaned data, using 100 by 100 km blocks. The blocks are clustered using WPGMA clustering based on species composition dissimilarity between blocks as measured by the Simpson's beta index. Both the (A) map and the (B) dendrogram are coloured to display the 9 highest level clusters, identified based on node length in the dendrogram.

existence of fungal bioregions. This is an important first step towards understanding continental-scale fungal biogeography in Australia and elsewhere. The shape and boundaries of fungal bioregions deserves further exploration, to enable research questions such as comparing fungal bioregional boundaries with those of other taxa (e.g. plants in González-Orozco et al., 2014). While our preliminary analysis using a clustering-based method could only determine regional geography in record-abundant portions of Australia, more complex regionalisation methods based on predictive statistical models may interpolate better in record-poor areas. Compared to clustering methods based only on species composition in blocks, model-based approaches, such as finite mixture models (Dunstan et al., 2011), also use environmental variables as inputs, to model the relationship between species occurrences and the environment. These models are then able to interpolate species composition profiles in unrecorded areas by predicting species occurrences as the response to local environment (Woolley et al., 2013). The suitability of such methods for our dataset is further supported by the good environmental coverage of our data (Fig. 8). However, because such methods are more complex than simple clustering (e.g. need to find appropriate environmental variables for predicting fungal occurrence across the continent, and the use of specialised multivariate modelling techniques), we will explore them in future works.

In conclusion, through our case study of Australian fungi, we showed that repository data need to be and can be cleaned, and we developed data cleaning scripts and methodologies applicable for use in any geographic region. The cleaned data in this study revealed interesting biological patterns across space, time and species, and also supported preliminary hypotheses on Australian fungal bioregions. Overall, our result provides support for the usefulness of repository data – when used with appropriate attention to data cleaning and exploration, repository data can be a valuable resource in biogeographic research, particularly for those under-studied taxa lacking alternative biological data sources.

Data availability statement

The cleaned dataset of Australian fungal occurrences and the R script for data cleaning are available at: https://doi.org/10.17632/38zfzr4z3w.1.

Acknowledgements

T. Hao is supported by a Melbourne Research Scholarship awarded by the University of Melbourne. We thank the Atlas of Living Australia, the Global Biodiversity Information Facility, MycoPortal, iNaturalist, and Fungimap Inc. for making data available and all the collectors and observers who created the data we analysed. We thank the editor, Dr. D. Bebber, reviewer Dr. S. Adamcik, and two other anonymous reviewers for their insightful feedback.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.funeco.2021.101097.

References

Anderson, R.P., Araujo, M.B., Guisan, A., Lobo, J.M., Martínez-Meyer, E., Peterson, A.T., Soberón, J., 2016. Are species occurrence data in global online repositories fit for modeling species distributions? The Case of the Global Biodiversity Information Facility (GBIF). Final Report of the Task Group on GBIF Data Fitness for Use in Distribution Modelling. Global Biodiversity Information Facility (GBIF).

Andrew, C., Diez, J., James, T.Y., Kauserud, H., 2019. Fungarium specimens: a largely untapped source in global change biology and beyond. Phil. Trans. Biol. Sci. 374 https://doi.org/10.1098/rstb.2017.0392.

Andrew, C., Halvorsen, R., Heegaard, E., Kuyper, T.W., Heilmann-Clausen, J., Krisai-Greilhuber, I., Bässler, C., Egli, S., Gange, A.C., Høiland, K., Kirk, P.M., Senn-Irlet, B., Boddy, L., Büntgen, U., Kauserud, H., 2018. Continental-scale macrofungal assemblage patterns correlate with climate, soil carbon and nitrogen deposition. J. Biogeogr. 45, 1942–1953. https://doi.org/10.1111/jbi.13374.

Andrew, C., Heegaard, E., Kirk, P.M., Bässler, C., Heilmann-Clausen, J., Krisai-Greilhuber, I., Kuyper, T.W., Senn-Irlet, B., Büntgen, U., Diez, J., Egli, S., Gange, A. C., Halvorsen, R., Høiland, K., Nordén, J., Rustøen, F., Boddy, L., Kauserud, H., 2017. Big data integration: pan-European fungal species observations' assembly for addressing contemporary questions in ecology and global change biology. Fungal Biol. Rev. 31, 88–98. https://doi.org/10.1016/j.fbr.2017.01.001.

Atlas of Living Australia occurrence download. https://doi.org/10.26197/5da44b73c57f9. (Accessed 14 October 2019).

Bissett, A., Fitzgerald, A., Meintjes, T., Mele, P.M., Reith, F., Dennis, P.G., Breed, M.F., Brown, B., Brown, M.V., Brugger, J., Byrne, M., Caddy-Retalic, S., Carmody, B., Coates, D.J., Correa, C., Ferrari, B.C., Gupta, V.V.S.R., Hamonts, K., Haslem, A., Hugenholtz, P., Karan, M., Koval, J., Lowe, A.J., Macdonald, S., McGrath, L., Martin, D., Morgan, M., North, K.I., Paungfoo-Lonhienne, C., Pendall, E., Phillips, L., Pirzl, R., Powell, J.R., Ragan, M.A., Schmidt, S., Seymour, N., Snape, I., Stephen, J. R., Stevens, M., Tinning, M., Williams, K., Yeoh, Y.K., Zammit, C.M., Young, A.,

2016. Introducing BASE: the Biomes of Australian Soil Environments soil microbial diversity database. GigaScience 5, 21. https://doi.org/10.1186/s13742-016-0126-5.
Breunig, M.M., Kriegel, H.-P., Ng, R.T., Sander, J., 2000. LOF: identifying density-based local outliers. In: Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, SIGMOD '00. Association for Computing Machinery, New York, NY, USA, pp. 93–104. https://doi.org/10.1145/342009.335388.
Callaghan, C.T., Poore, A.G.B., Major, R.E., Rowley, J.J.L., Cornwell, W.K., 2019a. Optimizing future biodiversity sampling by citizen scientists. Proc. Biol. Sci. 286, 20191487. https://doi.org/10.1098/rspb.2019.1487.
Callaghan, C.T., Rowley, J.J.L., Cornwell, W.K., Poore, A.G.B., Major, R.E., 2019b. Improving big citizen science data: moving beyond haphazard sampling. PLoS Biol. 17 https://doi.org/10.1371/journal.pbio.3000357 e3000357.
Cassis, G., Laffan, S.W., 2017. Biodiversity and bioregionalisation perspectives on the historical biogeography of Australia. In: Ebach, M.C. (Ed.), Handbook of Australian Biogeography. CRC Press, Boca Raton, USA, Melbourne, pp. 11–26.
Chapman, A.D., 2005. Principles and Methods of Data Cleaning. GBIF.
Cox, B., 2001. The biogeographic regions reconsidered. J. Biogeogr. 28, 511–523. https://doi.org/10.1046/j.1365-2699.2001.00566.x.
Cutler, D.R., Edwards Jr., T.C., Beard, K.H., Cutler, A., Hess, K.T., Gibson, J., Lawler, J.J., 2007. Random forests for classification in ecology. Ecology 88, 2783–2792. https://doi.org/10.1890/07-0539.1.
Dunstan, P.K., Foster, S.D., Darnell, R., 2011. Model based grouping of species across environmental gradients. Ecol. Model. 222, 955–963. https://doi.org/10.1016/j.ecolmodel.2010.11.030.
Ebach, M., 2012. A history of biogeographical regionalisation in Australia. Zootaxa 3392, 1. https://doi.org/10.11646/zootaxa.3392.1.1.
Elith, J., Kearney, M., Phillips, S., 2010. The art of modelling range-shifting species. Methods Ecol. Evol. 1, 330–342. https://doi.org/10.1111/j.2041-210X.2010.00036.x.
Elith, J., Leathwick, J.R., 2009. Species distribution models: ecological explanation and prediction across space and time. In: Annual Review of Ecology Evolution and Systematics. Annual Reviews, Palo Alto, pp. 677–697.
Gange, A.C., Allen, L.P., Nussbaumer, A., Gange, E.G., Andrew, C., Egli, S., Senn-Irlet, B., Boddy, L., 2019. Multiscale patterns of rarity in fungi, inferred from fruiting records. Global Ecol. Biogeogr. 28 (8) https://doi.org/10.1111/geb.12918.
GBIF Occurrence Download. https://doi.org/10.15468/dl.gqtesn. (Accessed 14 October 2019).
González-Orozco, C.E., Ebach, M.C., Laffan, S., Thornhill, A.H., Knerr, N.J., Schmidt-Lebuhn, A.N., Cargill, C.C., Clements, M., Nagalingum, N.S., Mishler, B.D., Miller, J. T., 2014. Quantifying phytogeographical regions of Australia using geospatial turnover in species composition. PLoS One 9. https://doi.org/10.1371/journal.pone.0092558 e92558.
Gotelli, N.J., Colwell, R.K., 2001. Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. Ecol. Lett. 4, 379–391. https://doi.org/10.1046/j.1461-0248.2001.00230.x.
Guillera-Arroita, G., Lahoz-Monfort, J.J., Rooyen, A.R. van, Weeks, A.R., Tingley, R., 2017. Dealing with false-positive and false-negative errors about species occurrence at multiple levels. Methods Ecol. Evol. 8, 1081–1091. https://doi.org/10.1111/2041-210X.12743.
Guisan, A., Thuiller, W., Zimmermann, N.E., 2017. Habitat suitability and distribution models: with applications in R, ecology. Biodivers. Conserv. https://doi.org/10.1017/9781139028271. Cambridge University Press.
Hao, T., Elith, J., Guillera-Arroita, G., Lahoz-Monfort, J.J., May, T.W., 2021. Curated Open-Access Fungi Occurrence Data for Australia. Mendeley Data. V1. https://doi.org/10.17632/38zfzr4z3w.1.
Hao, T., Guillera-Arroita, G., May, T.W., Lahoz-Monfort, J.J., Elith, J., 2020. Using species distribution models for fungi. Fungal Biol. Rev. 34 (2) https://doi.org/10.1016/j.fbr.2020.01.002.
Hastie, T., Tibshirani, R., Friedman, J.H., 2009. The Elements of Statistical Learning : data Mining, Inference, and Prediction, Springer Series in Statistics. Springer, New York c2009.
Laffan, S.W., Lubarsky, E., Rosauer, D.F., 2010. Biodiverse, a tool for the spatial analysis of biological and related diversity. Ecography 33, 643–647. https://doi.org/10.1111/j.1600-0587.2010.06237.x.
Lennon, J.J., Koleff, P., GreenwooD, J.J.D., Gaston, K.J., 2001. The geographical structure of British bird distributions: diversity, spatial turnover and scale. J. Anim. Ecol. 70, 966–979. https://doi.org/10.1046/j.0021-8790.2001.00563.x.
Liaw, A., Wiener, M., 2002. Classification and Regression by Random Forest, 2. R News, pp. 18–22.
Mackey, B.G., Berry, S.L., Brown, T., 2008. Reconciling approaches to biogeographical regionalization: a systematic and generic framework examined with a case study of the Australian continent. J. Biogeogr. 35, 213–229. https://doi.org/10.1111/j.1365-2699.2007.01822.x.
Madsen, J.H., 2018. DDoutlier: Distance & Density-Based Outlier Detection.
Mair, L., Ruete, A., 2016. Explaining spatial variation in the recording effort of citizen science data across multiple taxa. PLoS One 11. https://doi.org/10.1371/journal.pone.0147796 e0147796.
May, T.W., 2017. Biogeography of Australasian fungi: from mycogeography to the mycobiome. In: Handbook of Australasian Biogeography. CRC Press, pp. 165–240.
Morera, A., Martínez de Aragón, J., Bonet, J.A., Liang, J., de-Miguel, S., 2021. Performance of statistical and machine learning-based methods for predicting biogeographical patterns of fungal productivity in forest ecosystems. For. Ecosyst. 8, 21. https://doi.org/10.1186/s40663-021-00297-w.
Nguyen, N.H., Song, Z., Bates, S.T., Branco, S., Tedersoo, L., Menke, J., Schilling, J.S., Kennedy, P.G., 2016. FUNGuild: an open annotation tool for parsing fungal community datasets by ecological guild. Fungal Ecol. 20, 241–248. https://doi.org/10.1016/j.funeco.2015.06.006.
Oksanen, J., Blanchet, F.G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P.R., O'Hara, R.B., Simpson, G.L., Solymos, P., Stevens, M.H.H., Szoecs, E., Wagner, H., 2019. Vegan: Community Ecology Package.
Phillips, S.J., Dudik, M., Elith, J., Graham, C.H., 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. https://doi.org/10.1890/07-2153.1.
Pyšek, P., Richardson, D.M., Rejmánek, M., Webster, G.L., Williamson, M., Kirschner, J., 2004. Alien plants in checklists and floras: towards better communication between taxonomists and ecologists. Taxon 53, 131–143. https://doi.org/10.2307/4135498.
R Core Team, 2017. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
Robertson, M., 2016. Biogeo: Point Data Quality Assessment and Coordinate Conversion.
Serra-Diaz, J.M., Enquist, B.J., Maitner, B., Merow, C., Svenning, J.-C., 2018. Big data of tree species distributions: how big and how good? For. Ecosyst. 4, 30. https://doi.org/10.1186/s40663-017-0120-0.
Sokal, R.R., 1958. A statistical method for evaluating systematic relationships. Univ. Kans. Sci. Bull. 38, 1409–1438.
Speed, J.D.M., Bendiksby, M., Finstad, A.G., Hassel, K., Kolstad, A.L., Prestø, T., 2018. Contrasting spatial, temporal and environmental patterns in observation and specimen based species occurrence data. PLoS One 13. https://doi.org/10.1371/journal.pone.0196417 e0196417.
Thessen, A.E., Patterson, D.J., 2011. Data issues in the life sciences. ZooKeys 15–51. https://doi.org/10.3897/zookeys.150.1766.
Troia, M.J., McManamay, R.A., 2016. Filling in the GAPS: evaluating completeness and coverage of open-access biodiversity databases in the United States. Ecol. Evol. 6, 4654–4669. https://doi.org/10.1002/ece3.2225.
Weiss, D.J., Nelson, A., Gibson, H.S., Temperley, W., Peedell, S., Lieber, A., Hancher, M., Poyart, E., Belchior, S., Fullman, N., Mappin, B., Dalrymple, U., Rozier, J., Lucas, T. C.D., Howes, R.E., Tusting, L.S., Kang, S.Y., Cameron, E., Bisanzio, D., Battle, K.E., Bhatt, S., Gething, P.W., 2018. A global map of travel time to cities to assess inequalities in accessibility in 2015. Nature 553, 333–336. https://doi.org/10.1038/nature25181.
Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., Robertson, T., Vieglais, D., 2012. Darwin Core: an evolving community-developed biodiversity data standard. PLoS One 7. https://doi.org/10.1371/journal.pone.0029715 e29715.
Woolley, S.N.C., McCallum, A.W., Wilson, R., O'Hara, T.D., Dunstan, P.K., 2013. Fathom out: biogeographical subdivision across the Western Australian continental margin – a multispecies modelling approach. Divers. Distrib. 19, 1506–1517. https://doi.org/10.1111/ddi.12119.
Wüest, R.O., Zimmermann, N.E., Zurell, D., Alexander, J.M., Fritz, S.A., Hof, C., Kreft, H., Normand, S., Cabral, J.S., Szekely, E., Thuiller, W., Wikelski, M., Karger, D.N., 2020. Macroecology in the age of Big Data – where to go from here? J. Biogeogr. 47, 1–12. https://doi.org/10.1111/jbi.13633.
Zizka, A., Silvestro, D., Andermann, T., Azevedo, J., Ritter, C.D., Edler, D., Farooq, H., Herdean, A., Ariza, M., Scharn, R., Svantesson, S., Wengström, N., Zizka, V., Antonelli, A., 2019. CoordinateCleaner: standardized cleaning of occurrence records from biological collection databases. Methods Ecol. Evol. 10, 744–751. https://doi.org/10.1111/2041-210X.13152.
Zurell, D., Franklin, J., König, C., Bouchet, P.J., Dormann, C.F., Elith, J., Fandos, G., Feng, X., Guillera-Arroita, G., Guisan, A., Lahoz-Monfort, J.J., Leitão, P.J., Park, D.S., Peterson, A.T., Rapacciuolo, G., Schmatz, D.R., Schröder, B., Serra-Diaz, J.M., Thuiller, W., Yates, K.L., Zimmermann, N.E., Merow, C., 2020. A standard protocol for reporting species distribution models. Ecography 43. https://doi.org/10.1111/ecog.04960.