Overcoming Limitations in Biodiversity Data

Data cleaning solutions for macroecological diversity models and Drivers of the variation in the synonym numbers of angiosperm species names

Dissertation zur Erlangung des Doktorgrades
"Doctor rerum naturalium"
der Mathematisch-Naturwissenschaftlichen Fakultäten der
Georg-August-Universität Göttingen

vorgelegt von

Dipl. Biol.

Petra Führding-Potschkat

aus Maulbronn

Göttingen, 17. November 2022

Gutachter

Prof. Dr. Holger Kreft, Biodiversität, Makroökologie und Biogeographie, Universität Göttingen
Prof. Dr. Stefanie M. Ickert-Bond, University of Alaska Fairbanks, Department of Biology and Wildlife, Fairbanks, AK 99775

Tag der mündlichen Prüfung: 13. 10. 2022

"God is in the detail"

— Gustave Flaubert

Table of contents

List of Tables

List of Figures

Author contributions

CHAPTER 2

Influence of different data cleaning solutions of point-occurrence records on downstream macroecological diversity models

The following authors contributed to this chapter: Petra Führding-Potschkat[1], Holger Kreft[1] and Stefanie M. Ickert-Bond[2]

P.F.P. led this study; all authors designed this study; P.F.P. compiled the data; P.F.P. analyzed the data; P.F.P. led the writing with substantial contributions from all authors.

Fuehrding-Potschkat, P., Kreft, H., Ickert-Bond, S.M. (2022). Influence of different data cleaning solutions of point-occurrence records on downstream macroecological diversity models. *Ecology and Evolution* **12**: e9168. https://doi.org/10.1002/ece3.9168.

CHAPTER 3

Drivers of the variation in the synonym numbers of angiosperm species names

Petra Führding-Potschkat[1], Patrick Weigelt[1], Holger Kreft[1] and Stefanie M. Ickert-Bond[2]

H.K. conceived the idea, P.F.P. led this study; all authors designed this study; P.W. collected the data; P.F.P. compiled the data; P.F.P. analyzed the data; P.F.P. led the writing with major contribution from all authors.

Unpublished manuscript.

Author affiliations

[1] Biodiversität, Makroökologie und Biogeographie; University of Göttingen. Büsgenweg 1; 37077 Göttingen; Germany.

[2] University of Alaska Fairbanks, Department of Biology and Wildlife, Fairbanks, AK 99775.

Summary

Large-scale biodiversity data drive the research of the variation and inter-dependence among living organisms that sustain life. Taxonomic and point-occurrence validity of species, sample completeness, and consistency are essential aspects of these data. A convenient way to access biodiversity data is through digital specimen records stored with public data providers. However, recent evaluations of public provider data showed inconsistent quality derived from, e.g., misidentified species, incongruities in associating synonyms with their accepted species, coordinate errors, and missing values. Therefore, one should not assume that the quality of public provider data is suitable for immediate use. My thesis comprises two independent studies in which I examine taxonomic and spatial limitations in biodiversity data retrieved from two major public data providers.

I. Developing data cleaning (DC) strategies and tools to reproducibly generate consistent data from global public provider data is a long-standing goal of biodiversity informatics. Coded instructions and *R* packages to retrieve, evaluate, format, and organize data are examples of such developments. While newly programmed and recently updated automated methods and tools are promising to support public data users, their effect on downstream macroecological diversity models remains poorly examined.

In chapter 2, I introduce the first quantitative analysis of how data, processed in DC pipelines using popular DC methods and tools, influenced downstream species distribution models (SDM). I focused on two aspects. (1), I examined the standardization and error removal performance of six DC pipelines, using 46,384 North American *Ephedra* records as input from the Global Biodiversity Information Facility (GBIF). (2), I analyzed differences in the SDM and stacked SDMs (S-SDMs) of *Ephedra* species in North America (e.g., caused by retained errors in the pipeline data). To test the reliability of the results, I compared the pipeline data SDMs to corresponding expert data SDMs that represented the gold standard. (1) Depending on the pipeline, about one-third (GBIF-filtered) to two-thirds (*R* packages-processed) of the records were unsuitable for biodiversity analyses. While the *R* package-based pipelines offered automated data cleaning in a standardized and reproducible manner, the GBIF-filtered data still contained significant spatial and taxonomic errors. Major drawbacks emerged from the fact that no pipeline entirely discovered misidentified specimens without the assistance of expert taxonomic knowledge. These results support the hypothesis that different data cleaning

solutions provide different data qualities. (2) Differences in the pipeline data did not translate into significant differences in downstream SDMs and S-SDMs. However, the prediction that models and maps from public provider data would differ significantly from expert data was supported by respective correlations in the models and maps (using Pearson's *r*).

II. Synonyms are a common part of scientific progression in taxonomy and nomenclature. They can emerge for different reasons, for example, because taxonomists interpret and classify interspecific variations differently. Synonyms may cause severe taxonomic uncertainties in biodiversity repositories (e.g., confusing taxonomy when it is challenging to recognize whether a species' name is an alias of a more common species). Recent studies showed that some taxa's synonymy level is quite substantial. In this context, several causes, in addition to splitting and lumping, were suggested to lead to variation in synonym numbers; for example, taxonomists might show preferences toward attractive taxonomic entities.

In chapter 3, I present five drivers of synonym numbers I hypothesized to account for variation in global angiosperm synonym numbers. The drivers comprised higher taxa of a species (family and genus), the botanical continents where a species is present, the insularity of a species (defined as the occurrence on islands, the mainland, or both), a species' range size, and the age of its accepted name. Using multi-model inference, I quantified the relative importance of the drivers across 137,378 accepted names of 193 angiosperm families and 5,019 genera present in 355 TDWG countries and regions worldwide using data from the World Checklist of Selected Plant Families (WCSP). The synonym number was used as the response variable in the models for explanations and predictions; the synonymy rate allowed for a relative ranking in groups (e.g., order of genera in angiosperm families). I identified range size, the age of an accepted name, and insularity as the core drivers that positively affected the global variation of synonym numbers. After accounting for these three factors, the residual differences in the number of botanical continents and the interaction of insularity and the range size were less significant. The combined multi-predictor model explained about 41% of the global variation in angiosperm synonymy (96%, including the random effects of the botanical continents, genera, and families).

Two essential interpretations emerged from the studies. First, when consistent species information is critical, expert data unavailable, and public biodiversity providers are known for frequently storing data of poor quality, this should prompt users to improve the data within their control before use. However, this will usually happen locally at a user’s space using retrieved data from the providers. Second, when, in particular, taxonomic accuracy is essential, data from a public provider requires additional effort. In this case, the biodiversity data should be

thoroughly analyzed with expert help since dubious specimens can still hide even in the cleaned data.

Zusammenfassung

Umfangreiche Biodiversitätsdaten treiben die Erforschung der Variation und gegenseitigen Abhängigkeit zwischen lebenden Organismen voran, die das Leben erhalten. Wesentliche Aspekte dieser Daten sind die Validität der Arten und Punktvorkommen, und die Vollständigkeit und Konsistenz von Zufallsstichproben zur Vermeidung von "Bias". Ein komfortabler Weg, auf Biodiversitätsdaten zuzugreifen, sind digitale Datensätze von Specimens, die bei öffentlichen Datenanbietern gespeichert sind. Jüngste Auswertungen von Daten öffentlicher Anbieter zeigten jedoch eine uneinheitliche Qualität, die z. B. falsch identifizierte Arten, Unstimmigkeiten bei der Zuordnung von Synonymen zu ihren akzeptierten Arten, Koordinatenfehler und fehlende Werten. Daher sollte man nicht davon ausgehen, dass die Datenqualität öffentlicher Anbieter für die sofortige Nutzung geeignet ist. Meine Dissertation umfasst zwei voneinander unabhängige Studien, in denen ich taxonomische und räumliche Mängel in Biodiversitätsdaten untersuche, die ich zwei großen öffentlichen Datenanbietern abgerufen habe.

I. Die Entwicklung von Datenbereinigungs (DC) Strategien und Werkzeugen zur reproduzierbaren Generierung konsistenter Daten aus Datenbeständen globaler öffentlicher Anbieter ist ein langjähriges Ziel der Biodiversitätsinformatik. Codierte Anweisungen und *R*-Pakete zum Abrufen, Auswerten, Formatieren und Organisieren von Daten sind Beispiele für solche Entwicklungen. Während neu programmierte und kürzlich aktualisierte, automatisierte Methoden und Werkzeuge vielversprechend sind, um die Nutzer öffentlicher Daten zu unterstützen, ist ihre Wirkung auf nachgelagerte makroökologische Diversitätsmodelle noch wenig untersucht.

In Kapitel 2 stelle ich die erste quantitative Analyse vor, wie Daten, die in DC-Pipelines mit gängigen DC-Methoden und -Werkzeugen verarbeitet wurden, nachgelagerte Artenverteilungs-modelle (SDM) beeinflussten. Ich habe mich auf zwei Aspekte konzentriert. (1) untersuchte ich die Standardisierungs- und Fehlerbeseitigungsleistung von sechs DC-Pipelines unter Verwendung von 46.384 nordamerikanischen *Ephedra*-Aufzeichnungen, abgerufen aus der Global Biodiversity Information Facility (GBIF). (2) analysierte ich Unterschiede in den SDMs und gestapelten SDMs (S-SDMs) von *Ephedra*-Arten in Nordamerika (z. B. verursacht durch zurückbehaltene Fehler in den Pipeline-Daten). Um die Zuverlässigkeit der Ergebnisse zu testen, habe ich die Pipeline-Daten-SDMs mit entsprechenden Expertendaten-SDMs verglichen

(Die Expertendaten repräsentierten den Goldstandard). (1) Je nach Pipeline waren etwa ein Drittel (GBIF-gefiltert) bis zwei Drittel (von *R*-Paketen verarbeitet) der Aufzeichnungen für Biodiversitätsanalysen ungeeignet. Während die auf R-Paketen basierenden Pipelines eine automatisierte Datenbereinigung auf standardisierte und reproduzierbare Weise boten, enthielten die GBIF-gefilterten Daten immer noch erhebliche räumliche und taxonomische Fehler. Große Nachteile ergaben sich aus der Tatsache, dass keine Pipeline vollständig die fehlbestimmten Specimen ohne die Unterstützung von taxonomischem Expertenwissen entdeckte. Diese Ergebnisse stützen die Hypothese, dass verschiedene Datenbereinigungs-lösungen unterschiedliche Datenqualitäten liefern. (2) Unterschiede in den Pipelinedaten führten nicht zu signifikanten Unterschieden in nachgelagerten SDMs und S-SDMs. Die Vorhersage, dass sich Modelle und Karten aus Daten öffentlicher Anbieter signifikant von Expertendaten unterscheiden würden, wurde jedoch durch entsprechende Korrelationen in den Modellen und Karten (unter Verwendung von Pearson's r) gestützt.

II. Synonyme sind ein üblicher Bestandteil der wissenschaftlichen Weiterentwicklung in Taxonomie und Nomenklatur. Sie können aus unterschiedlichen Gründen entstehen, zum Beispiel weil Taxonomen interspezifische Variationen unterschiedlich interpretieren und klassifizieren. Synonyme können schwerwiegende taxonomische Unsicherheiten in Biodiversitäts-Repositorien verursachen (z. B. eine künstliche Erhöhung der Anzahl von Artnamen, Verwechslungen in Taxonomien, wenn es schwierig ist zu erkennen, ob der Artname ein Alias einer häufigeren Art ist). Neuere Studien haben gezeigt, dass das Synonymieniveau einiger Taxa ziemlich beträchtlich ist. In diesem Zusammenhang wurden neben dem Aufteilen und Zusammenfassen mehrere Ursachen vorgeschlagen, die zu einer Variation der Synonymzahlen führen. Beispielsweise könnten Taxonomen Präferenzen gegenüber attraktiven taxonomischen Einheiten zeigen.

In Kapitel 3 stelle ich fünf Synonymietreiber vor, von denen ich angenommen habe, dass sie die nicht-nomenklaturbedingten Variation in den globalen Angiospermen-Synonymzahlen erklären. Die Treiber umfassten höhere Taxa einer Art (Familie und Gattung), die botanischen Kontinente, auf denen eine Art vorkommt, die Insellage einer Art (definiert als das Vorkommen auf Inseln, dem Festland oder beiden), die Größe des Verbreitungsgebiets einer Art und das Alter seines akzeptierten Namens. Mittels Multi-Modell-Inferenz habe ich die relative Bedeutung der Treiber unter Verwendung von Daten aus der World Checklist of Selected Plant Families (WCSP) quantifiziert (für 137.378 akzeptierte Namen von 193 Angiospermenfamilien und 5.019 Gattungen, die in 355 TDWG-Ländern und -Regionen weltweit vorkommen). Als

Antwortvariable wurde in den Modellen die Synonymzahl verwendet (für "response" und "prediction"); die Synonymierate ermöglichte eine relative Rangfolge in Gruppen (z. B. für die Reihenfolge der Gattungen in Angiospermenfamilien). Ich identifizierte die Bereichsgröße, das Alter eines akzeptierten Namens und die Insellage als die Haupttreiber, die sich positiv auf die globale Variation von Synonymnummern auswirkten. Nach Berücksichtigung dieser drei Faktoren waren die verbleibenden Unterschiede in der Anzahl der botanischen Kontinente und der Wechselwirkung von Insellage und Verbreitungsgröße weniger signifikant. Das kombinierte Multi-Prädiktor-Modell erklärte etwa 41 % der globalen Variation der Angiospermen-Synonymie (96 % einschließlich der zufälligen Effekte der botanischen Kontinente, Gattungen und Familien).

Zwei weitere wichtige Aspekte kristallisierten sich aus den Studien heraus. Erstens, wenn konsistente Arteninformationen kritisch und Expertendaten nicht verfügbar sind und öffentliche Biodiversitätsanbieter dafür bekannt sind, dass sie oft Daten von schlechter Qualität speichern, sollte dies die Benutzer dazu veranlassen, Daten unter ihrer Kontrolle vor der Verwendung zu verbessern. Dies geschieht jedoch in der Regel lokal bei abgerufenen Anbieterdaten. Zweitens, wenn es insbesondere auf taxonomische Genauigkeit ankommt, erfordern Daten eines öffentlichen Anbieters zusätzlichen Aufwand. In diesem Fall sollten die Biodiversitätsdaten mit Hilfe von Experten gründlich analysiert werden, da sich auch in den bereinigten Daten immer noch zweifelhafte Specimen verbergen können.

1

Introduction

Research background

Anthropogenic transformations worldwide, including land degradation (Scholes et al., 2018), deforestation in the tropics (Myers, 1995), the loss of species (e.g., Koh et al., 2004, Powers & Jetz, 2019), and regional floristic uniqueness (Qiang et al., 2021), indicate a compelling need to study biodiversity. Biodiversity refers to the variation and the inter-dependence among living organisms from genes, traits, and species to ecosystem, as well as cultural processes that sustain life in a particular region or globally (Whittaker et al., 2005, Chandra & Idrisova, 2011, McGill et al., 2015). Biodiversity research reveals the services that intact ecosystems provide to all living organisms on earth (e.g., Morton & Hill, 2014). The services provide food and water, influence climate and diseases, nutrients and crop pollination, and enable many cultural services like recreational benefits (Morton & Hill, 2014). Thus, it is vital to intensify research and collect and preserve all attainable data of the organisms involved, i.e., their biodiversity data. Biodiversity data sit at the core of many knowledge areas, like taxonomy, ecology, and evolutionary biology, where they are used to describe organisms and the organisms' distributions, functions, and phylogenies (Soberón & Peterson, 2004, Hamilton, 2005). Digitally available specimen records are the most common representation of public biodiversity data, mainly stemming from field collections, herbarium inventories, and citizen scientist's observations (Graham et al., 2004). Figure 1.1 presents the general data flow of transforming raw biodiversity data into biodiversity data fit for use. Significant data quality aspects are the validity of the species and point-occurrences and the completeness and consistency of sampling to avoid bias. The data's timeliness is also essential, as historical specimen records are invaluable for tracking changes in biodiversity. The figure also shows where data errors can enter the data flow, which is particularly essential as it may trigger two questions by public data users: 1. are the retrieved data reliable, and 2. are they fit-for use? Answering the first question will lead to data evaluation. Answering the second question initiates customized data cleaning to standardize the data and correct errors (e.g., Araujo et al., 2019, Zizka et al., 2020).

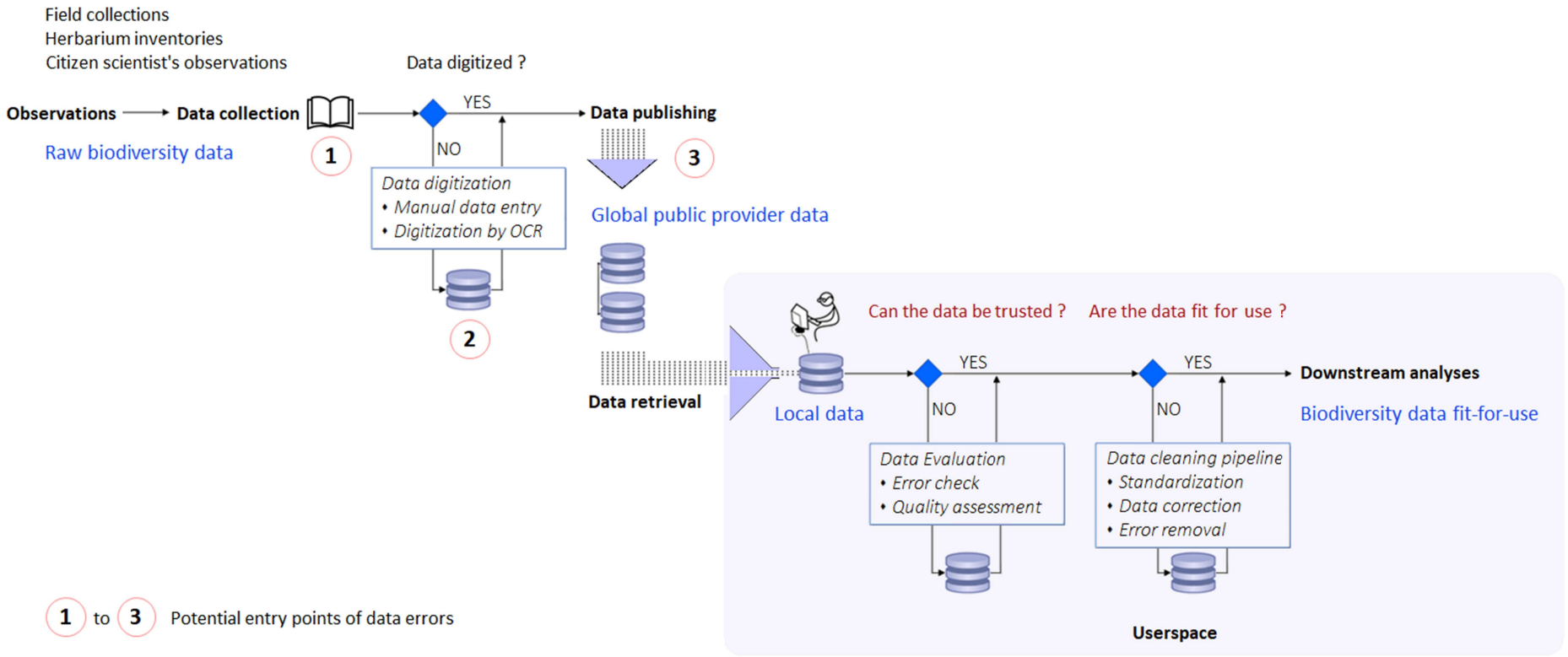


Figure 1.1. **Data flow of the raw biodiversity data transformation into biodiversity data fit for use.** Data errors can enter the data flow at three places: (1) At data collection (e.g., misidentifications, measurement errors, missing values), (2) data digitization (e.g., OCR reading errors, missing values, misspellings), and (3) the data exchange between public data providers. Thus, after data retrieval into the local, private computer workplace (Userspace), the local data must be considered from two perspectives: 1. are the data reliable, and 2. are they fit for use? Answering the first question will lead to data evaluation (e.g., questionable data sources, the validity of the taxa and respective point-occurrences, sampling completeness, and data consistency, e.g., Araujo et al., 2019, Guisan et al., 2017). Answering the second question initiates a customized data cleaning pipeline to standardize the data and correct errors. (e.g., Belbin et al., 2013, Chapman et al., 2020, Zizka et al., 2020). Data evaluation and cleaning pipeline details, see: Figure 1.2. Abbreviation: OCR, optical character recognition. Drawing: courtesy C. Meyer.

The demand for high-quality biodiversity data is substantial (e.g., Guisan et al., 2017, Raes & Aguirre-Gutierrez, 2018, Araújo et al., 2019). However, high-quality data remain limited even when combining all available data from various sources. In recent years different and sometimes contradicting approaches were described to accomplish high-quality biodiversity data. One approach was based on experienced users' knowledge of the design and use of "gold standards" in field studies. Depending on the targeted biodiversity model, the chosen data standardizations and improvements should be implemented in userspace (Chapman, 2005, Araújo et al., 2019, Chapman et al., 2020). Other authors (e.g., Yesson et al., 2007, Mesibov, 2013) expressed regularly improving biodiversity data at the data provider's sites, performed by experts, as their favored approach. It was also suggested to improve the biodiversity data by the data user community interacting with the global public data providers. Processing data that contain limitations to achieve biodiversity data fit for use would depend on the scrutiny of both data providers and expert data users (e.g., Belbin et al., 2013, Costello et al., 2013, Chapman et al., 2020). In addition, taxonomic experts were asked to improve data of their area of expertise at museums and global public data providers (e.g., Belbin et al., 2013, Ickert-Bond, 2003). Hereafter, I will apply the term "fit for use" (Chapman et al., 2020) to biodiversity data that were already explored, evaluated, standardized, and cleaned and are ready for downstream biodiversity analyses. The term "public biodiversity data provider" refers to sources accessible to anyone without particular qualifications or authorizations to retrieve the data.

Global public provider data

The Global Biodiversity Information Facility (GBIF) is a well-known global biodiversity data provider, funded by national governments (e.g., Australia, Brazil, Denmark, South Africa, and the United States). It presently comprises the most notable global infrastructure, storing more than 2.2 billion specimen records of many scientific sources (e.g., other data providers, herbaria, botanical gardens, and citizen scientists) in its data warehouses (GBIF.org, 2020). A second, and also important, example is the World Checklist of Selected Plant Families (WCSP, wcsp.science.kew.org), which is maintained by the Royal Botanic Gardens, Kew. The primary information in the WCSP records is the taxonomically assessed and standardized species names of currently 270 plant families, including their complete classification, the name status, description information, and the taxa's countries of occurrence.

Limitations in public provider data

Evaluations of GBIF-hosted biodiversity data using characteristics of high-quality data for comparison showed that the GBIF data were of inconsistent quality (e.g., Wieczorek et al.,

2012, Sousa-Baena et al., 2014, Meyer et al., 2016). For example, the circumstances and standards under which the data were collected and digitized were undocumented (e.g., Sterner & Franz, 2017). Thus, it can be assumed when, e.g., specimen records are retrieved from public data providers to userspace, their quality is generally not fit for use. Limitations in the public providers' data occur mainly along three dimensions: taxonomy, space, and time (Meyer et al., 2016). They comprise, e.g., incorrect taxonomic hierarchies, misidentifications, implausible point-occurrences and taxon ranges. However, significant problems for analysis and model building that were difficult to resolve are rather of taxonomical or geographical nature, which I will discuss in detail below.

The knowledge of species and their taxonomy are constantly progressing. Thus, taxonomic errors are difficult or impossible to identify or estimate, particularly in large amounts of data (commonly from global data providers). Such identifications usually involve labor-intensive re-evaluations of the original documentation and metadata (Belbin et al., 2013). Taxonomic errors comprise, e.g., misidentified or incompletely identified species (e.g., only to the genus), uncertainties whether the names are accepted or synonyms, incorrect linking of synonyms to the accepted name and incorrect spelling of taxon names (classification of errors, see Kutsch & Hall, 2010). It is unknown how often species are misidentified. Different authors estimated misidentification rates between < 1 and 17% (Bisang & Urmi, 1994, Scott & Hallam, 2002, Ahrends et al., 2011). Frequently quoted studies assumed that more than 50% of tropical specimens, on average, were likely incorrectly named (Goodwin et al., 2015) and that incorrectly named specimens in the Zoological Record database ranged from 5% to nearly 60% (Meier & Dikow, 2004). It was also found that taxonomic misidentification errors in differently determined specimens of the same origin were only visible when recognizing them as duplicates given to other institutes and handled in isolation from their parent specimens (Nicolson, 2019). Synonyms that are incorrectly or ambiguously linked to the parent species name cause inaccurate references in taxonomic checklists and can compromise floras and checklists. Unrecognized synonyms, orthographic variations, and incorrectly spelled names (Jansen & Dengler, 2010) lead to inflated species numbers, which may influence conservation efforts for species that do not exist or are more frequent than initially believed (Linnéan shortfall; Lomolino, 2004, Hortal et al., 2015, Ickert-Bond et al., 2019).

Point-occurrences are spatially accurate when details of specimen locations are given with consistent accuracy (Yesson et al., 2007). Spatial errors in biodiversity data, including spatial biases, make views on species distribution difficult or unattainable (Mitchell et al., 2017). The

errors include invalid, inaccurate, and imprecise locations for specimens. For example, zeroes and missing values in coordinates cause invalid specimen locations. Still, whether such invalid specimen locations result from protecting sensitive species against exploitation ('dark' fields) or as a consequence of spatial errors is usually not indicated in the provided records (Anderson et al., 2016, Chapman et al., 2020). It is also unknown how often direction and distance to reference points were measured incorrectly and coordinates were incorrectly recorded from GPS devices (Murphey et al., 2004). Furthermore, coordinate values that, e.g., originated from species range maps instead of georeferencing drive inaccurate coordinates for locations (Zizka et al., 2019, 2020). In addition to spatial errors, geographic sampling inconsistencies, or sampling biases, are a common phenomenon in species distribution data. Sampling biases contribute to the lack of knowledge about the true geographical distribution of a species (Wallacean shortfall; Lomolino, 2004, Hortal et al., 2015). Geographical biases include species that are mistakenly thought to be present or absent, and presence records that rather reflect survey effort than occurrence (Hortal et al., 2007). Other important sources of bias are opportunistic collections (Pyke & Ehrlich, 2010; Ter Steege et al., 2011) that aim at maximizing taxonomic diversity in herbaria or botanical gardens rather than reflecting biogeographical or ecological reasons (e.g., the broad data gap in tropical countries: Prance, 1977; Collen et al., 2008). The roadside bias is a well-studied case of opportunistic collecting behaviour (e.g., Reddy & Dávalos, 2003, Kadmon et al., 2004), where the frequency of observations near roads was consistently greater than expected from a spatially random distribution. Species with valid coordinates outside their native ranges are a particular case. They might be either misidentified specimens, or alien species artificially introduced into an atypical area. Frequent examples are coordinates of implausible and dubious sites (herbaria, botanical gardens, museums, herb shops, etc.) for implausible species or improbable specimens with no distribution status (native or non-native species).

The studies revealed that the limitations in the provider data represent a wide range of errors, constitute a significant challenge to users who require high-quality biodiversity data and that they also pose a central problem to knowledge areas that depend on high-quality data (e.g., Kadmon et al., 2004, Araújo & Guisan, 2006, Despot-Belmonte et al., 2017). The limitations can increase the likelihood that the data will be misinterpreted. When used by decision-makers in, e.g., conservation, they are highly likely to lead to problematic management decisions. Examples of problematic decisions may be, e.g., duplication of efforts and accidental oversight (e.g., CBD, 2009) and conservation priorities that lack sufficient reliable information (CBD, 2009, p. 39, Scholes et al., 2018). The following two chapters in the Research Background will

address important challenges to biodiversity data fit for use concerning taxonomic and spatial errors and data uncertainties and options to overcome them.

Data cleaning (DC) solutions for macroecological diversity models

Although significant limitations in the GBIF data were reported (e.g., Meyer et al., 2016), GBIF declines responsibility for the quality of the content and shifts potential problems related to the data users (Terms of use: Data agreement. GBIF, 2021c). For a data user, manually cleaning is time-consuming and often unfeasible, given that the data sets may contain thousands or millions of records. Therefore, powerful, automated, and locally implementable DC solutions that evaluate, standardize, and clean biodiversity data in high demand. The importance of comprehensive DC solutions is particularly the case if the data errors may hamper downstream analyses and diversity models.

In recent years, standardizing and cleaning methods and tools were designed to support organizations and users in obtaining consistent and integrated biodiversity data. Key considerations when integrating biodiversity data included task-specific evaluation, standardization and cleaning rules as well as instructions (Chapman, 2005, Zizka et al., 2019). With their help, the user could, with the hardware and software that is in line with the requirements of the task, develop sound taxonomic, spatial, and temporal data for downstream analyses (e.g., Guralnick et al., 2018, Araújo et al., 2019, Hijmans & Elith, 2019). Yet, the majority of the methods and tools still comprise single solutions (e.g., Chapman, 2005, Chapman et al., 2020) like instructions and *R* packages supporting data cleaning (general data cleaning: Hijmans & Elith, 2019, Wickham et al., 2019; biodiversity data-specific data cleaning: Zizka et al., 2019). Only the web-based GBIF occurrence-search application to manually filter record subsets (GBIF.org, 2020) corresponds most closely to an end-to-end DC pipeline. Therefore, it is consequential to integrate already existing and ready-to-use DC tools into powerful pipelines to jointly achieve synergies.

DC pipelines are significant in the scientific domains when, for example, biodiversity data from different sources such as herbarium vouchers, observations, and expert data need to be combined. Four steps are suggested for a pipeline (Gueta & Carmel, 2016, GBIF.org, 2020, Zizka et al., 2020): Data retrieval from the source (Local data), evaluation of data errors that might influence the quality of the downstream models, standardization of records (Data evaluation), and correction or removal of errors (Data cleaning). Figure 1.2 generically shows data evaluation and cleaning processes where task-specific methods and tools are used (e.g., suitable *R* packages such as the *CoordinateCleaner*, Zizka et al., 2019).

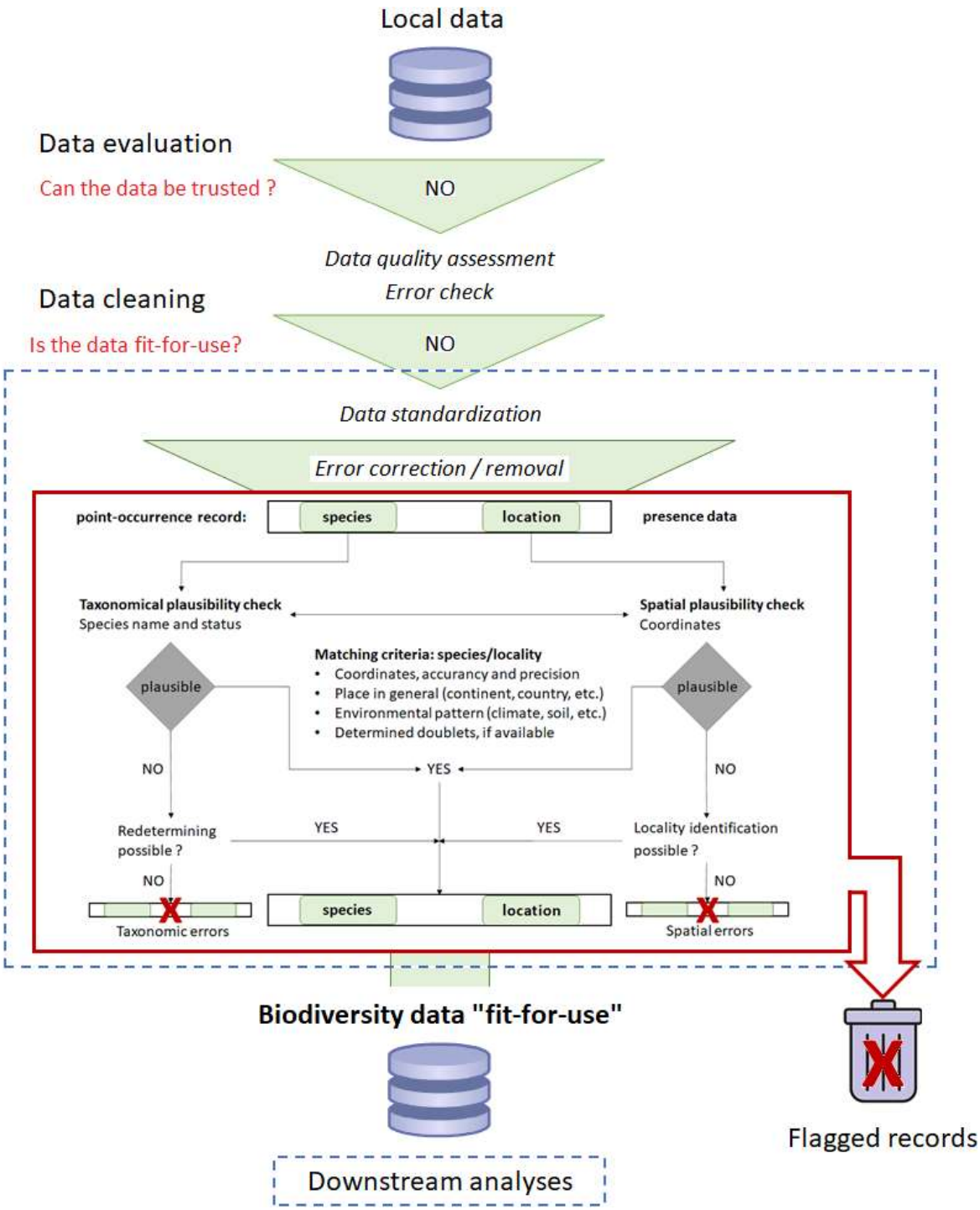


***Figure 1.2*. Overview of the data evaluation step and data cleaning pipeline to identify and address specific data limitations** (Extract from Figure 1.1)**.** Processual challenges comprise assessing how far the data can be trusted and the degree to which the biodiversity data are fit for use. Evaluating the trustworthiness of the data includes weighting taxonomic and spatial errors present in the data based on the specifications of the targeted biodiversity model. Moreover, it includes examining the possibly achievable quality level of the data. Judging whether the biodiversity data are fit for use primarily addresses how to resolve the inherent data issues (based on Araújo et al., 2019; red box in blue dashed box).

Data cleaning typically leads to the reduction of available records (Figure 1.2, Flagged records). However, errors may still remain in the analysis data set (Figure 1.2, Biodiversity data fit for use), which might influence the analyses and models. Spatial errors were investigated in tutorials of various *R* packages (e.g., Hijmans & Elith, 2019), and the importance of filters on the loss of data in DC pipelines (Zizka et al., 2020). In GBIF, it was found that more than 3.4 million records (3.7%) have spatial errors which are potentially problematic (Zizka et al., 2019). Thus, independent data are also required to test the cleaned data against it to avoid such effects (expert data from field studies, herbarium voucher analyses, and distribution maps, or all combined). Biodiversity data evaluation and cleaning results were described for a wide variety of organisms (see section: Limitations in public provider data). Functionalities and capacity of available DC methods and tools were also described and tested (e.g., Hijmans & Elith, 2019, Zizka, 2020). To create effective and meaningful biodiversity models and assemble the right set of DC methods and tools for a pipeline providing biodiversity data fit for use, we must understand which details in the biodiversity data influence models. However, no study has systematically examined the influence of different data cleaning tools in pipelines on macroecological diversity models. Also, correlations between biodiversity models from data of different pipelines and expert data were not yet assessed. I hypothesize that different DC pipelines show (a) different cleaning performances, and (b) that this influences conventional species distribution models (SDM) of a model organism. In addition, I assume that (c) models from provider data will differ significantly from expert data.

Ephedra as the model genus

Ephedra is a popular study object due to interesting ecological and non-ecological traits, highlighted below. North American *Ephedra* which I present in more detail serve as the model group in the thesis. *Ephedra* is adapted to dry environments, it shows a relatively uniform morphology, but different dispersal syndromes which explain specific distributions and species ranges. *Ephedra* is also attractive for taxonomic and phylogenetic studies and pharmaceutical investigations of neuro-pharmaceutical secondary metabolites. Specimens are collected quite frequently, as shown by the record numbers of the public providers (e.g., GBIF, as of November 18, 2021: 46,384 presence records worldwide), and high-quality expert data is available for the New World species (Ickert-Bond, 2003). *Ephedra* is the only genus of the family Ephedraceae. There are about 65 extant *Ephedra* species worldwide (Ickert-Bond & Renner, 2016). The plants are profusedly branched with green photosynthetic stems, mostly with only rudimentary leaves (Hunziker, 1949, Stevenson, 1993, Freitag & Maier-Stolte, 2003, Ickert-Bond &

Wojciechowski, 2004). Most *Ephedra* species grow as shrubs, and a few species are small trees and climbers (Stapf 1889, Freitag and Maier-Stolte, 1994). About 40 species of *Ephedra* are found in the Old World extending westwards from Central Asia across southwest Asia and into Mediterranean Europe, up to the Swiss and Italian Alps, and North Africa (e.g., Freitag & Maier-Stolte, 1994, 2003, Kozhamzharova et al., 2013, Huang et al., 2005). In the New World, thirteen species occur in North America ranging from the southwestern United States to the central plateau of Mexico (e.g., Cutler, 1939, Stevenson, 1993, Ickert-Bond & Wojciechowski, 2004). Twelve more species are found in South America occurring from Ecuador to Patagonia (e.g., Hunziker, 1949, Peinado et al., 2006). The genus *Ephedra* occupies a wide range of habitats, its distribution spans from narrow endemics to widely distributed species. The species are adapted to semiarid and desert conditions, as well as to seasonally dry habitats, such as Mediterranean-type evergreen or deciduous woodlands and subtropical thorn scrub (e.g., Ickert-Bond, 2003, Loera et al., 2015, 2017). *Ephedra* ranges from depressions below sea level (Death Valley, Dead Sea area) to more than 5000 m above sea level (Andes of Ecuador, Himalayas) (Fu et al., 1999, Ickert-Bond, 2005, Ickert-Bond & Renner, 2016). The genus is absent in sub-Saharan Africa and Australasia.

The majority of the *Ephedra* species are dioecious and wind-pollinated, but a few monoecious taxa with bisexual organ complexes are known to be insect-pollinated (Endress, 1997, Rydin & Bolinder, 2015). The bracts of the female strobilus (*s*, strobili, *pl*) are described as succulent, papery, and coriaceous strobili (Figure 1.3, A to C). These strobilus types are related to particular seed dispersal syndromes: endozoochory (succulent, e.g., birds and less commonly lizards, and coriaceous, e.g., Rodriguez-Pérez et al., 2012, and seed-caching rodents, e.g., Hollander and VanderWall, 2009, Loera et al., 2015), and anemochory (papery, Stapf 1889). Recently, an intermediary type (papery/coriaceous) was defined, that is also dispersed by rodents but shares traits with wind-dispersed species, respectively (Hollander & VanderWall, 2009, Loera et al., 2015). Dispersal by wind and frugivores occurs in both Old World as well as New World species, whereas seed-caching-rodent dispersal is restricted to the North American species. (Figure 1.3, D).

Commercial applications of *Ephedra* extracts derive from the ephedrine alkaloids found in the dried stems in some Eurasian species (e.g., *E. sinica*, *E. equisetina*, *E. intermedia*, White et al., 1997, Zhu, 1998, Caveney et al., 2001). The best-documented drug made from *Ephedra* is Ma-huang, used in Chinese medicine for about 5000 years to treat fever, nasal congestion, and asthma (Zhu, 1998).

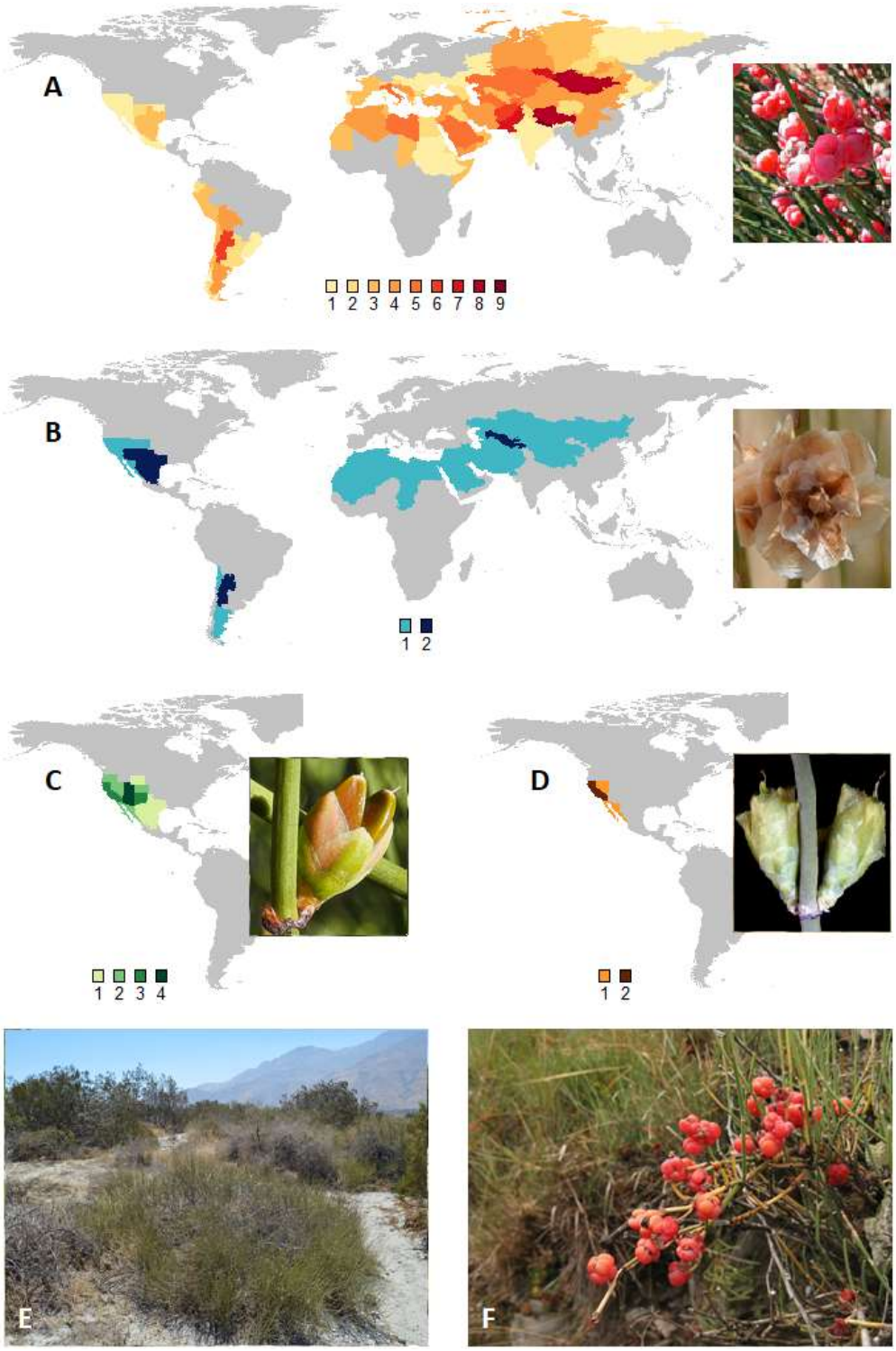


***Figure 1.3. Ephedra* global distribution overview, by strobilus type.** Color keys: number of co-occurring species per strobilus type in a particular country. Strobilus types: A, fleshy (e.g., *E. distachya*); B, papery (e.g., *E. torreyana*); C, coriaceous (e.g.,: *E. viridis*); D, intermediary (papery/coriaceous) (e.g.,: *E. funerea*). Photos A–D show female specimens with seeds (C, D: photographs by M. Baker). E and F: Life form examples of *Ephedra* plants: nanophanerophyte (*E. viridis*, left), chamaephyte (*E. distachya* ssp. *helvetica*, right). The World Checklist of Selected Plant Families (WCSP, 2018) provided the *Ephedra* country occurrences. For the maps, I used *R* (*R* Core Team, 2013), a world shapefile (TDWG, 2021), and the *R* packages *sp* (Pebesma and Bivand, 2005), *RcolorBrewer* (Neuwirth and Brewer, 2014), and *ggplot2* (Wickham, 2016).

All ephedrine-containing species are restricted to Old World *Ephedra*, which grow in the drier regions of China, North West India, and Pakistan. In New World *Ephedra* species, ephedrine alkaloids are not detectable (Caveney et al., 2001).

Among seed plants, the phylogenetic position of the Ephedraceae and related families of the Gnetales is still contentious (Rydin, 2018, Zumajo-Cardona & Ambrose, 2021). According to morphological data, Gnetales were traditionally placed as a sister to angiosperms (Crane, 1985, Doyle & Donoghue, 1986, 1992, Loconte & Stevenson, 1990), but the molecular data do not support this hypothesis. Instead, Gnetales, seem to be sister to different groups of conifers, Pinaceae and non-Pinaceae (Ruhfel et al., 2014, Wickett et al., 2014, Zumajo-Cardona & Ambrose, 2021).

Drivers of synonym numbers

The principle of priority (Turland et al., 2018), important to naming organisms, states that the accepted name is the earliest validly published name for a given species (e.g., Nicolson, 1991, Rao, 2004, Mori, 2013). The younger names are considered synonyms if more than one name describes the same species. One of four reasons to change a species' name is synonymy; and synonyms may emerge for different reasons. Synonyms are defined if more than one published name describes the same species. Following the Principle of Priority, the first species described becomes the accepted species, and the synonyms are linked to it as aliases. Name changes from synonymy usually derive either from an improved understanding of taxonomic relationships, differently interpreted interspecies variations, and from recognizing that different scientists described the same species independently (e.g., Rao, 2004, Franz et al., 2008, Turland et al., 2018). Due to improved understanding of taxonomic relationships, species are transferred from one genus to another. The reorganization resulted in a name change and a new synonym. It was, for example, diagnosed that the traits of South American *Chytroma brancoensis* R. Knuth were the same as those used to define the genus *Lecythis* (Lecythidaceae, Brazil nut trees; Mori, 2013). Thus, he established the new combination *Lecythis brancoensis* (R. Knuth) S. A. Mori. *Chytroma brancoensis* is now a synonym of *Lecythis brancoensis.*

Differently interpreted interspecific variation (morphological, ecological, and geographical differences) by different botanists are another reason for name changes and synonymy (e.g., (Valdecasas et al., 2008, Mori, 2013). *Hesperis matronalis* L. (Brassicaceae, Dame's rocket) illustrates different color variants in even side-by-side growing plants that do not merit the variants being recognized as separate species. However, S.A. Mori in 2013 split *Gustavia macarenensis* Philipson subsp. *paucisperma* S. A. Mori (Lecythidaceae, Venezuelan

population) into two species. The first species included the populations from the Andean slopes of Colombia and Ecuador, *G. macarenensis* Philipson, and the second comprised the Venezuelan population, now known as *G. paucisperma* (S. A. Mori) S. A. Mori. The split was justified based on significant morphological trait differences (Mori, 2013).

Synonyms may also emerge from different taxonomists interpreting and classifying interspecific variation differently; the two resulting philosophies are referred to as 'splitting' and 'lumping'. Consequently, if splitters work during an earlier period, some of their created species will most likely be associated and joined if other botanists who work with the same taxonomic group are lumpers. This may create confusion when species once recognized as different become synonyms of a more widely circumscribed species. The opposite happens when a splitter separates a more broadly defined species into different species (e.g., Valdecasas et al., 2008, Ickert-Bond et al., 2019). If a splitter and a lumper classify species of the same genus, the former will usually recognize more species than the latter.

The most common reason for changing a species name is that botanists described the same species more than once (Mori, 2013). For example, John Dwyer described *Gustavia superba* (Kunth) O. Berg var. *puberula* Dwyer in 1965. The same entity was described in 1974 as *G. grandibracteata* by Croat and Mori. The common characters of both specimen groups justified a rank change to species, the Principle of Priority did not apply due to rank differences, and consequently, *G. grandibracteata* Croat & S.A. Mori is the accepted name. *Gustavia superba* (Kunth) O. Berg var. *puberula* Dwyer is placed in synonymy. Likewise, *Lecythis elliptica* Kunth (published in 1825) is a synonym of the earlier published name *L. minor* Jacq. (Published in 1763).

Recent studies suggest a greater variety of reasons other than taxonomy and nomenclature why a species was possibly described more than once. It was suggested that wide-ranging species might have higher synonym numbers, as such species are often described independently and unknowingly under different names (Explanatory variable: range size; Baselga et al., 2010, Fenneman, 2017). A study presented the synonym numbers for a range of newly updated angiosperm families (Lughadha et al., 2016). The authors showed that synonymy was unevenly distributed among the studied families, and high synonym numbers were only concentrated in a few families (e.g., in the daisy family, Asteraceae, the orchid family, Orchidaceae, and the grass family, Poaceae). They explained that the results might show scientists' preferences for appealing families and genera (Pillon & Chase, 2006, Lughadha et al., 2016). In other studies, it was argued that historically described seed plant species had more time to accumulate

synonyms than species that were described more recently (Explanatory variable: Age of an accepted name, as the proxy for the time passed since the publication of its publication; Alroy, 2002, Baselga et al., 2010). Over time, when the taxonomic relationships were identified and ordered, the Principle of Priority dictated the accepted species and their aliases. As a result, the synonym number is the sum of the aliases that have become known during these identification and structuring processes. However, there might be as yet unidentified synonyms of evaluated species.

Resulting from previous studies, and the varying number of already identified synonyms per species, I hypothesize that there are more factors than taxonomy and nomenclature which drive the variation of synonyms of species. Table 3.2 provides an overview of 15 angiosperm species that accumulated high numbers of names, and some potentially important drivers of synonym numbers (Family, economic significance, botanical continent, first published [year] to calculate the age of the accepted name).

Angiosperms as the model group

The Angiosperm Phylogeny Group (APG), a collaborating group of international systematic botanists established a taxonomy of flowering plants (angiosperms) reflecting the most current knowledge about the taxonomic relationships by continuously incorporating molecular data from phylogenetic studies into long-held views of relationships based on morphology, by expert consensus (Christenhusz et al., 2015, Stevens, 2016). The development of the present APG IV phylogeny started in 1998, where a set of experts were asked to re-classify the angiosperms with the aim of avoiding taxonomic confusion. In 2008, the APG could accomplish a significant pact. In the course of physical reorganizations and moves of major European herbaria (e.g., the Natural History Museum and the Royal Botanic Gardens of Kew, London, and the Muséum National d'Histoire Naturelle, Paris; Wearn et al., 2013), a committee established to lead this project decided to follow the then APG III phylogeny (Haston et al., 2009). The decision included organizing the collections at the member herbaria in accordance to the APG. Until present, the continuous and consent-based APG re-classifying approach resulted in common and homogenously structured data across the APG and major museums worldwide (Christenhusz et al., 2015). Because of the Royal Botanical Gardens, Kew, that participated in the 2008 reorganizations, this data situation, therefore, also applies to the World Checklist of Selected Plant Families (hereafter: WCSP), which I selected as the data provider for the angiosperm analyses. In February 2020, the WCSP maintained 530,000 APG-aligned, homogenously structured, and thoroughly scrutinized plant name records from 270 seed plant

families worldwide, including 200 angiosperm families, which I used for the synonymy driver analyses and models.

My Ph.D. thesis comprises two independent studies in which I examine taxonomic and spatial limitations in biodiversity data retrieved from two major public data providers: I. Influence of different data cleaning solutions on downstream macroecological diversity models (Chapter 2), and II. Drivers of the variation in synonym numbers of angiosperm species names (Chapter 3).

The main goals of my thesis were to

1., provide the first quantitative analysis of how public provider data cleaned by different DC pipelines (pipeline data) influenced downstream species distribution models (SDM),
2., understand how the downstream SDMs and stacked SDMs (S-SDM) from pipeline data differ from the respective models from expert data that represent the gold standard, (1&2: Chapter 2), and
3., identify drivers affecting the variation in synonym numbers across angiosperm species, and the extent to which the drivers explain the synonymy in the employed angiosperm species (3: Chapter 3).

Study outline

In my thesis, I address causes, manifestations, and effects of taxonomic and spatial limitations (which are mainly, but not exclusively, data errors) in data from public providers. The individual studies include data analyses using data from the Global Biodiversity Information Facility (GBIF) and the World Checklist of Selected Plant Families (WCSP) as public providers. The different resulting datasets are applied in macroecological diversity models.

In chapter 2, I focus on North American *Ephedra* from the GBIF to analyze two aspects of pipeline-cleaned biodiversity data from global public providers. (1), I examine provider data cleaning in different DC pipelines and, subsequently, analyze performance differences (measured as data cleaning steps processed and errors identified and removed). Each DC pipeline comprises different DC methods and tools. (2), I analyze how the differences in the retained taxonomic and spatial errors per pipeline data translates into differences in the macroecological biodiversity models (single species SDMs and S-SDM) and maps.

In chapter 3, I focus on angiosperm species from the World Checklist of Selected Plant Families (WCSP) to analyze the role of five drivers of synonym numbers (higher taxa of species: families

and genera, the botanical continents where the species are present, insularity, range size of a species, and the age of their accepted name.

Research chapters

Fuehrding-Potschkat, P., Kreft, H., Ickert-Bond, S.M. (2022): Influence of different data cleaning solutions on downstream macroecological diversity models. Ecology and Evolution **12**: e9168. https://doi.org/10.1002/ece3.9168.

2

Influence of different data cleaning solutions on downstream macroecological diversity models

Abstract

Digital point-occurrence records from the Global Biodiversity Information Facility (GBIF) and other data providers enable a wide range of research in macroecology and biogeography. However, data errors may hamper immediate use. Manual data cleaning is time-consuming and often unfeasible, given that the databases may contain thousands or millions of records. Automated data cleaning pipelines are therefore of high importance. Taking North American *Ephedra* as a model, we examined how different data cleaning pipelines (using, e.g., the GBIF web application and four different *R* packages) affect downstream species distribution models (SDMs). We also assessed how data differed from expert data. From 13,889 North American *Ephedra* observations in GBIF, the pipelines removed 31.7% to 62.7% false positives, invalid coordinates, and duplicates, leading to datasets between 9,484 (GBIF application) and 5,196 records (manual-guided filtering). The expert data consisted of 704 records, comparable to data from field studies. Although differences in the absolute numbers of records were relatively large, species richness models based on stacked SDMs (S-SDM) from pipeline and expert data were strongly correlated (mean Pearson's *r* across the pipelines: 0.9986, versus the expert data: 0.9173). Our results suggest that all *R* package-based pipelines reliably identified invalid coordinates. In contrast, the GBIF-filtered data still contained both spatial and taxonomic errors. Major drawbacks emerge from the fact that no pipeline fully discovered misidentified specimens without the assistance of taxonomic expert knowledge. We conclude that application-filtered GBIF data will still need additional review to achieve higher spatial data quality. Achieving high-quality taxonomic data will require extra effort, probably by thoroughly analyzing the data for misidentified taxa, supported by experts.

Introduction

Digitally accessible species records from global data-sharing networks like the Global Biodiversity Information Facility (GBIF) provide the basis to address a wide range of biodiversity-related questions in ecology, biogeography, and other disciplines (e.g., Soberon & Peterson, 2004, Guralnick et al., 2007, Meyer et al., 2016). Such databases and data-sharing networks represent a valuable source of knowledge in which individual researchers and institutions worldwide invested considerable amount of time and resources (Wieczorek et al., 2012, Baskauf et al., 2016, Guralnick et al., 2018). However, since the circumstances and standards under which these records were collected and digitized are usually unknown, a user must assess whether the data quality provided meets the requirements of the research question (Beck et al., 2013, Sterner & Franz, 2017). Consequently, this demands data cleaning tools (hereafter: DC tool) to standardize data and identify and remove data errors. Thus, developing appropriate DC tools is a long-standing goal of biodiversity informatics (e.g., Chapman et al., 2000, Kadmon et al., 2004, Araújo & Guisan, 2006).

Data errors occur mainly along three dimensions: taxonomy, space, and time (Meyer et al., 2016). They may significantly affect common downstream analyses such as the accuracy of species distribution models (SDMs, e.g., Gueta & Carmel, 2016, Tessarolo et al., 2017, Hijmans & Elith, 2019, Zizka et al., 2019). In the taxonomic dimension, resolving misspellings (Zermoglio et al., 2016) and reconciling the synonymy of taxonomic names (Alroy, 2002, Wortley & Scotland, 2004) pose a significant challenge. The related widespread and particularly challenging problem is misidentified specimens, estimated at 50% for tropical plant specimens (Goodwin et al., 2015) and ranging from 5% to nearly 60% in the Zoological Record database (Meier & Dikow, 2004). In the spatial dimension, errors in and low precision of coordinates, e.g., from rounding of the decimal digits, swapped latitude and longitude, missing coordinates, or coordinates with zero-values are common data-quality problems (e.g., Yesson et al., 2007, Otegui et al., 2013, Topel et al., 2017). Lower geospatial accuracy is frequently assumed for older records than for those collected more recently (Tessarolo et al., 2017, Zizka et al., 2020). Stropp et al. (2016) showed, for instance, that conspicuous records of flowering plants collected in Africa before the 1960s were filtered out due to poor data quality. Another issue associated with older records is that the probability increases that populations no longer exist at a given sampling location over time due to natural or anthropogenic reasons (Meyer et al., 2016).

Even for experts, identifying and resolving data quality issues manually is in many cases unfeasible, given that datasets typically contain thousands to millions of records. Therefore, selective DC strategies based on well-explained instructions and automated DC tools that reproducibly generate high-quality data are especially in high demand for inexperienced users (Zizka et al., 2019). Downstream applications such as conventional SDMs depend on this data quality (e.g., Guisan et al., 2017, Raes & Aguirre-Gutierrez, 2018, Araújo et al., 2019). Data scientists and biodiversity informaticians approached the development of DC solutions from several angles: (1) DC tools that generally solve thematically limited requirements, like retrieving, evaluating, formatting, completing, and organizing data. This type of DC solution was implemented in the widely used *Tidyverse* "umbrella" package (Wickham et al., 2019). The solution was also included in specialized packages such as *CoordinateClearer* (Zizka et al., 2019), *rgbif* (Chamberlain et al., 2020), and the GBIF web application (GBIF.org, 2020). (2) Manuals supporting the preparation of data for SDMs. Particular *R* packages are an integral part of such manuals (e.g., Chapman, 2005, Guisan et al., 2017, Hijmans & Elith, 2019). The manuals consist of verbal explanations and coded instructions, which the user can apply (e.g., per package *dismo,* Hijmans et al., 2020). While the newly developed and recently updated methods for automated cleaning of records are promising, their effect on commonly applied SDMs remains poorly examined (see Schmidt-Lebuhn et al., 2013, Hijmans et al., 2017, Zizka et al., 2020).

Pipelines play an important role in the scientific domain when, for example, biodiversity data from different sources such as herbarium vouchers and observations need to be combined for analysis. In this study, we investigated the performance of six pipelines (P1 to P6) using various DC tools and how these pipelines affected downstream SDMs. We used North American *Ephedra* species as the model organisms (Ephedraceae, Gnetales; Cutler, 1939; Stevenson, 1993, Figure 2.2, A to C; Appendix, Table A1) and GBIF as the data source. With over 2.1 billion species records worldwide, GBIF is the largest and one of the most frequented public providers of biodiversity data. It is often the primary data source for many researchers (Guralnick et al., 2018, Hobern et al., 2019, Zizka et al., 2020). Thus, we selected the GBIF records as input to the pipelines. In this context, we address three questions:

1. How do the pipelines differ in their performance? We expect that different DC tools will generate different result datasets.

2. How do differences in pipeline data affect downstream diversity models and maps (observed, predicted)? We expect the pipeline datasets to differ in the resulting models (single species- and stacked SDMs, hereafter: S-SDM) and maps.

3. How does the pipeline data - after being cleaned by the pipelines – differ from the expert data (observed and predicted), assuming that the expert data represent the most accurate *Ephedra* environmental and geographical range? We expect the quality of the pipeline data to differ from the expert data. The differences will be measurable (occurrences and correlations) in the models and maps.

We analyzed to which extent the data from the different pipelines led to different species constellations and numbers in the grid cells and visualized the differences in diversity maps created from S-SDMs. Finally, we discuss how realistic the results from GBIF data and expert data reflect the environmental or geographical extent of the *Ephedra* species' ranges.

Material and methods

In North America, *Ephedra* species are characteristic components of arid and semi-arid regions of the southwestern USA and Mexico (Hollander & VanderWall, 2009, Loera et al., 2015). They occur from the Death Valley to about 2,500 m in the Rocky Mountains (Stevenson, 1993). The species share a morphologically reduced, uniform growth habit with mostly leafless, photosynthetic stems (Ickert-Bond & Renner, 2016). Specimens are collected frequently, as shown by the record numbers of the public providers (e.g., GBIF: 46,384 records worldwide), and high-quality expert data is available for the New World species (Ickert-Bond, 2003). The coordinates served as the proxy for the *Ephedra* species' characteristic locations (response variables), from which we developed species SDMs and genus S-SDMs for North America.

We monitored changes in similarities and correlations using the validated records from P1 to P6 and the expert data (observed occurrences, hereafter: L1; Table 2.2). From L1, we developed L2 and L3 data of the North American *Ephedra* species and their occupied grid cells (per pipeline and the expert data). L2 included the grid cell numbers an *Ephedra* species occupied, and L3 counted the concurrent *Ephedra* species per grid cell. L4 data comprised the correlations of the observed occupied grid cells. The L5 data (pipeline and expert) included the predicted distribution in S-SDMs across the pipelines and expert data (L2/L4, and L5: Spatial autocorrelation by Moran's *I* and correlation between two random variables by Pearson's *r*). (Figure 2.3).

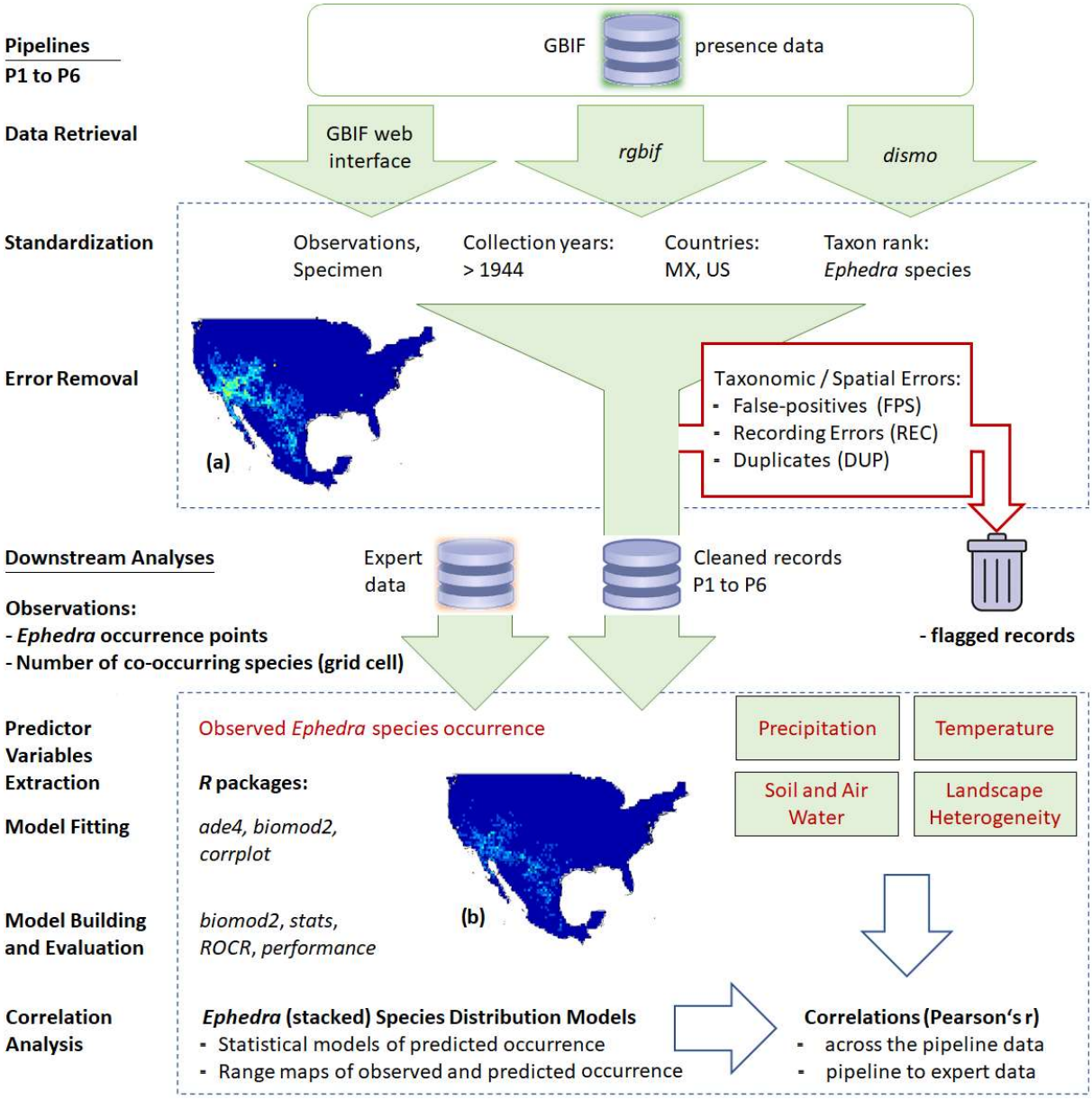


Figure 2.1. **Workflow of the pipelines and the downstream analyses.** The pipelines' part comprised the following sections: Data Retrieval, Standardization, and Error Removal. The Downstream Analysis featured the Predictor Variables Extraction, the Model Fitting, the Model Building (SDMs, S-SDMs) and Evaluation, and the Correlation Analysis, developed from the pipeline data P1 to P6 and the expert data. *R* packages used in the course of the workflow are in italics. (a) Observed species distribution from GBIF P1 data. (b) Observed species distribution from expert data. Filter categories: DUP = Duplicate records, FPS = False-positives, REC = Recording Errors.

Data pipelines

Ensuring comparability across six pipelines, the process chain of filters provided identical conditions to optimize the provider data (See Table 2.1, the filters of the pipelines). The chain consisted of (1) selecting and retrieving data from GBIF, (2) standardizing the records by filtering, and (3) correcting or removing data errors (Figure 2.1, Table 2.2. At each pipeline

step, we employed one or more DC tools (e.g., Chapman, 2005, Hijmans & Elith, 2019, Zizka et al., 2019a). The selected tools (e.g., GBIF web application, written instructions, or *R* packages) or their most recent updates were released between 2005 and 2020 and are free of charge. In some pipelines, the three steps were performed by one ("three-in-one") DC tool. In the setup of the process chain, we followed the data cleaning recommendations given by the respective DC tool's authors and pertinent best-practice guidelines (Araujo et al., 2019, Guisan et al., 2017).

We retrieved data from GBIF (gbif.org, 2020) on November 18, 2020, in four different ways: (1) The filter "*Ephedra* L." (hereafter: GBIF (I)) retrieved 46,384 records for P5, P6, and the P0 benchmark data using the "three-in-one" GBIF web application (GBIF, 2020a). (2) The filter set "*Ephedra* L. specimens of North America, from 1945 to 2019" (hereafter: GBIF (II)) selected 9,484 records for the P1 process chain using the web application (GBIF, 2020b). In both cases, the data were download with the web application. (3) *rgbif*, a "three-in-one" tool, employed its integrated functionality to standardize the P2 and P3 data and retrieved 6,687 GBIF records into the userspace. (4) *dismo* selected 46,384 GBIF records for P4 and retrieved them into the userspace. (Details see Table 2.2).

We created the P0 data for comparison. It served as the benchmark of standardization and errors, delivered by the GBIF data, which the DC tools could have removed in the pipelines. However, P0 was not itself a pipeline nor was it part of any pipeline. We performed an inventory of the dataset and the data errors that might influence the quality of the downstream models (Table 2.2, P0 column). Using P0, we could identify questionable records and the degree of feasibility to which each pipeline removed such records. After data retrieval, further data cleaning was performed in P3, P4, P5, and P6 by basic *R* code, the *dplyr* package (of *Tidyverse*, Wickham et al., 2019), and the *CoordinateCleaner* (Zizka et al., 2019), in different combinations (Table 2.2). We selected records of taxon rank "species" (Claridge et al., 1997, Reydon, 2019), filtered for North America (Mexico, USA) and collection years 1945 to 2020 (Zizka et al., 2020). As the basis of records, we selected specimens and observations. During error removal, we focused on taxonomic and spatial errors (Meyer et al., 2016), such as non-native specimens, missing or zero values, and sea coordinates. We also removed false-positive records reporting, e.g., occurrences at biodiversity institutions, and geographic outliers. From the P0 evaluation, we were aware of two false-positive occurrences (Figure 2.2, Marker 2) hidden in the data. We found these errors challenging to be recognized by any tool. Therefore, we removed one of these errors in P4, and two in P5 and P6, using basic *R* code.

Table 2.1. **Pipeline filter summary for standardization and error removal.**

Categories	Filter	Requirement	Rationale
STD	Country range	Spatial	North America: Mexico and the USA
STD	Infraspecific rank	Taxonomic	Required rank: species (Claridge et al., 1997, Reydon, 2019), infraspecific ranks (e.g., subspecies, hybrids) to be omitted.
STD	Collection years	Temporal	1945 to 2020, as older records are more likely to contain erroneous coordinates (Zizka et al., 2020).
STD	Basis of record	Consistency	Specimens and observations.
STD	Occurrence status	Consistency	Presence data.
FPS	Non-North America-native *Ephedra* species	Taxon	All non-native *Ephedra* species that are allocated to the North American countries either by mistake or are artificially introduced, e.g., to Botanical Gardens.
FPS/REC	Zero or missing coordinates	Spatial	Zeroes and missing values may represent records with data entry errors. Missing values will cause error messages in *ade4*.
REC	Longitude and latitude are equal	Spatial	Equal longitude and latitude may represent records with data entry errors.
DUP	Duplicate records	Consistency	Duplicate records that may represent e.g., record copy errors.
FPS	Country capitals	Spatial	Records that may contain the coordinates of the country capital.
FPS	Country centroids	Spatial	Records that may contain the centroid coordinates of the country.
FPS	GBIF headquarters	Spatial	Records that may contain the coordinates of the GBIF headquarters.
FPS	Biodiversity institutions	Spatial	Records that may contain the coordinates of biodiversity institutions where the herbarium voucher is stored.
FPS	Geographic outliers	Spatial	Geographic outliers that may represent misidentified specimens.
REC	Urban areas	Spatial	Records from urban areas that may represent old data or vague locality descriptions.
REC	dd.mm to dd.dd conversion errors	Spatial	Records with ddmm to dd.dd conversion error (misinterpretation of the degree sign as decimal delimiter).
REC	Rasterized collections	Spatial	Records with a significant proportion of coordinates that might have a low precision.
FPS	"Manual" removal of false-positives	Consistency	False-positives that have been overlooked by automatted error-removal, based on the knowledge that they are in the records.

Categories: DUP, duplicate records; FPS, false-positives; REC, recording errors; STD, standardization.

As coordinates with three or fewer decimal places often indicate they were obtained from grid maps (Zizka et al., 2019), we permitted only validated coordinates with no less than four decimal places. However, this precision was not required for the modeling. The *CoordinateCleaner* identified specimens of urban areas and flagged them for scrutiny. We searched for duplicates based on the variables: species, coordinates, and collection date, respectively, and removed them. Finalizing the process chains, we excluded native species for which the sample size was lower than fifty occurrences to avoid biased models and maps (Guisan et al., 2017, Hijmans & Elith, 2019). (Usage of the tools in the pipelines, see Table 2.2). At the end of the pipelines, we examined the retained records and errors in the pipelines' datasets in comparison to P0 (data at L1).

Downstream analysis

Data from examination of physical herbarium specimens and field studies (Ickert-Bond, 2003) represented the most realistic environmental and geographical range ("gold standard", Araujo et al., 2019) of the genus *Ephedra* in North America. The expert dataset comprised 4,081 records of New World *Ephedra* specimens from herbaria with large holdings of *Ephedra* in both North and South America (e.g., ARIZ, ASU, HUH, NY, RM, SGO, SI, TEX, UC, UNAM, US; herbarium acronyms according to Thiers, 2022). 704 records of twelve *Ephedra* species (L1) were selected for North America; however, they were not processed in a pipeline. We applied standardization conditions only for comparability. The records contained confirmed taxa, examined coordinates, and detailed locality descriptions comparable to field-collected data. We considered an overlap of 90 records of 13,889 from GBIF and the expert dataset negligible. As *Ephedra* is adapted to dry environments, we imported nineteen temperature and precipitation variables from the CHELSA climatology (Karger et al., 2017), elevation data as a proxy for landscape heterogeneity (GMTED, 2020) and plant-available water data (Zhang et al., 2018). From their habitat description (e.g., Cutler, 1939, Stevenson, 1993), we assumed the selected environmental data being ecologically relevant.

For the SDMs and S-SDMs, we created a grid of 4017 cells across Mexico and the USA (30 arc minutes, WGS84) using wrld_simple (*R* package *maptools*, Bivand & Lewin-Koh, 2017) and *raster* (Hijmans et al., 2016). The grid size reasonably showed the co-occurring species, which was not the case on different scales. We aggregated the environmental data to the grid resolution (*sp* package, version 1.4-5, Pebesma & Bivand, 2005, Bivand et al., 2013) and extracted the values for each occurrence (*raster*). We built a presence-absence table, creating a random selection of pseudo-absences for each *Ephedra* species using the *R* package *biomod2*

(Thuiller et al., 2016). We tested the localities where *Ephedra* species were not recorded (*R* package *ecospat*, Di Cola et al., 2017). We anticipated environmental conditions to cause absence (Stevenson, 1993, Loera et al., 2015), making sure that the localities used for fitting the model represented the requirements of the species across North America (Training area, Guisan et al., 2017). We summed-up the species present in the grid cells as the number of co-occurring species. (L2, L3).

We identified the contributing predictors (using *R* packages *ade4*, Bougeard & Dray, 2018 and *corrplot*, Wei et al., 2017). From the 21 variables, we selected a subset of reasonably uncorrelated variables per species using *biomod2* (Appendix, Table A2; Thuiller et al., 2014, Guisan et al., 2017). Reasonably uncorrelated refers to being below the recommended threshold of 0.7 (Dormann et al., 2013). As goodness-of-fit evidence we used the Akaike Information Criterion (AIC; Johnson & Omland, 2004), and Tjur's R^2 (Coefficient of Discrimination for binary outcomes; *R* package *performance*, Lüdecke et al., 2021) to identify the variables with the highest impact (Table A2). Finally, we fitted logistic regression models for the *Ephedra* occurrences using glm as the model and "binomial" as the distribution family. The threshold value of a high-performance index (0.9, Guisan et al., 2017) was used to evaluate the predictive accuracy of the model, particularly the Receiver Operating Characteristic Curve (ROC) and the area under the curve (AUC) (*R* packages *biomod2* and *ROCR*, Sing et al., 2005). We stacked the predictions of the twelve *Ephedra* species resulting from the different pipelines as well as the expert data to S-SDMs (without using thresholds; Calabrese et al., 2014, Guisan et al., 2017, Biber et al., 2020). The correlations between the observed and the predicted *Ephedra* occurrences informed how strongly the differences between the pipelines and the expert data affected the respective SDMs and S-SDMs (L5).

We inspected spatial autocorrelation (L2/L4: grid occupation, L5: predicted distributions) using the Moran's *I* coefficient (*R* package *spdep*, Bivand et al., 2015). We computed the correlations of the observed and predicted *Ephedra* occurrences in two pipelines (the least-cleaned data, P1, and the most cleaned data, P6) and the expert data using Pearson's *r* (*R* package *rstatix*, Kassambara, 2020). Ultimately, we visualized them as map pairs (Figure 2.4); and to adequately represent the species richness in the maps, we chose eleven breaks (R package *classInt*, Bivand et al., 2015) for the maximum possible co-occurring species.

Results

The GBIF web interface using GBIF (I) filters and *dismo* retrieved 46,384 unstandardized and uncleaned, globally distributed *Ephedra* records. The GBIF web interface using GBIF (II) filters retrieved 9,484 partially standardized *Ephedra* records from North America. *rgbif* retrieved 6,687 somewhat standardized specimen records from North America and already removed significant spatial errors. (Download results see Table 2.2). The three tools stopped after the data retrieval.

P0 benchmark data

13,889 P0 records represented the unstandardized and uncleaned GBIF North American *Ephedra* data. 1,979 specimens were collected or observed in Mexico (14.2%) and 11,910 in the USA (85.8%). The majority of species records consisted of North America-native *E. viridis* (19.0%), *E. aspera* (14.4%), *E. californica* (14.1%), *E. nevadensis* (13.3%), *E. trifurca* (11.9%), and *E. torreyana* (8.7%), a total of 81.4% for six species. Another six native species, *E. antisyphilitica* (4.4%), *E. funerea* (2.4%), *E. fasciculata* (1.8%), *E. pedunculata* (1.5%), *E. compacta* (1.3%), and *E. cutleri* (1.1%) totaled 12.5%. The remaining 6.1% were non-native (55 taxonomic false-positives of South American and Eurasian origin) or indeterminate specimens (499 specimens of genus *Ephedra* L.). Several standardization conditions and errors coincided in the same record. Thus, the number of removed records did not correspond to the sum of the identified errors. 5,187 records (37.3%) were flagged as fit for use for the downstream analyses. 8,702 records (63.7%) were marked for removal due to one or more significant errors. Missing coordinates (5,978 records, 43.1%) represented the majority of identified data errors, followed by the sampling year (4,329 records, 31.1%, were older than 1945) and the duplicate records (3,584 records, 25.8%). 220 records showed coordinates in bodies of water. With two exceptions, the non-native *Ephedra* species were, e.g., found in botanical gardens and scientific institutes (e.g., Atlanta Botanical Garden; Figure 2.2D, locality markers 3, 4, 10, and 11). As a few non-native species contain medicinally active substances, they were reported with two records from a shop in Berkeley (*E. sinica*, Figure 2.2D, locality markers 8 and 9) and one record from an herbal product shop in Seattle (*E. sinica*, Figure 2.2D, locality marker 1). We detected *E. nevadensis* at the University of Connecticut (Figure 2.2D, locality marker 2), yet this species is native to the Southwestern United States. Three records revealed misplaced taxa by comparing the verbatim locality description with the coordinates. These errors were not identified by a tool, only by scrutiny.

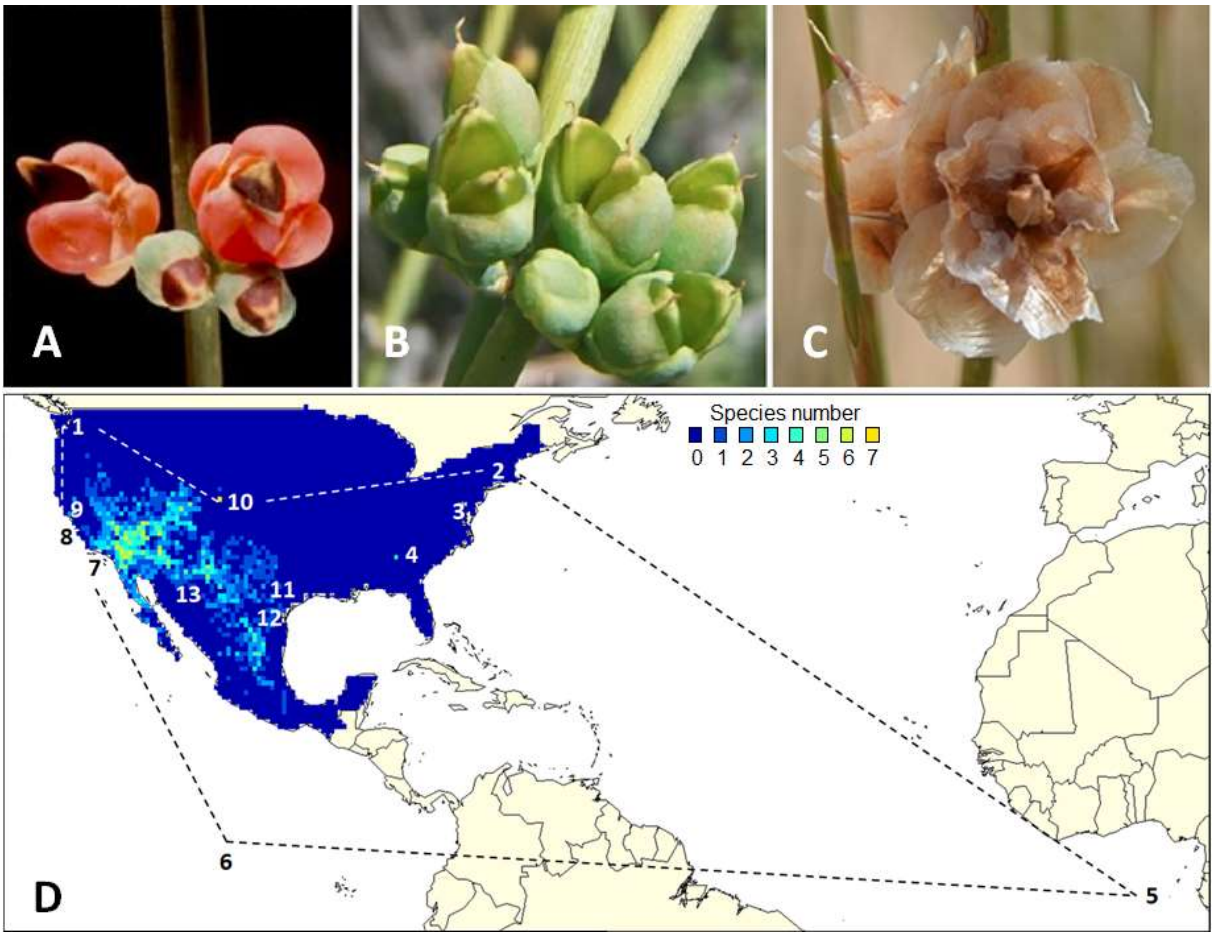


Figure 2.2. **A-C. North America-native *Ephedra* specimens (female specimens with seeds).** *E. antisyphilitica*, *E. nevadensis*, and *E. trifurca* (left to right**). D. Representation of false-positive taxonomic and spatial errors in the *Ephedra* dataset (Examples)**. Markers 1, 8, and 9 were specimens from shops in Seattle and Berkeley. Markers 3, 4, 10, and 11 were non-native species from botanical gardens and scientific institutes. Marker 2 pointed to a North America-native species at the University of Connecticut, NY. Markers 5 to 7 showed coordinate errors that the verbatim locality description can only identify. The species at markers 12 and 13 were misidentified, as the documented species do not occur naturally at these localities. Data for the map: P1 (L3, number of co-occurring species). Color coding of the map: P1 observed distribution (Figure 2.4).

Locality marker 12 referenced a misidentified specimen (*E. distachya*, Figure 2.2D) that does not naturally occur in Coahuila, Mexico. The specimen that locality marker 13 referenced (*E. trifurcata*, Figure 2.2D) might be a misspelling of *E. trifurca* (P0 results, see Table 2.2, Table A1).

Expert data

577 of 2,251 specimens were collected or reviewed from Mexico (25.3%) and 1,674 specimens from the US (75.4%). After standardization, 704 records remained (210 records of Mexican specimens, 494 records of US specimens). After standardization, the majority of records

(65.2%) were allocated to *E. aspera* (22.3%), *E. trifurca* (21%), *E. fasciculata* (11.4%), and *E. antisyphilitica* (10.5%). The other eight species, *E. viridis* (7.1%), *E. californica* (5.4%), *E. torreyana* (5.1%), *E. funerea* (4.5%), *E. compacta* (3.7%), *E. pedunculata* (3.3%), *E. nevadensis* (2.3%), and *E. cutleri* (1.4%) totaled 32.8% of the standardized records. The remaining 14 records (2%) were of other taxonomic ranks.

Effects of differences in the pipeline data on diversity models

P1and P2 were partly standardized in their process chain. GBIF (II) of P1 met four out of five standardization requirements. Explicit error removal did not occur; however, P1 implicitly removed 3,386 missing coordinate records as a side effect of the standardization. It left 2,592 missing coordinates records, 296 indeterminate records, and 33 South American and Eurasian species in the P1 dataset. P2's *rgbif* met three standardization requirements but the resulting data still contained infraspecific ranks. *rgbif* standardized the P2 data partly, using the parametrized standardization criteria, and, in addition, the built-in error exclusion parameter of invalid coordinates was employed. Except for excluding missing values in the coordinates, P2 removed no other spatial errors. P3, P4, P5, and P6 continued their respective process chains. The pipelines removed between 43.1% and 45.3% of all spatial error types (e.g., the complete subset of 5,986 missing coordinates records, see Table 2.2). P3 used the *dplyr* and *CoodinateCleaner*, providing 5,189 records to the downstream analyses. In P4, we fully standardized the data, using instructions explained in a tutorial (Hijmans & Elith, 2019) and basic *R* code. P4 provided 5,387 records to the downstream analyses. In P5, we standardized the data and removed errors, using basic *R* code and the *dplyr*. P5 provided 5,386 records to the downstream analyses. P6 used instructions from Chapman (2005) translated to basic *R* code and *dplyr* functionality to handle taxonomic errors. The *CoordinateCleaner* removed spatial errors. P6 identified 5,187 fit-for use records for the downstream analyses. Due to not meeting the sampling size criteria, we manually removed *Ephedra coryi* records from the pipelines. At the end of the pipelines, the records for the downstream analyses varied considerably and ranged from 9,484 (P1) to 5,187 (P6) (L1). (Table 2.2).

The cleaned datasets differed by 4,288 (P1 versus P6), and the number of occupied grid cells by 26 grid cells (maximum). We observed similarly clustered occupancy patterns in the distribution maps regardless of the pipeline since most records were allocated to the same grid cells per species. The occupied grid cells in the stacked *Ephedra* range maps varied between 636 and 610 (P1 versus P6 data).

***Table 2.2*. Results of the pipelines' data cleaning performance, compared to the P0 benchmark dataset (Summary table).** The color-coded cells of P1 to P6 datasets indicate the activity of a particular DC tool (color code see below). The blue cells of the P0 benchmark indicate the number of *Ephedra* records in GBIF, quantified by standardization and error category. Records which did not comply to the standardization conditions or were erroneous in the context of this study, were flagged (flg). Since several standarization conditions and errors coincided in the same record, the number of removed records did not correspond to the sum of the identified errors. The P1, P2, and P3 data retrieval tools partially standardized the data and eliminated several errors ("three-in-one" tools). Thus, the number of records retrieved differed significantly from P4 to P6, and P0. The removed records in these pipelines could only be reconstructed as differences of sub-categories (e.g., in-scope countries, collection year, null and zero coordinates) in comparison to P0. The difference between P3 and P2 resulted from the added *dplyr* and *CC* packages, which increased standardization and removed still more erroneous records. Using the added packages ensured more insight into data cleaning. Abbreviation: *CC* (→ P3/P6) = *R* package *CoordinateCleaner*.

Pipeline datasets	**P1**	**P2**	**P3**	**P4**	**P5**	**P6**	**P0 benchmark**
Input: Data retrieved by	**GBIF (II)**	***rgbif***	***rgbif***	***dismo***	**GBIF (I)**	**GBIF (I)**	**GBIF (I)**
Number of records retrieved	9,484	6,687	6,687	46,384	46,384	46,384	46,384
Non-native *Ephedra* species outside North America	NA	NA	NA	**32,495, rem**	**32,495, rem**	**32,495, rem**	**32,495, flg**
Number of records passed to the standardization	**NA**	**NA**	**6,687**	**13,889**	**13,889**	**13,889**	**13,889**
Data standardized by	**GBIF**	***rgbif***	***rgbif, dplyr***	**R code**	**GBIF (I), *dplyr***	**GBIF (I), *dplyr, CC*, R code**	
North America-sampled *Ephedra* specimens (MX, US)	9,484	6,687	6,687	13,889	13,889	13,889	13,889
Occurrence status: presence	default	default	default	default	default	default	default
Non-native *Ephedra* specimens in North America	31, ret	0	0	**55, rem**	**55, rem**	**55, rem**	**55, flg**
Not identifiable specimens in North America (e.g., genus level, fossil)	296, ret	0	0	**501, rem**	**501, rem**	**501, rem**	**501, flg**
North America-native, taxon rank: species	9,010	6,687	6,678	13,240	13,240	13,240	13,240
Infraspecific ranks	147, ret	0	**0**	**704, rem**	**704, rem**	**704, rem**	**704, flg**
Collection years: >1944	9,484	6,687	6,687	9,560	9,560	9,560	9,560
Collection years: < 1945	NA	NA	NA	**4,329, rem**	**4,329, rem**	**4,329, rem**	**4,329, flg**
Basis of record: observations, specimens	9,484	6,560	6,560	13,762	13,762	13,762	13,762
Other basis of records	NA	127, ret	**127, rem**	NA	**127, rem**	**127, rem**	**127, flg**
Number of records passed to the data cleaning	**NA**	**NA**	**6,560**	**8,300**	**8,173**	**8,173**	

Pipeline datasets	P1	P2	P3	P4	P5	P6	P0 benchmark
Data cleaned by	**NA**	***rgbif***	***rgbif*, *CC***	**R code**	***dplyr*, R code**	***CC*, R code**	
NULL coordinates (Missing values)	2,592, ret	**rem**	**rem**	**1,852, rem**	**1,758, rem**	**1766, rem**	**5,978, flg**
Zero coordinates	8, ret	**rem**	**rem**	**8, rem**	**8, rem**		**8, flg**
Longitude and latitude are equal	8, ret	8, ret	**8, rem**	**12, rem**	**12, rem**	**12, rem**	**22, flg**
Duplicate records (species,longitude,latitude,year,month,day)	1,086, ret	1,226, ret	**1,182, rem**	**1,031, rem**	**998, rem**	**1,000, rem**	**3,584, flg**
Country capitals	1, ret	NA	NA	1, ret	1, ret	**1, rem**	**1, flg**
Country centroids	9, ret	8, ret	**23, rem**	23, ret	23, ret	**23, rem**	**23, flg**
GBIF headquarters	NA	NA	NA	NA	NA	NA	NA
Biodiversity institutions	33, ret	19, ret	**19, rem**	36, ret	36, ret	**36, rem**	**36, flg**
Geographic outliers	12, ret	12, ret	**12, rem**	35, ret	35, ret	**35, rem**	**35, flg**
Sea coordinates	146, ret	67, ret	**67, rem**	228, ret	228, ret	**61, rem**	**228, flg**
Urban areas	193, ret	165, ret	165, ret	298, ret	298, ret	2, ret	**298, flg**
dd.mm to dd.dd conversion errors	202, ret	202, ret	0	278, ret	278, ret	**0, rem**	**278, flg**
Rasterized collections, possibly reduced coordinate precision	56, ret	56, ret	**56, rem**	56, ret	56, ret	**41, rem**	**56, flg**
Unidentified false-positives (manually identified and removed)	2, ret	2, ret	2, ret	**1, rem**	**2, rem**	**2, rem**	**2, flg**
Number of records passed to the data finalization	**NA**	**NA**	**5,198**	**5,396**	**5,395**	**5,196**	
Data standardized and finalized by	**NA**	**NA**	**R code**	**R code**	**R code**	**R code**	
Native *Ephedra* species, sample size < 50 occ points	53, ret	9, ret	**9, rem**	**9, rem**	**9, rem**	**9, rem**	**93, flg**
Output: Final number of cleaned records	**9,484**	**6,687**	**5,189**	**5,387**	**5,386**	**5,187**	**13,889**

Color code key	Data retrieval	Data cleaning
dismo	Y	NA
basic *R* code	NA	Y
coordinateCleaner	NA	Y
dplyr	NA	Y
GBIF (I)	Y	Y
GBIF (II)	Y	Y
rgbif	Y	Y

Comparisons of highly correlated occupied grid cells (mean Pearson's *r* across the pipelines: 0.9956) were confirmed by highly correlated maps of observed *Ephedra* distribution with well-defined clusters (Figure 2.3, and Figure 2.4, P1 and P6 map pairs). Moran's *I* confirmed the spatially clustered patterns of the *Ephedra* species (observed P1/P6 Moran's *I*: 0.144, observed expert data's Moran's *I*: 0.087, p-value: significant) (L2/L4). *Ephedra californica* occurrences occupied identical grid cells across all six pipelines; therefore, the Pearson correlation coefficient was 1. For the other eleven *Ephedra* species, the occupancy of the grid cells varied slightly across the pipelines, depending on the respective pipelines compared. For example, in *E. fasciculata*, P1 differed from P6 with 49 versus 53 occupied grid cells (92.5% identical occupancy), while the occupancy in P2 and P3 in *E. antisyphilitica* was again identical (Pearson's $r = 1$). The evaluation of the S-SDMs showed that the grid cell occupancy patterns (observed occurrences) continued in the species distribution maps (predicted occurrences). Correlograms based on residual analysis are listed in Appendix Figure A4.

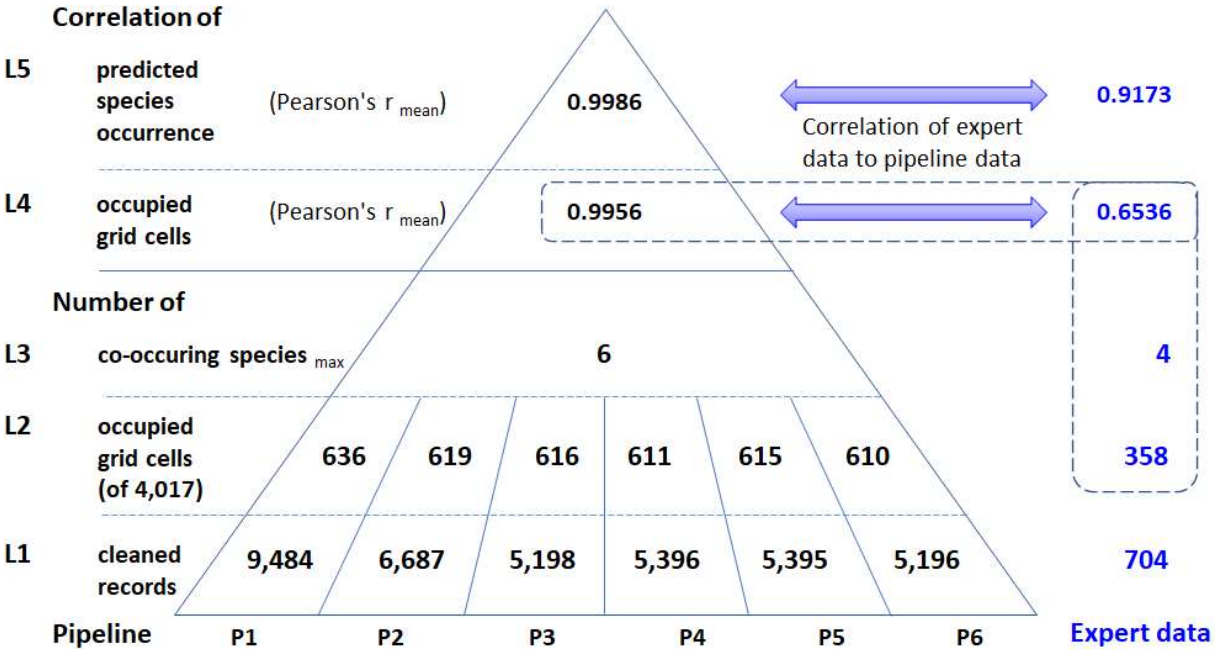


Figure 2.3. **Information-condensing pyramid of the pipelines and the expert data (L1 to L5: Condensing levels of the data).** The data show an increasingly higher correlation from the bottom to the top of the pyramid, which results from data transformations into an increasingly higher-condensed species occurrence information state. The 704 expert data occurrences (L1) were allocated into 358 grid cells (L2, with a maximum of four co-occurring species, L3). The correlation of 0.6536 (L4, mean Pearson's *r* of pairings (P1 to P6/expert)) was compared to the mean of the pairings P1 to P6. At this level (L4), the minimum Pearson's *r*-value of the occupied grid cells from pipeline data was 0.9920 (pair: P1/P6), and the maximum Pearson's *r* value was 0.9999 (pair: P4/P5). At the L5 level, the minimum Pearson's *r* value was 0.9951 (pair: P1/P6), and the maximum Pearson's *r* value was 1.0000 (pair: P4/P5). Dashed box: Expert data comparison numbers, L2 to L4.

Post-pipelines, we found that the *ade4* indicated coordinates with missing values as invalid in records containing this error type, hence, may also be regarded as a testing point for missing values in the coordinates. (Note that we did not intervene in the data cleaning in P1 by GBIF (II). Thus, records with missing values in coordinates were preserved).

The final number of predictors for the species ranged from four (*Ephedra aspera*) to ten (*Ephedra viridis*) (Table A2). The area-under-the-curve (AUC) scored from 0.9355 (*Ephedra antisyphilitica*) to 0.9990 (*Ephedra nevadensis*) (AUC mean: 0.9825). The AIC decreased to a stable minimum value in the variable's combination tests, indicating the best possible model performance compared to the other variable combinations. Therefore, we considered our models as adequately accurate to describe the distribution of the *Ephedra* species with the identified explanatory variables. The differences in the pipelines had a minor effect on the correlations, models, and maps at L4 and L5. At level L4, the mean Pearson's *r* of the occupied grid cells across the pipelines was 0.9956 (P1/P6 pair: 0.9920, minimum; P4/P5 pair: 0.9999, maximum). The high correlation led to maps of observed *Ephedra* distribution that showed also only insignificant differences (Figure 2.4, P1 and P6 observed distribution). Across the six pipelines, the predicted probability of occurrence from the S-SDMs indicated high correlations (mean Pearson's $r = 0.9986$, Figure 2.3, L5). Figure 2.4 displays the maps of the predicted distribution based on the S-SDMs.

Differences between pipeline data and expert data

The 704 expert data occurrences (L1) were allocated into 358 grid cells, with a maximum of four co-occurring species (L3). Across the pipelines, 294.5 of the average 630.5 grid cells (46.7%) showed occupancy by one species, compared to 265 of 358 grid cells (74.0%) of the expert data. 42.6 of the grid cells showed occupancy by four species (6.7%), compared to the maximum of four species (1.1%) of the expert data. Ten grid cells showed occupancy by the maximum of six species (1.6%) in the pipeline data (L2). The correlations differed clearly between the pipelines and the expert data. At level L4, the mean Pearson's *r* of the occupied grid cells for pipeline data correlated to the expert data was 0.6536 (L4: Figure 2.3). The correlation of the predicted occurrence probabilities in the S-SDMs showed a mean Pearson's *r* of 0.9173. Across the different pipelines and the expert data, the observed diversity in the maps from the S-SDMs showed a large *Ephedra* diversity center in Southern California. It continued to the North into Arizona and Nevada, and to the South into the states of Baja California and Sonora, Mexico with a predicted *Ephedra* diversity greater than seven species.

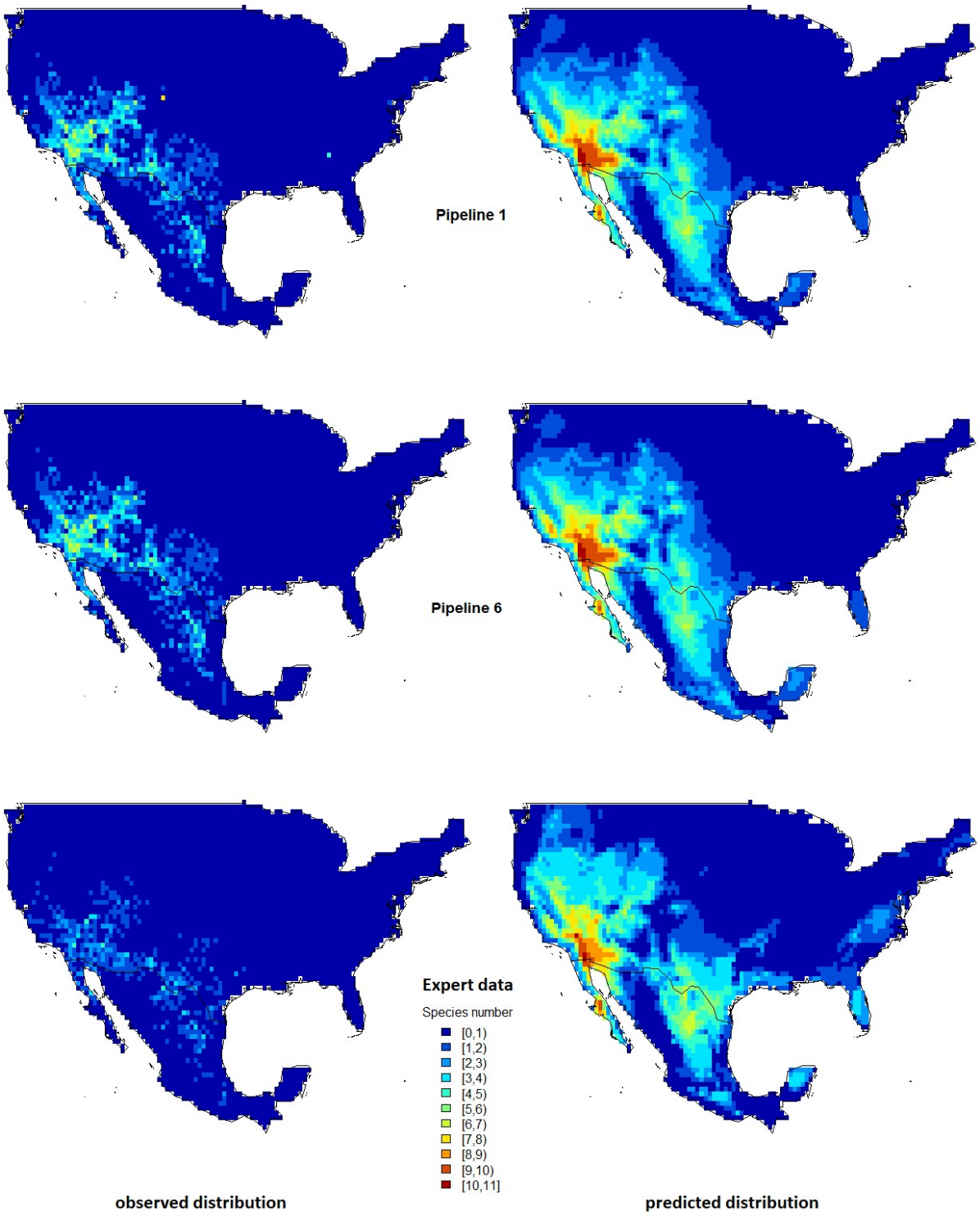


Figure 2.4. **Stacked species distribution maps based on cleaned GBIF data from pipelines P1, P6, and expert data.** Depicted are the maps of the least-cleaning P1 and the most-cleaning P6 that show only minor differences (the maps from the other pipeline data are close to P6). The control data map from the expert data shows differences to the pipelines. Left: Observed distribution (L2 data). Point-occurrences after passing the pipelines, allocated to grid cells of a stacked range map of all *Ephedra* species. The expert map shows less occupied grid cells (n = 358) than P1 (n = 636) resulting in a smaller range. Right: Map of the predicted probability of species from S-SDMs (L5 data). The colour keys show highly correlated patterns of each data quality (P1, P6, and expert data: 0 to 12 species, Pearson's $r = 0.9173$).

A second diversity center emerged across the state of Texas, USA, and continued into the states of Chihuahua, Coahuila, Nuevo Léon, and Tamaulipas, Mexico, with a predicted *Ephedra* diversity of up to seven species. (L5). The diversity patterns in the expert data, although similar in shape, were less distinct (Figure 2.4).

Discussion

We analyzed the data cleaning performance of six different pipelines for digital point-occurrence records and their effects on species distribution models, a common downstream application in macroecology. The six pipelines differed significantly in the number of accepted species, errors removed, and remaining records for analysis (Table 2.2, Table A1). For example, P6 removed the most significant number of records, approximately twice as many records as the least-cleaning pipeline P1. Data from P1 differed from the other group by hosting seventeen non-native species in addition to the twelve natives, all of which were removed by the other pipelines. P1 also retained false-positive coordinates (e.g., sea, country capitals and centroids, biodiversity institutions, herbal shops), geographic outliers, and duplicates, which were removed to different degrees by the pipelines of the other group (Table 2.2). (Question 1).

Due to the low complexity of the data cleaning environment, P1 and P2 required only little effort to get their pipelines installed. Both pipelines did not achieve the standardization and error elimination anticipated to reduce unwanted effects in the downstream analyses. P1 identified potential shortcomings in the data only in a few cases due to the limited options of the GBIF filter application. In contrast, P3 to P6 were more demanding in the required know-how, mainly when using the *R* packages and preparing the respective user environments but offered a more substantial functionality (Table 2.2). The *R* packages performed the data cleaning well for coordinate errors that rendered records unusable for use in diversity models. Generalist packages like the *dplyr* and specialists like the *CoordinateCleaner*, especially in combination, reliably identified problematic records with missing values and false-positive occurrences such as biodiversity institutes or country centroids. Accurate distribution data are essential for any SDM and the many comparable downstream analyses (Chapman et al., 2000, Kadmon et al., 2004, Araújo and Guisan, 2006, Zizka et al., 2020). Therefore, the main aim of well-designed pipelines is to efficiently and automatedly generate cleaned data tailored to the specific research question (Zizka et al., 2020; Table 2.1). We mainly focused on comparing the outcomes of different pipelines that used well-known data retrieval or DC tools to answer this question. The standardization filters served to unify the record structure across the pipelines. Although older herbarium vouchers or observations are as valuable as recent vouchers since

they may document both a historical status and biodiversity changes over time (Meyer et al., 2016), the "collection year, older than 1945" filter, for example, was implemented to standardize the data but also to reduce expected general coordinate imprecisions up-front. However, removing taxonomic and spatial errors was at the core of the pipeline data for the model-fitting and -building and the respective tools.

Influence of different data cleaning solutions on downstream analyses

Removing the non-native species, which consisted of only a few specimens, reduced the number of cleaned records only slightly (per species and overall). The non-native *Ephedra* species had no noticeable effect in the occupied grid cells as co-occurring species. They were concentrated in a few places and in small numbers of species only (P1, Figure 2.3, Figure 2.4: observed distribution). The low level of differences was confirmed by reasonably high correlation coefficients, which continued to even higher correlation coefficients regarding the predicted probability of species in S-SDMs (L1 to L5: Figure 2.3). Removing the missing value records in the pipelines was essential for the downstream analyses. The model fitting tool issued error messages when identifying any in the provided data (*ade4*). Although we included the duplicate records filter in determining the number of duplicate records in the data, duplicate records did not affect the fitted models. (Question 2).

The tested pipelines offer automated data cleaning in a standardized and reproducible manner. Pipeline P1 supports all users but produces data that still contains serious taxonomic and spatial errors. In contrast, the pipelines P2 to P6, which help users with some programming experience (Zizka et al., 2019, Zizka et al., 2020), produce data qualities where many errors were eliminated and which seem suitable for diversity model use (SDMs and S-SDMs).

Significant differences of the expert data and the GBIF data

The P1 data differed noticeably from the expert data, e.g., in the species composition (P1 data: 29 species versus expert data, and P2 to P6 data: 12 species), the number of records per species, the number of occupied grid cells after the observations were allocated to gridded range maps (Figure 2.3, L2), and the number of co-occurring species. P2 to P6 differed less from the expert data. (Question 3). The aim of collating data for SDMs is to avoid bias and inaccuracies in taxonomic and distribution data, and an effective means of overcoming bias and inaccuracies is to build data from field studies (Chapman, 2005, Araujo et al., 2019). Well-maintained expert data support both the aims and provide an alternative to field studies. A less-maintained data alternative, biodiversity records from GBIF, are free of charge but with limitations in data

quality due to several known and unknown errors. Expert and GBIF data form the data layer (Vetter, 1990, Bakshi, 2012). However, the critical difference between expert data and GBIF data is that the expert data may be used unprocessed as input to the data modeling workflow as there are no data errors to be expected. For the GBIF data, an additional data cleaning process chain needs to be included in the workflow so that the data modeling can be meaningfully linked to the data layer. Consequently, a user of GBIF data always has to plan for an additional effort for the data cleaning design, which includes the functional structure of the target data that is fit for use, and a pipeline to obtain it (Wirth & Hipp, 2000, Zizka et al., 2019).

A major issue: Misidentified specimens that still hide in the dataset

Comparing the quantities of the GBIF pipelines' analysis data and the expert data shows that the expert data is roughly 11.8 % or about 1/8th of the GBIF data (mean). From this ratio, we may assume that there are still many errors in the pipeline data, hence, the visible differences in the maps (Figure 2.4). This point opens the question of how realistic the GBIF data is. No pipeline detected taxonomic issues such as misidentifications or false positives like non-native specimens in the data due to a lack of information about their distributional status. For differently determined specimens of the same origin, given to other institutes and handled in isolation from their parent specimens, Nicolson (2019) provided a technical solution. We used expert know-how to assess the likeliness of taxonomic identities in recorded localities as there presently is no tool that possesses this functionality (Appendix, Figure A3). Developing a tool that resolves this issue might be challenging considering the many names, from synonyms to misspellings (Zermoglio et al., 2016). A correction method that was already introduced is that a data owner directly changes false positives identified in individual cases by notifying the provider. Generally, with the present interfaces to GBIF, it cannot be avoided that misidentified taxa enter into the databases by, e.g., citizen scientists. Interfaces that prevent taxonomic or spatial errors before entering a public provider must be designed.

Conclusion

Our results suggest that the P1 data shows more differences from P2-P6 data than within this group. Depending on the pipeline, one-third (P1) to two-thirds (P6) of the GBIF records were classified as unsuitable for biodiversity analyses. Importantly, differences in the pipeline data did not translate into significant differences in downstream SDMs and S-SDMs, suggesting remarkable robustness of these analyses towards data cleaning differences. The increasingly condensed information from the occurrence data led to ever-stronger correlations across the

pipelines. Three aspects emerged from the study. First, data from the GBIF web application requires further cleaning. Second, the *R* packages reliably removed incorrect or dubious coordinates. Therefore, choosing the right DC tools depends on the researcher's skills. Third, it is challenging to detect misidentified specimens in the public data providers. To overcome this difficulty, we suggest new processes to detect misidentified specimens or prevent new misidentified specimens from being entered into the public data providers. Consequently, programmers developing new data cleaning packages should consider the functionalities required for data cleaning, notably as the *CoordinateCleaner* eliminates most spatial errors.

Acknowledgement

We thank Pedro Tarroso and an anonymous reviewer for their helpful suggestions and comments on the earlier versions of the chapter. We also acknowledge statistical advice of Patrick Weigelt and fruitful discussion with the members of the Biodiversity, Macroecology and Biogeography group.

Fuehrding-Potschkat, P., Weigelt, P. Kreft, H., Ickert-Bond, S.M.: Drivers of Variation in Synonym Numbers of Angiosperm Species Names. Unpublished manuscript.

3

Drivers of variation in synonym numbers of angiosperm species names

Abstract

Synonyms are part of the scientific progression in taxonomy and nomenclature and reflect the evolving knowledge about species based on revisionary systematics. However, synonyms frequently cause problems in biodiversity repositories, so understanding the causes of the variation of botanical synonyms is essential. Recent studies attribute variation in synonyms to intrinsic and extrinsic drivers, such as nomenclature, taxonomic group membership (e.g., of orchids), and the age of the accepted name. Here, we examine the drivers of the synonyms for a large global subset of all angiosperms. Across 137,378 accepted names of 193 angiosperm families and 5,019 genera present in 355 botanical countries and regions worldwide, range size, the age of the accepted name, and insularity (insular or mainland occurrence, or occurrence on both) emerged as drivers with a positive effect on angiosperm synonyms. After accounting for these three factors, the residual differences in the number of botanical continents and the interaction between insularity and the range size became less significant. The combined multi-predictor model explained about 41% of the global variation in angiosperm synonymy (96%, including the random effects of the families, genera, and the presence patterns of accepted species on one or more botanical continents). We suggest that geographic distance between taxonomists enables wide-ranging species and species with insular distributions to accumulate more synonyms. Also, the age of an accepted name plays a vital role in synonym accumulation. Our results can help to set priorities in revising floras and checklists and to resolve synonymy problems in biodiversity databases, likely leading to more realistic global species numbers. As the drivers may also impact other plant taxa, the study likely has implications for a wider range of families and genera.

Introduction

Taxonomy aims at identifying, characterizing, and classifying living organisms and thereby sets the foundation for hypothesis-driven research in ecology, biogeography, and conservation biology (e.g., Isaac et al., 2004, Wilson, 2004, Thomson et al., 2018). Taxonomists use morphological, genetic, behavioral, and biochemical characters to identify and describe taxa following specific nomenclatural principles and rules. The principle of priority (Turland et al., 2018), essential to naming organisms, states that the accepted name is the earliest validly published name for a given species; younger names are considered synonyms if more than one name describes the same species. Synonyms may emerge for different reasons, for instance, from different taxonomists interpreting and classifying interspecific variation differently; the two resulting philosophies are referred to as 'splitting' and 'lumping'. If a splitter and a lumper classify species of the same genus, the former will usually recognize more species than the latter.

It was suggested that if multiple names exist for the same species, these were not solely caused by altered taxonomic relationships, e.g., that the natural variation of a species was unknown, and various forms of the same species were given different names (e.g., Mori, 2013). For example, it was speculated that taxonomists might show preferences toward attractive taxa and that this would increase synonym numbers (Pillon & Chase, 2006, Lughadha et al., 2016). An uneven distribution of synonymy among families and high concentrations in a few large families like Asteraceae, Orchidaceae, and Poaceae was detected by Lughadha et al. (2016). Other studies explored cases of taxonomists describing unknowingly and independently the same species more than once (e.g., Valdecasas et al., 2008, Joppa et al., 2011, Ickert-Bond et al., 2019). For example, due to the geographic distance between taxonomists, wide-ranging species may accumulate more synonyms (e.g., Baselga et al., 2010, Mori, 2013, Fenneman, 2017). This assumption might also be realistic for continents separated by large bodies of water that expand the range, like the Americas and Africa. Also, species with island distributions might accumulate more synonyms because of a higher number of endemic species (Kier et al., 2009) and complex distributional ranges or because of a researcher's assumption that a species discovered on an island is endemic. Other studies proposed that the time passed since the original description of the accepted name plays a crucial role in the accumulation of synonyms (Alroy, 2002, Baselga et al., 2010, Joppa et al., 2011). Finally, some taxonomists noted that other taxonomists were creating species' names as if to 'retain a place in posterity' through authorship of taxa (Bruun, 1950, Pillon & Chase, 2006, Dubois, 2008, Evenhuis, 2008).

Synonyms are an integral part of the natural progression of taxonomy and nomenclature and reflect the ever-changing knowledge about species (Valdecasas et al., 2008, Mori, 2013). Revealing synonyms helps to deepen our understanding of organisms by better understanding otherwise hidden properties of organisms (Holman, 1987). Recent studies, however, showed that the degree of synonymy is quite substantial for some taxa. In some insect groups, the observed ratio of synonyms to accepted names plus synonyms (synonymy rate) exceeds 50% (Gaston & Mound, 1993, Wells et al., 2019). Similarly, it was estimated that around 66% of all published seed plant names are synonyms (Wortley & Scotland, 2004).

Taxonomic uncertainties resulting from the inconsistent treatment of species' delineation and synonymy represent a major challenge for integrating biodiversity data in public data repositories and may lead to erroneous results (Alroy, 2002, Gotelli, 2004, Dubois, 2008, Jansen & Dengler, 2010). For instance, unresolved synonyms artificially increase the number of names in biodiversity repositories. Synonyms also confuse taxonomy when, for example, it is difficult to recognize whether a species' name in a repository is simply an alias of a more common species (Gaston & Mound, 1993). The same applies to a synonym that cannot correctly relate to an accepted parent name. When taxonomic sources do not consistently identify a scientific name as a synonym, the likelihood for misinterpretation in checklists and other floristic and faunistic treatments increases (Gotelli, 2004, Jansen & Dengler, 2010, Meyer et al., 2016). As a result, thousands of floras and checklists used worldwide are rarely congruent in their taxonomy (Dubois, 2008, Jansen & Dengler, 2010).

Here, we analyzed the variation of synonym numbers in angiosperm names worldwide and tested five competing but not mutually exclusive hypotheses contributing to synonymy (Table 3.1). We examined the variation in synonym numbers across families and genera. Furthermore, we explored the variation in synonymy across botanical continents where the species were distributed, species' insularity (defined as a species occurrence on islands, the mainland, or both), and the species' range sizes. Finally, we tested the age of the accepted name as a proxy for the time passed since the description of an accepted name. Our results can be used to identify plant taxa that may have an increased probability of unidentified and unresolved synonyms, and to set priorities in revising checklists, floras, or biodiversity databases. The identified name discrepancies can also be further tracked for negative effects across related floras, checklists, and repositories. The outcome of this study likely has implications for a broader range of plant families and genera beyond those examined in the current study.

Material and methods

Data cleaning and preparation of the analysis file

On February 13, 2020, we retrieved 537,000 seed plant name records of 270 families from the World Checklist of Selected Plant Families (hereafter: WCSP; WCSP, 2020), including species' accepted names, synonyms, and publication information. We removed 27,538 non-angiosperm names (Stevens, 2016), 12,927 erroneous, and 7,974 unplaced names (both categories were already flagged by the WCSP). 161,392 accepted names and 340,271 synonyms from 193 angiosperm families remained. An additional 18,262 lower-level names (e.g., subspecies) with 27,453 synonyms were removed, leaving 143,130 accepted names and 312,791 synonyms for a total of 455,921 angiosperm names. We also removed 24,542 synonyms containing nonsensical publication year values (e.g., 0 (zero), 3- or 5-digit years, multiple years, and comments) that could not be matched to an accepted name. For 475 accepted names with nonsensical years (1,577 assigned synonyms), it was impossible to derive the age of the accepted names even from the oldest synonym. We used parent-dependant relationship information to link the remaining 279,694 synonyms (dependants) to their respective accepted parent names (142,655 records) and counted them (synonym number: *synNum*; hereafter, variable names in italics). The *synNum* served as the response variable during hypothesis testing. The publication year was unavailable for 3,694 accepted names (1.1%). In this case, we used the oldest synonym. The oldest publications date to 1753 (Linnaeus, 1753) and end in 2019 (spanning a total of 267 years). Therefore, we calculated the *age of an accepted name* by subtracting the publication year from 2020. For further analyses, we used the *full accepted name*, *family* (predictor variable of hypothesis H1a, Table 3.1), *genus* (H1b), *age of an accepted name* (H5), and the *synNum*.

In addition, we used occurrence information for 143,130 accepted seed plants at the species level (in one or more of the 378 TDWG countries and regions, hereafter: TDWG entity, level 3, indicated by 1, presence, and 0, absence; Brummitt, 2001, WCSP, 2020a). From this, second, WCSP file (hereafter: occurrence file) and a spatial polygon (TDWG, 2021), we prepared the predictor variables *botanical continent, where a species is present* (hereafter: "*BC*"; H2a) and *number of botanical continents on which a species occurs* (hereafter: *BCNum*, H2b). In addition, we established the predictor variable *insularity* of a species and computed the *range size* by summing the areas of the respective TDWG units.

Table 3.1. **Hypotheses summary: Drivers of synonym numbers, affecting the variation in synonym numbers and synonymy rates (*synRate*).** Species which are affected by the drivers described in the hypotheses below are expected to accumulate synonyms more frequently.

H1. Synonym number and *synRate* vary among families or genera. Species belonging to particular angiosperm families and genera – regardless of the higher taxonomic level's species number – have an increased probability of being described as different species (Pillon & Chase, 2006, Lughadha et al., 2016). Thus, we expected species of these families and genera to accumulate synonyms more frequently. Explanatory variables: *family* and *genus* (both categorical).
H2a. Synonym number and *synRate* vary across the botanical continents / continent combinations. Species present in specific continents and continent combinations have an increased probability of being described as different species. We expected species being present in particular botanical continents and continent combinations to accumulate synonyms more frequently. Explanatory variable: *occurrence on TDWG botanical continent(s)* (except Antarctica) (eight-digit variable, binary: Y = 1 / N = 0). **H2b. Synonym number and *SynRate* vary with the number of botanical continents where species are present**. Species present on more than one botanical continent have an increased probability of being described as different species. We, thus, expected species present on many botanical continents to accumulate synonyms more frequently than species occurring on only one or few continents (extension of H4, below). Explanatory variable: *number of botanical continents a species occurs* (numerical: 1 to 8).
H3. Synonym number and *synRate* vary with the insularity of a species. Species present on islands have an increased probability being described as different species than species occurring on the mainland. We expected species present on islands to accumulate synonyms more frequently than species occurring on the mainland only. Explanatory variable: *insularity* of a species, on islands only, on the mainland only, and both on islands & the mainland (using the TDWG classification of the respective botanical country as island or mainland).
H4. Synonym number and *synRate* vary among species range sizes. Wide-ranging species are more likely to be described as different species (proxy for the geographic distance between taxonomists) than species with small ranges. We expected species with large ranges to accumulate synonyms more frequently than species with small ranges (Baselga et al., 2010, Fenneman, 2017). Explanatory variable: *range size*, computed as the sum of TDWG countries where a species occurs (Source: TDWG shapefile data frame).
H5. Synonym number and *synRate* vary with the age of a species' accepted name. Species' accepted names validly published a long time ago had more time to accumulate synonyms than recently published accepted names (Alroy, 2002, Baselga et al., 2010). We expected early published names to accumulate synonyms more frequently than recently published names. Explanatory variable: *age of a species' accepted name* or – if not available or younger than the first published synonym – its oldest synonym.

We considered eight botanical continents (by TDWG Level 1 code, Brummitt, 2001), excluding Antarctica, to avoid the bias of a large continent with very few species (WCSP, 2020a). If a species was reported in a TDWG unit (given in the occurrence file for each accepted name), we marked the corresponding botanical continent ('1', presence and '0', absence). Furthermore, we summed up the *BCNum*. We concatenated the species' occurrences on the eight botanical continents into an eight-digit *BC* string (presence-absence patterns). The TDWG continent number was the position number in the string, determined from left to right. Examples for presence-absence patterns were, e.g., for the presence in South America only: '00000001', and presence in Europe, Africa, and South America: '11000001'. We determined the *insularity* of a species by their respective TDWG classification (Brummitt, 2001). For example, Australia and its continental subunits (e.g., Western Australia, Queensland) were classified as mainland, and Tasmania as an island. Depending on the determined species insularity type, we set *insularity* to 'I' (islands), 'M' (mainland), or 'A' (island and mainland) (factor with three levels). We regarded a species' *range size* as a proxy for the physical distance between taxonomists. As an estimate for *range size*, we computed the sum of all country areas where a species was reported.

We merged the continent-related explanatory variables (the presence-absence string *BC* and *BCNum*), *insularity*, and the total *range size* to the initial part of the analysis file, achieving a final set of nine variants of five putative drivers of synonym numbers in angiosperms. Resulting from the merge, we identified 2,058 accepted names with 6,592 synonyms that were not associated with a TDWG unit or Antarctica. We also identified 3,219 records of accepted species with 11,278 synonyms that had not all predictor variables filled with values and therefore had to be removed. The data cleaning process resulted in 137,378 accepted angiosperm names with 261,824 synonyms.

Statistical modelling

Collinearity among predictor variables was tested using the *R* package *rstatix* (Kassambara, 2020) and visualized using the *GGally* package (Schloerke et al., 2018; Appendix, Figures A1(a) to (f)). We examined skewness and kurtosis of the data using the package *moments* (Komsta et al., 2015, Appendix, Table B2), and nested, multilevel structures with *lmerTest* (Kuznetsova et al., 2017). Structural details of the data were visualized using *ggplot2* (Wickham, 2016, e.g., Appendix, Table B2) and the *ggpubr* function *ggdensity* (Kassambara, 2020a, Appendix, Table B2(a), Density diagram).

We used generalized linear mixed effects models (GLMM) to examine the drivers of synonym numbers. We analyzed the linear relationships of *synNum*, including interactions of explanatory variables and assessed variable performances using *R* packages *ROCR* (Sing et al., 2015) and *performance* (Lüdecke et al., 2021). We natural log-transformed *range size* to approximate its observed distribution to a normal distribution. The explanatory variables were standardized (z-transformation, using the *rescale* function) to improve the linearity and comparability of coefficient estimates. We analyzed the suitable error distribution for the count data and the appropriate link functions (Garson, 2013). Frequent issues to be handled in count data are zero-inflation (e.g., Hartig, 2019) and overdispersion (causing incorrect standard errors, e.g., Bell & Grunwald, 2011, Meyer, 2021). In terms of error distribution, Poisson, Poisson/zero inflation, and negative binomial, Poisson/zero inflation, the employed logit link function provided the best-fitting models (Tlhaloganyang & Sakia, 2020, Appendix, Table B2).

The variables *range size* and *age of an accepted name*, *insularity*, and *BCNum* were used as fixed factors in the GLMM model. The other variables, *family* (193 levels), *genus* (5,019 levels), and *BC* (217 levels) were used as random factors (McGill, 2015, Appendix, Figure B3). All variables showed significant effects (SE < 0.013, *p*-values < 0.001) in the GLMM analyses, suggesting they were predictive (Bell et al., 2019). In addition to the single predictor variables, we tested how the interaction of species occurring on islands, the mainland, or both related to their range size (hypotheses H3 and H4) influences the accumulation of synonyms (Hox et al., 2017, partial correlation analysis: R-package *ppcor*, Kim 2015; $p = 0$).

We fitted multi-level regression models using the *R* package *glmmTMB*, which minimized overdispersion and zero inflation (Bolker, 2016; see: Table B2). We used three distinct goodness-of-fit measures for the model selection: the Akaike information criterion (AIC; Burnham & Anderson, 2004), the root-mean-square error (RMSE), and the marginal and conditional pseudo-R^2 (Nagakawa & Schielzeth, 2013, Johnson, 2014, Schielzeth et al., 2020). We computed models of the individual predictors in all possible combinations (Stoffel et al., 2021; predictors: four fixed factors, one interaction, and three random factors). The possible combinations were determined by the mandatory specifications of the used algorithm. At least one random factor was compulsory for *glmmTMB*. The computations delivered the R^2 proportions of the fixed and random factors for the models (as the conditional and marginal R^2s). We decomposed the R^2s per explanatory variable as described in the computation procedure of the R packages *PartR2* (Stoffel et al., 2021) and *rptR* (Stoffel et al., 2017). We selected four models, all with an almost identical AIC at a stable minimum and a maximized

pseudo-R^2 (model selection criteria; Myung, 2000. Appendix, Table B4(a) and (b)). We evaluated the model performances with the packages *jtools* (Long 2017), *sjPlot* (Lüdecke, 2021), and residual information. We also analyzed the models with the *DHARMa* diagnostics package (Hartig, 2020). We performed a Kolmogorov-Smirnov (KS) test (normal distribution of the residuals), an overdispersion, and an outlier test. The *p*-values were calculated for each model (Appendix, Figure B5, Appendix, Table B4(a) based on 500 replications.

While we counted *synNum* per accepted species, we computed a *synonymy rate* (*synRate*) from the *sum of accepted species* and their collective *synNum*s (both from the counted, hereafter: observed, and predicted by the GLMM model) of a given group (*synRate* = (*synNum*/(*sum of accepted names* + *synNum*)) * 100 [%]; Lughadha et al., 2016). The predicted values were reverted from the natural log using the R *exp* function. The *synRate*s allowed a species richness-independent ranking of each categorical predictor level based on the observed or predicted *synNum* (predicted: from the model) they accumulated or computed for their accepted names (*family*, *genus* and *BC*). The predicted *synNum* were higher than the observed *synNum*. For example, we extracted the observed and predicted synonyms per botanical continent using the variable *BC* as a presence indicator. We summed the synonym numbers per botanical continent according and analyzed the variation between observed and predicted synonym numbers (a) per botanical continent and (b) across botanical continents (Figure 3.1).

For the data retrievals, manipulations, analyses, and modeling in this study, we employed *R* Studio and *R* versions 3.0.2-3.2.1 (*R* core team, 2013).

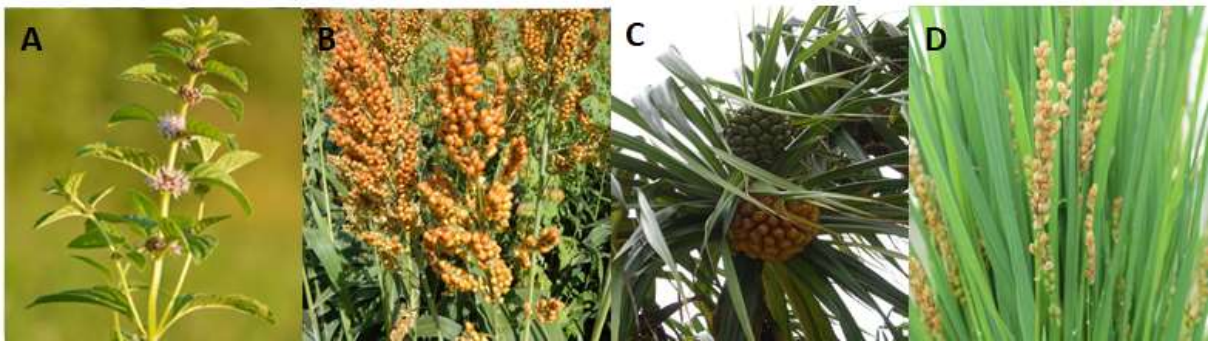
Results

Data basis for model fitting

The data cleaning exhibited out-of-scope species records, i.e., non-angiosperms (27,538 records, including 12,928 erroneous records and 176 unplaced records, respectively marked by the WCSP; 5.1%, based on the initial 537,000 WCSP seed plant records), unplaced angiosperm names (7,799 records, 1.5%), subspecific angiosperm names including their synonyms (45,715 records, 8.5%), and species, occurring on the continent of Antarctica (1,608 records, 0.3%). Among the species of interest, we also found records lacking correct values for essential variables. This category contained 2,052 accepted names and synonyms where no oldest name was available (leaving the *publishing year* and, subsequently, the *age of an accepted name* variable empty; 0.4%). This category comprised erroneous synonym records containing nonsensical data in variables essential for our study (24,542 records, 4.6%, e.g., zeroes, text in

the *publication year*) that could not be matched to an accepted name. This category also included 1,661 accepted names with 5,381 synonyms that were not assigned to a BC (1.3%) and 3,219 accepted names with 17,863 synonyms that were missing proper values in one or more of the relevant predictor variables (3.9%). During data cleaning, we removed a total of 137,798 records (25.7%, summed from the individual percentages).

Table 3.2. **Summary of the fifteen accepted species names with the highest synonym numbers among the angiosperms studied.** Images A to D show the four species with the highest synonym number per species name. (Images: A, *Mentha arvensis*; B, *Sorghum bicolor*; C, *Pandanus tectorius*; D, *Oryza sativa*). Column pubYear (Publication year: For each scientific name marked with *, the publication year was determined using the oldest synonym in the absence of the publication year of the accepted name. (Image credit: A: Ivar Leidus, B: Forest & Kim Starr. C: Judgefloro. D: C.T. Johansson. Creative commons licences: A, B, and D: CC BY-SA 3.0; C: CC BY-SA 4.0.)



Family	Scientific name	synNum	pubYear	BotCont	Human use
Lamiaceae	*Mentha arvensis* L.	377	1753	1 to 5,7,8	medicinal, spice
Poaceae	* *Sorghum bicolor* (L.) Moench	344	1753	1 to 8	staple food (crop)
Pandanaceae	*Pandanus tectorius* Parkinson ex Du Roi	321	1774	2 to 8	food, building
Poaceae	*Oryza sativa* L.	320	1753	1 to 8	staple food (crop)
Lamiaceae	*Mentha aquatica* L.	302	1753	1 to 3,7,8	medicinal, spice
Asparagaceae	* *Cordyline fruticosa* (L.) A.Chev.	233	1754	2 to 8	ornamental gardening
Poaceae	*Festuca rubra* L.	222	1753	1 to 8	ornamental gardening
Euphorbiaceae	*Ricinus communis* L.	212	1753	1 to 8	medicinal
Poaceae	*Agrostis stolonifera* L.	209	1753	1 to 8	ornamental gardening
Rubiaceae	*Kadua affinis* Cham. & Schltdl.	200	1829	6	ornamental gardening
Oleaceae	*Phillyrea latifolia* L.	187	1753	1 to 3	ornamental gardening
Campanulaceae	*Campanula rotundifolia* L.	179	1753	1,3,5,7,8	ornamental gardening
Myrtaceae	* Myrcia splendens (Sw.) DC.	170	1788	7,8	medicinal, fruits, timber
Poaceae	*Festuca ovina* L.	168	1753	1 to 4,7,8	-
Cannaceae	*Canna indica* L.	166	1753	1 to 8	ornamental gardening

BotCont, botanical continent: 1 = Europe, 2 = Africa, 3 = AsiaTemperate, 4 = AsiaTropical, 5 = Australasia, 6 = Pacific , 7 = Northern America, 8 = Southern America.

We ultimately obtained 137,378 accepted names (25.6%) with a total of 261,824 synonyms (48.8%) present in 355 TDWG units. The *synNum* varied strongly between zero and 377, mean *synNum* was 1.904, median *synNum* was 1, and the distribution was strongly right-skewed (skewness coefficient: 17.58) with a steep kurtosis (685.65). Natural log-transforming the *synNum* led to a slightly right-skewed, approximated normal distribution (skewness coefficient: 1.33, kurtosis: 4.68, Appendix, Table B2(a)). 68,979 of 137,378 accepted angiosperm names (50.2%) had no synonyms, while five names accumulated more than three hundred synonyms each since 1753. The five accepted species with the highest synonym numbers were *Mentha arvensis* L. (377 synonyms), *Sorghum bicolor* (L.) Moench (344 synonyms), *Pandanus tectorius* Parkinson *ex* Du Roi (321 synonyms), *Oryza sativa* L. (320 synonyms), and *Mentha aquatica* L. (302 synonyms). Table 3.2 lists the top-fifteen accepted names with the highest synonym numbers among angiosperms available in the WCSP.

The synonym numbers differed significantly across families and genera. Synonym numbers per family varied from no synonyms (in sixteen out of 193 families) to more than 40,000 in the Poaceae (47,443 synonyms) and Orchidaceae (43,839 synonyms). Cannaceae exhibited the highest *synRate* (95.2%) of all families for the twelve accepted names and 238 synonyms (*synNum* [mean]: 19.83). The relatively small family Potamogetonaceae took second place (88.2%) with 106 accepted names and 790 synonyms (*synNum* [mean]: 7.19). Large families such as the Poaceae ranked 12th with a *synRate* of 80.4% (*synNum* [mean]: 4.11). The Orchidaceae ranked 99th with a *synRate* of 60.3% (*synNum* [mean]: 1.5). (Details: Table B6(a)). At the generic level, *synNum* varied by four orders of magnitude, ranging from zero synonyms (in a total of 578 genera out of 5,019) up to more than 4,000 (*Carex* L., *synNum* [mean]: 2.42) and 3,000 (*Dendrobium* Sw., *synNum* [mean]: 1.95, *Euphorbia* L., *synNum* [mean]: 2.46, and *Cyperus* L., *synNum* [mean]: 3.54). The highest *synRates* were found for *Ricinus* L. (*R. communis* L., one accepted name and 212 synonyms) and *Phillyrea* L. (two accepted names and 247 synonyms). Both had a *synRate* of more than 99% (Details: Table B6(b)).

The *synRates* also varied among the botanical continents (Figure 3.1, Table 3.3). Europe, Pacific, and North America emerged as the continents with the highest *synRates* from observed synonym numbers (90.7%, *synNum* [mean]: 9.79; 85.6%, *synNum* [mean]: 5.96; 84.2%, *synNum* [mean]: 5.31).

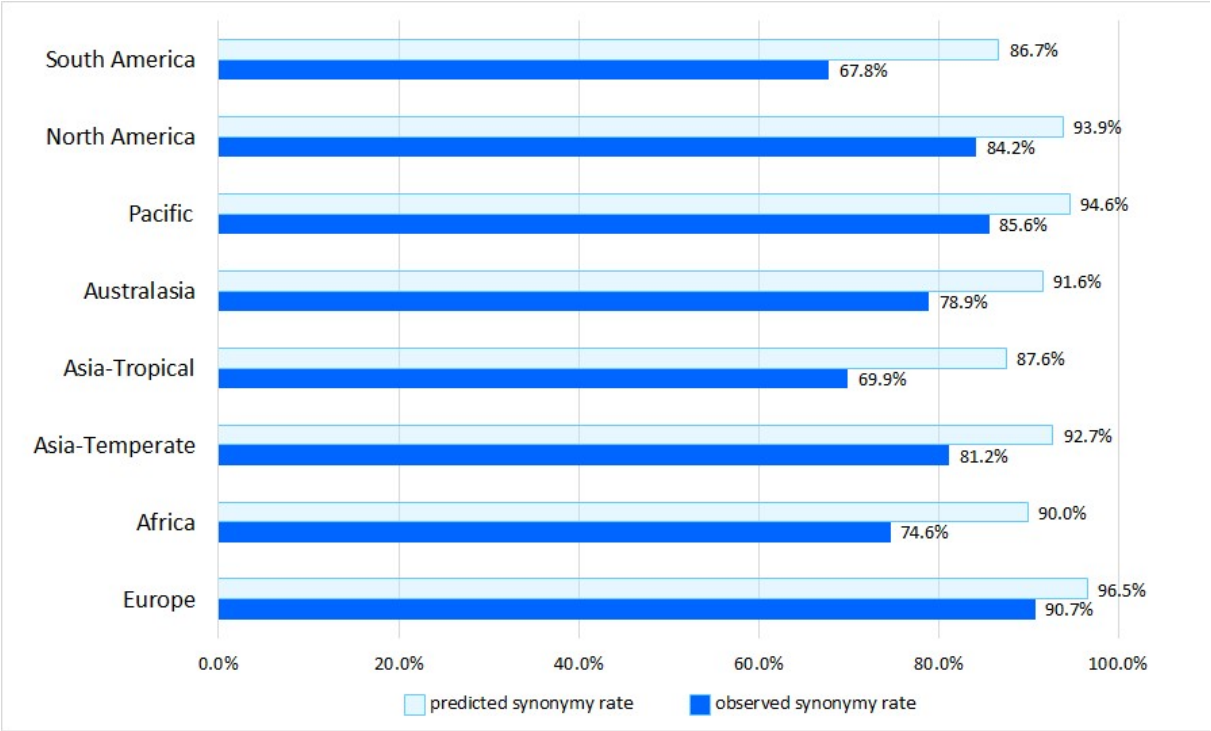


Figure 3.1. Variation of synonym numbers across botanical continents (random factor), by synonymy rates (observed: blue, predicted: light blue). We extracted the observed and predicted synonym numbers per botanical continent using the *BC* (eight-digit presence-absence pattern) as a presence indicator. We summed the synonym numbers per botanical continent and analyzed the variation between observed and predicted (a) per botanical continent and (b) across botanical continents. Observed synonym numbers were counted when linking synonyms to their parent species name. Thus, the observed *synonymy rates* (*synRate*) are derived from this number for the *synRate* formula. The predicted synonym number is derived from the glmmTMB model. Thus, the predicted *synRate* is used in the formula. Both the observed and the predicted *synRates* were computed as: *synRate* = (*synNum*/(accepted names + *synNum*)) * 100 [%] (Lughadha et al. 2016). The *synRate*s allowed for a relative level ranking independent of each continent's absolute synonym number. Europe (90.7%), Pacific (85.6%), and North America (84.2%) emerged as the continents with the highest *synRates* from observed synonym numbers. The predicted *synRates* were even higher. For Europe, a *synRate* of 96.5% was computed, followed by Pacific with 94.6%, and North America with 93.9%. The observed *synRate* of South America increased from 67.8% to a predicted *synRate* of 86.7%, similar to Asia-Tropical, where the observed *synRate* increased from 69.9% to 87.6% (predicted).

Drivers of synonym numbers

Collinearity among the explanatory variables was generally low and highest between the numerical variables *range size* and *age of the accepted name* (absolute Pearson correlations of 0.35). Repeated testing of different error distributions and the *DHARMa* diagnostics showed that the best model performances were obtained using the *glmmTMB* package, the Poisson/zero-inflation error distribution, and the logit-link function (Appendix, Table B2). Iterative GLMM analyses resulted in four fitted models of very similar model parameters, performing nearly

equally well. As a result, the AIC values and the conditional and marginal R^2s of the four models were also similar (Appendix, Table B4(a)), ranging from 4.880E+05 to 4.882E+05. The RSME of 4.005 revealed a high predictive accuracy with a quantified average error of 4%. The conditional R^2s ranged from 0.958 to 0.964, and the share of the fixed factors (marginal R^2) ranged from 0.396 to 0.414. Only Model 4 met the equidispersion requirement (conditional R^2 of 0.989, fixed factors: 0.414). According to *DHARMa* diagnostics (Appendix, Figure B5), the models did not show significant zero inflation (i.e., given the fitted model, the expected and modeled zeroes were in the same range, Hartig, 2019). Thus, we selected Model 4 as the final model to explain the combined drivers of synonym numbers.

The combined multi-predictor GLMM model 4 explained about 41% of the global variation in angiosperm synonym numbers (96% including the random effects; Table 3.3). The model included the range size (explaining 21.0% of the variation in synonym numbers), the age of an accepted name (11.6%), and insularity (5.6%) as main predictors (Table 3.3). We observed root-mean standard errors (RMSE) between 4 to 5% suggesting that the variables were highly predictive.

Table 3.3. **Global model of angiosperm synonymy.** Selection conditions of the model were: (1) AIC at a stable minimum, (2) maximized pseudo-R^2, (3) *DHARMa* performance tests successful. Result of GLMM of a combined nine-predictor model, by random factors and fixed factors. H1 to H5: Hypotheses (see: Table 3.2). ***, $p < 0.001$. The table below is an extract of Table B4(b) (Appendix).

Hypothesis	Combined model	R^2 share	$RMSE_{mean}$	z	Variation
	Random Factor R^2 share	**0.544**	**-**		**54.4%**
H2(a)	Botanical continents (Presence on particular continents)	0.302	4.857	-	30.2%
H1	Genus of species	0.223	4.404	-	22.3%
H1	Family of species	0.019	4.930	-	1.9%
	Fixed Factor R^2 (Marg.)	**0.414**	**-**		**41.4%**
H4	Range size	0.201	4.621	69.7 ***	21.0%
H5	Age of accepted name	0.111	4.698	139.9 ***	11.6%
H3	Insularity	0.054	4.800	-31.9 ***	5.6%
H2(b)	No. of botanical continents (a species is present)	0.029	4.763	3.6 ***	2.9%
H3*H4	Range size * Insularity	0.019	5.078	17.8 ***	1.9%
	Total R^2 (Cond.)	**0.958**	**4.005**		**95.8%**

Range size had a positive effect on the accumulation of synonym numbers. The larger the range size of an accepted species, the more synonyms it accumulated (Rank 1, Figure 3.2B – with insularity, Table 3.3). The age of an accepted species had a positive effect on the species'

accumulated synonym numbers (Figure 3.2C): The more time had passed since the description of a species name, the more synonyms it accumulated. (Rank 2, Figure 3.2C, Table 3.3). The three insularity types showed a positive effect on species' accumulated synonym numbers, albeit to different extents, as displayed in the regression lines with varying points of intersection and slopes (Rank 3, Figure 3.2B). Species found on islands had a significantly lower *synRate* than those found on the mainland or even both islands and the mainland (99 percent confidence interval: $p < 2.2e\text{-}16$). Species observed only on islands showed a *synRate* of 51.0% (*synNum* [mean]: 1.04), and species only present on the mainland showed a *synRate* of 58.2% (*synNum* [mean]: 1.39). Yet, species present in both showed a *synRate* of 89.0% (*synNum* [mean]: 8.09). The working residuals (Hardin & Hilbe, 2007) varied somewhat for the range size and the age of the accepted name, and they varied slightly within the insularity based on the range size (Figures B7a-c). For the age of an accepted name, the working residuals varied only slightly. We also found differences for the *BCnum* and the interaction of the range size and insularity (Table 3.3). The number of botanical continents on which a species is present had a positive effect on accumulated synonym numbers, but showed only weak effects on global synonym numbers (Rank 4, Figure 3.2A, Table 3.3). The interaction of insularity and the range size showed very weak effects (Rank 5, Figure 3.2B, Table 3.3). The botanical continent's *synRates* (predicted synonym numbers from the patterns, split per botanical continent: *BC*) confirmed the ranking from the observed synonym numbers, but were higher. For Europe, a predicted *synRate* of 96.5% was computed, followed by Pacific with 94.6%, and North America with 93.3%. The observed *synRate* of South America increased from 67.8% to a predicted *synRate* of 86.7%, similar to Asia-Tropical, where the observed *synRate* increased from 69.9% to 87.6% (Figure 3.1).

Discussion

In this study, we analysed geographical and taxonomical patterns and drivers of synonymy of 137,378 accepted angiosperm names and 261,824 synonyms from 5,019 genera and 193 families on eight botanical continents. We examined five competing but not mutually exclusive hypotheses of synonym numbers (Hypotheses H1 to H5, Table 3.1). We observed a large variation in synonym numbers in the used global subset of angiosperms ranging from zero (about 50% were accepted names without synonyms) to 377 synonyms. Variation in synonym numbers was associated with all drivers investigated, which positively affected the accumulation of synonym numbers, but range size, the age of an accepted name, and insularity

emerged as the primary drivers. Together, these three drivers explain about 41% of the global variation in angiosperm synonymy.

Drivers of synonym numbers discussed in order based on their relative importance

Among all analyzed factors, the range size (H4) emerged as the driver with the highest predictive power for the accumulation of synonyms among the three primary drivers. This finding supports the hypothesis that widespread angiosperm species collect more synonyms than range-restricted species as the geographical distance between taxonomists is large (Baselga et al., 2010, Fenneman, 2017, Figure 3.2B).

The age of an accepted name (H5) served as the proxy for the time that passed since the publication of an accepted angiosperm name. This variable also positively affected synonym numbers and ranked second in predictive power. The result corroborates the "historical accumulation of names" hypothesis which states that the more time had passed since the description of a species' name, the more synonyms it accumulated (e.g., Baselga et al., 2010, Joppa et al., 2011, Figure 3.2C).

In our analyses, the three insularity types positively affected the accumulated synonym numbers (H3: rank 3). However, we found differences between the types' accumulation extent (Figure 3.2B). Computed from counted *synNum*, the *synRate* of species present on islands and the mainland show 89.0%, compared to the *synRates* of species restricted to islands (51.0%) and the mainland (58.2%). Computed from predicted *synNum*, the *synRate* of species present on islands and the mainland still shows 63.1%, while the *synRates* of islands and mainland species drop to 2.2% and 6.3%, respectively. The results are probably due to extended species' ranges, and the ranges increased complexity.

Synonym numbers were unevenly distributed among the studied families and genera and differed significantly (Hypothesis H1). The rank positions of families and genera (by *synRate*) may hint at particular taxa being more notable than others. For example, some families are morphologically difficult (e.g., Poaceae), others tend to produce hybrids (e.g., Betulaceae). Also, the attractiveness of a taxon may have a decisive impact on taxonomists' motivation in general (Henrich & Gil-White, 2000, Pimm & Joppa, 2015, Jensen, 2019). However, attractiveness is subjective and difficult to quantify. Thus, our results cannot support findings in previous studies which suggested that particular families, like Orchidaceae, accumulate more synonyms due to being more attractive to researchers than others (Pillon & Chase, 2006, Lughadha et al., 2016, Tables B6(a) and (b)).

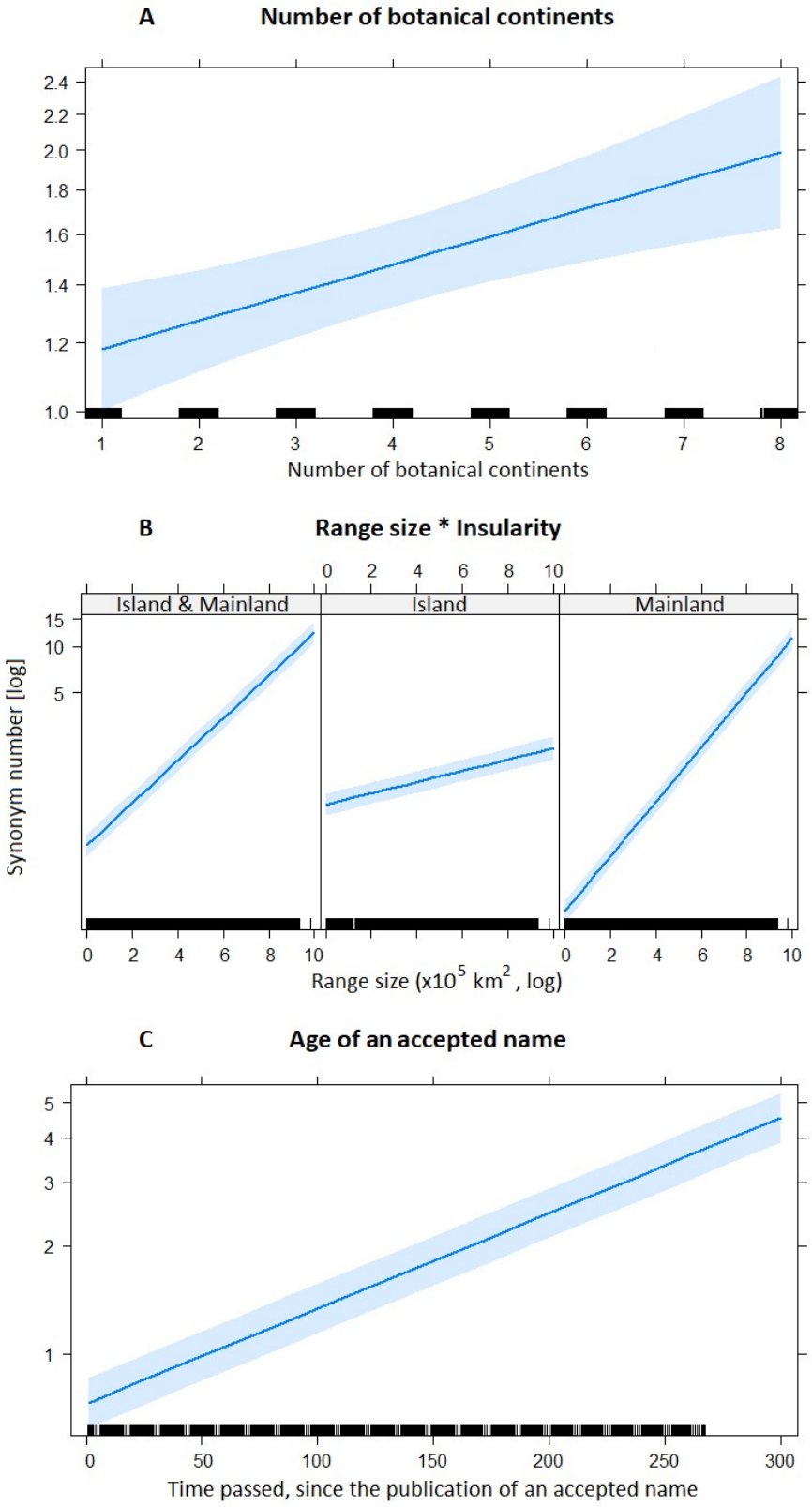


Figure 3.2. **Variation of synonymy rates (from the predicted fixed factor synonym numbers):** A: Number of botanical continents, a species is present, rank 4. B: "Range size" and "Insularity" (H3/H4), individual predictor ranking: "Insularity", rank 5, "Range size", rank 1, interaction: rank 5. The working residuals vary by the species' insularity. C: "Age of an accepted name" (H5), rank 2. The plots were prepared using the *effects* package (Fox et al. 2016).

Yet, the attractiveness of taxonomic study objects in the selection process of researchers (e.g., due to specific pollination mechanisms, ecology, and horticultural value, Heß, 1990, Lughadha et al., 2016) may warrant a study on their consequences on research biases.

The expectation that species in particular botanical continents and continent combinations will accumulate synonyms more frequently was confirmed. However, the botanical continent proved to be a contradicting driver when comparing the predictive power of the continent patterns to the number of continents a species is present. With almost 32%, the botanical continent accounted for a dominant proportion as a random effect (H2a, continent patterns: high conditional pseudo R2 share). The number of continents, a species is present, had a positive effect on the accumulation of synonyms, albeit with very low predictive power, accounting only for 3% of the global variation (H2b, number of continents: very low marginal pseudo R2 as a fixed effect).

Synonym numbers varied systematically by botanical continent (Hypothesis 2a). Europe, Pacific, and North America emerged as the continents with the highest *synRate*s based on observed synonym numbers (from nearly 85% to more than 90%). The *synRate*s from predicted synonym numbers confirmed these results, although predicted synonym numbers were slightly higher (by about 10%) as compared to observed synonym numbers. Contrary to this overall trend, the observed *synRate* of South America and Asia-Tropical, however, increased by 15 to 20% (Figure 3.1). Overall, these results are consistent with the notion that numbers of invalid, infraspecific, and hybrid names are significantly higher in Europe than in surrounding areas, which coincides with the high number of systematists working there (Pillon & Chase, 2006). (Hypothesis 2b). Also, for the Eupelmidae (family of parasitic wasps), it was found that the larger their species range size and the more western a Eupelmid species was located, the earlier a species was described both in Afrotropical and in the Palearctic biogeographical regions (Baselga et al., 2010).

Considering species with high synonym numbers, it is striking that mostly their range is recorded across a higher number of botanical continents. In addition, some of these species are native only to one or a few continents. The botanical continent's predictive power was possibly influenced by such species introduced to new continents. The extension of species' range sizes due to cultivation or invasiveness may have created new opportunities for species to accumulate additional synonyms outside their native range. For example, *Mentha arvensis* with 377 synonyms was a taxon in our analysis that accumulated the highest number of synonyms. The

species is native to Africa and Asia-Tropical, and was introduced to Europe in the 16th century as a pharmaceutical (Roy et al. 2020). *Mentha arvensis* was introduced to at least ten countries (GBIF, 2022), and recorded for seven out of eight continents by the WCSP (WCSP, 2020a). Its European relative *M. aquatica* (302 synonyms) is likewise pharmaceutically significant. It was introduced to at least seven countries (GBIF, 2022), and recorded for five out of eight continents by the WCSP (WCSP, 2020a). Today known as one of the most important crops worldwide (Dial 2012), *Sorghum bicolor* (344 synonyms) is an even more extreme example than *M. arvensis*, having been introduced to 54 countries or islands on eight out of eight continents (WCSP 2020a, GBIF 2022). *S. bicolor* originated in the savannahs of north-eastern Africa (De Wet & Harlan, 1971). Effects from such events were not considered in the models. Taking all findings into account, it may be interesting to investigate the role of the botanical continent on the accumulation of synonyms further.

Conclusion

In our study, we identified range size, the age of an accepted name, and insularity as the main drivers that positively affected the global variation of synonym numbers. Residual differences in the number of botanical continents and the interaction of insularity and the range size became less significant. Our combined multi-predictor model explained about 41% of the global variation in angiosperm synonymy. Four main interpretations emerged from the study. First, the geographic distance between taxonomists caused widespread and insular species to accumulate more synonyms. Also, the time passed since the publication of an accepted species played a dominant role – a trend that is expected by chance. Second, the rank positions of families and genera may hint at particular taxa being more appealing than others. Thus, the attractiveness of taxonomic study objects in the selection process of researchers and the associated research bias may warrant further study. Third, the predictive power of the continent patterns (high) and the number of continents a species is present (low) contradict each other. Also, the artificial extension of species on the botanical continents due to cultivation or invasiveness needs more attention. Therefore, it may be interesting to further explore the botanical continent's role. Fourth and finally, the outcome of this study likely has implications for a wider range of plant families and genera and might also extend to other groups of organisms.

4

Synopsis

Introduction and methods

Comprehensive global biodiversity data are an essential resource to biodiversity studies and, as such, must be clean, consistent, and complete. The required quality is usually obtained in the course of field studies (Chapman, 2005, Araujo et al., 2019). A more convenient way for scientists to access biodiversity data is through aggregated digital specimen records stored with public data providers. However, according to recent analyses, public provider data are generally heterogeneous and limited (e.g., Meyer et al., 2016, Nicolson, 2019, Zizka et al., 2020).

Identified limitations comprise data errors mainly along three dimensions: taxonomy, space, and time (Meyer et al., 2016). Taxonomic errors include, e.g., misidentified species (e.g., Nicolson, 2019, Zizka et al., 2020), unrecognized synonyms or synonyms linked to the incorrect parent species (Dubois, 2008, Jansen & Dengler, 2010), outdated names, and orthographic variations (e.g., Jansen & Dengler, 2010). While the exact number of affected specimens is unknown, estimates range between <1 and 17% (Bisang & Urmi, 1994, Scott & Hallam, 2002, Ahrends et al., 2011) but also extend to more than 50% (incorrectly named tropical specimens: Goodwin et al., 2015; Zoological Record database: Meier & Dikow, 2004). Point-occurrences of misidentified specimens may lead to doubtful species ranges in macroecological diversity models (Nicolson, 2019, Zizka et al., 2020). Unrecognized synonyms, orthographic variations, and misspelled names result in non-existent species or inflated species counts (Linnéan shortfall: The discrepancy between formally described species and the number of species that actually exist; e.g., Lomolino, 2004, Hortal et al., 2015, Ickert-Bond et al., 2019). In the spatial dimension, missing locality data, low precision of coordinates, and false-positives pointing to dubious places were commonly found data quality problems (e.g., Yesson et al., 2007, Otegui et al., 2013, Töpel et al., 2017). More than 23% of their removed records contained invalid coordinates pointing to the sea (Meyer et al., 2016). Taking the analyses of the identified data limitations into account, my thesis aimed to

1., provide the first quantitative analysis of how public provider data cleaned by different DC pipelines (pipeline data) influenced downstream species distribution models (SDM),
2., understand how the downstream SDMs and stacked SDMs (S-SDM) from pipeline data differ from the respective models from expert data that represent the gold standard, (1&2: Chapter 2), and
3., identify drivers affecting the variation in synonym numbers across angiosperm species, and the extent to which the drivers explain the synonymy in the employed angiosperm species (3: Chapter 3)
In Chapter 2, I focused on North American *Ephedra*. For this, I retrieved 46,384 records from the GBIF and processed them in pipelines. The metrics used quantified the ability of the pipelines to standardize and clean the input data (number of steps of the data cleaning process performed, and type and number of errors removed). In the subsequent steps, I compared the results of the pipeline data with the expert data. I analyzed the *Ephedra* species observed and predicted occurrences in North America using SDMs and S-SDMs. Here, the metrics quantified the number of grid cells occupied and the observed co-occurring species in the grid cells (Figure 2.3, Figure 2.4). Pairwise correlations across the various pipelines and expert data were computed as Pearson's *r* and Moran's *I* to identify differences and similarities in the models and maps.

In chapter 3, I focused on 399,202 angiosperm name records (accepted names and synonyms) from the World Checklist of Selected Plant Families (WCSP) to analyze the role of five drivers of synonym numbers (higher taxa of species (families and genera), the botanical continents where the species are present, insularity, range size of a species, and the age of their accepted name; hypotheses: Table 3.1). Using the name records, I tested the global variation in angiosperm synonym numbers related to these drivers of synonym numbers, including the drivers' relative importance. I combined name data, species range data, and independent distribution maps and computed synonym numbers per species and synonymy rates per driver. The combined data served as the input to the model fitting, assessment, and prediction steps.

Results and discussion

Chapter 2. 13,889 non-standardized and uncleaned North American Ephedra records from GBIF included 8,702 taxonomically and spatially erroneous records (63.7%), identified in an independent data analysis step (Table 2.2). Thus, the error proportion in the *Ephedra* data is substantially higher than that reported in previous studies for other taxa. The high error

proportion may be related to the fact that much of the study's time was explicitly dedicated to detecting errors.

The pipeline data confirmed the hypothesis that different DC tools would generate result datasets, which differed significantly in the number of accepted species, errors removed, and remaining records for analysis. For example, P6 removed the most significant number of records, approximately twice as many as the least-cleaning pipeline, P1. P1 retained false-positive coordinates (e.g., sea, country capitals, centroids, biodiversity institutions, herbal shops), geographic outliers, and duplicates that were removed to different degrees by P1–P6.

The resulting models (single species- and stacked SDMs, hereafter: S-SDM) and maps were also expected to differ. But surprisingly, differences within the pipeline data did not translate into significant differences in downstream SDMs and S-SDMs. Correlations of the observed and predicted occurrences differed clearly between the pipelines and the expert data (based on the mean Pearson's r). For example, diversity patterns, although similar in shape, were less distinct in maps from expert data.

Chapter 3. The analysis data resulted in 137,378 accepted angiosperm names with 261,824 synonyms of 193 angiosperm families and 5,019 genera present in 355 TDWG units. The results showed range size, the age of an accepted name, and insularity as the main drivers which positively affected the global variation of synonym numbers.

The combined multi-predictor model explained about 41% of the global variation in angiosperm synonymy (96% including the random effects of the botanical continents, genera, and families). Here, the geographic distance between taxonomists caused widespread and insular species to accumulate more synonyms. Also, the time passed since the publication of an accepted species played a dominant role – a trend that is expected by chance.

Regarding the random effects, synonym numbers were unevenly distributed among the studied families and genera and differed significantly. However, the rank positions of families and genera may hint at particular taxa being more appealing than others. Synonym numbers varied systematically by botanical continent. Europe, Pacific, and North America emerged as the continents with the highest *synRate*s based on observed and predicted synonym numbers. However, the predictive power of the continent patterns (32% as a random factor) and the number of continents a species is present (3% as a fixed factor) contradict each other.

Conclusion and outlook

I aimed to address causes, manifestations, and effects of taxonomic and spatial limitations (mainly, but not exclusively, data errors) in data from public providers. I focused on manifestations of taxonomic and spatial errors (e.g., false-positive species and localities). Spatial errors were straightforwardly detected, particularly with automated evaluation. However, taxonomic errors needed laborious scrutiny and expert support. Remaining errors within the pipeline data did not translate into significant differences in downstream models. I discussed five drivers of synonym numbers which proved to influence global angiosperm synonymy. Although synonymy is not a limitation per se, it likely impacts public repositories and checklists adversely.

Chapter 2. Three crucial aspects emerged from the study. First, the used *R* packages reliably removed incorrect or dubious coordinates. However, no package could identify misidentified specimens. Second, differences in the pipeline data did not translate into significant differences in downstream SDMs and S-SDMs. This suggests a remarkable robustness of these analyses towards data cleaning differences. Third, GBIF data requires further cleaning, and taxonomic data may need even more attention, possibly from experts, as detecting misidentified specimens in the public data providers proved to be challenging. However, automated detection of dubious taxa was recently started, using machine learning and artificial intelligence concepts (Nicolson, 2019). In the course of these activities, we suggest developing new processes and tools to detect misidentified specimens.

Chapter 3. Four essential interpretations came up from the study. First, the geographic distance between taxonomists caused widespread species and species with complex ranges to accumulate more synonyms. Also, the time passed since the publication of an accepted species played a significant role. Second, the rank positions of families and genera may hint at particular taxa being more appealing than others. Thus, the attractiveness of study objects in the selection process of researchers and the associated research bias may warrant further study. Third, the predictive power of the botanical continent patterns (high) and the number of continents a species is present (low) contradict each other. Therefore, it may be interesting to investigate the role of the botanical continent further. Fourth and finally, the outcome of this study likely has implications for a wider range of plant families and genera and might also extend to other groups of organisms.

Appendices

A

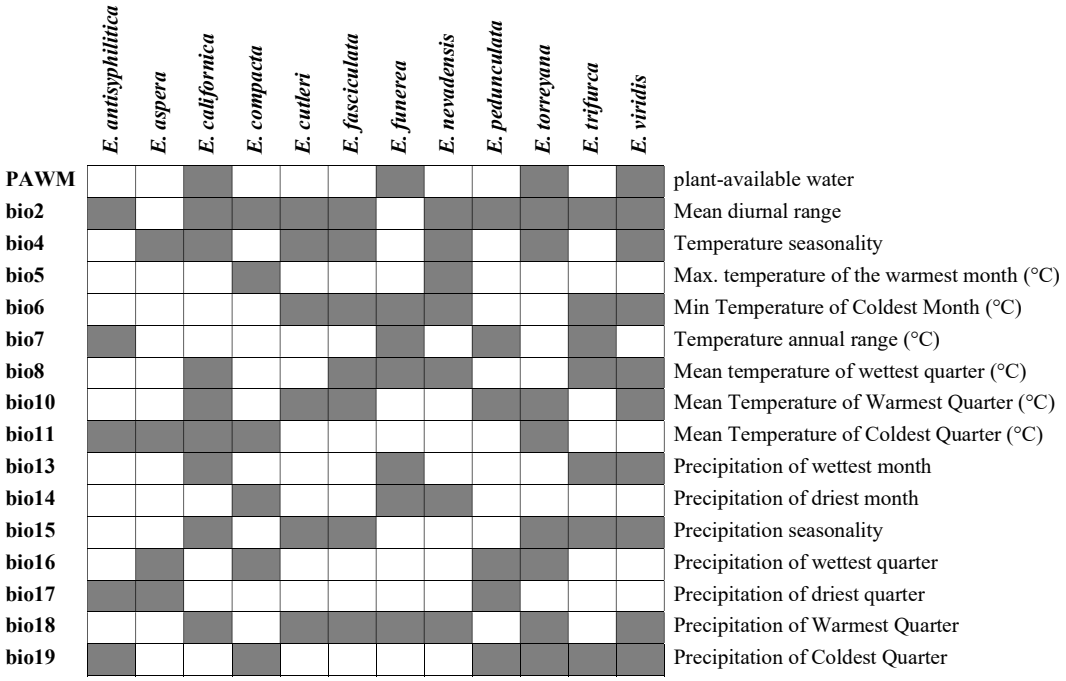
Supporting information to Chapter 2

Table A1. **Summary of the *Ephedra* species at the end of the pipelines (L1, record numbers per species).** We retained the North America native species (Stevenson, 1993) in the box for the downstream analyses and removed species outside the box, depending on exclusion criteria (Table 2.1). P0 data included all *Ephedra* records allocated to North America and served as the uncleaned control data to which we compared the other pipelines P1 to P6.

***Ephedra* species sums**		**9,484**	**6,687**	**5,198**	**5,396**	**5,395**	**5,196**	**13,889**
Taxon	**GEO**	**P1**	**P2**	**P3**	**P4**	**P5**	**P6**	**P0**
Ephedra antisyphilitica Berland. ex C.A. Mey.	NAm	375	211	185	187	187	184	612
Ephedra aspera S. Watson	NAm	1,478	1,166	835	920	919	837	1,994
Ephedra californica S. Watson	NAm	1,325	1,045	846	854	854	846	1,959
Ephedra compacta Rose	NAm	152	128	117	119	119	117	183
Ephedra cutleri Peebles	NAm	116	96	64	64	64	64	158
Ephedra fasciculata A. Nelson	NAm	184	158	119	129	129	119	245
Ephedra funerea Coville & C.V.Morton	NAm	198	146	113	113	113	113	328
Ephedra nevadensis S. Watson	NAm	1,264	952	666	672	672	664	1,845
Ephedra pedunculata Engelm. ex S.Watson	NAm	103	75	66	70	70	66	211
Ephedra torreyana S. Watson	NAm	811	571	435	445	445	435	1,210
Ephedra trifurca S. Watson	NAm	1,094	849	683	731	731	682	1,658
Ephedra viridis Coville	NAm	1,857	1,281	1,060	1,083	1,083	1,060	2,632
Ephedra coryi E. L. Reed	NAm	53	9	9	9	9	9	93
Ephedra miocenica Wodehouse (fossil)	NAm							2
Ephedra L. (indeterminates)	NAm	296						499
Ephedra hybrid	NAm							9
Ephedra form and variety	NAm	147						196
Ephedra altissima Desf.	EAs	2						4
Ephedra distachya L.	EAs	3						4
Ephedra equisetina Stapf	EAs	3						3
Ephedra fedtschenkoae Paulsen	EAs	1						1
Ephedra fragilis Desf.	EAs	1						3
Ephedra gerardiana Wallich ex C. A. Mey.	EAs	4						4
Ephedra major Host	EAs	1						1
Ephedra monosperma J. G. Gmel. *ex* C. A. Mey.	EAs	1						1
Ephedra przewalskii Stapf	EAs	2						2
Ephedra regeliana Florin	EAs	1						1
Ephedra sinica Stapf	EAs	3						8
Ephedra americana Humb. & Bonpl. ex Willd.	SAm	2						3
Ephedra andina Poepp & Endl.	SAm							2
Ephedra chilensis C. Presl.	SAm	2						2
Ephedra frustillata Miers	SAm	1						1
Ephedra triandra Tul.	SAm							2
Ephedra trifurcata Zöllner	SAm	3						11
Ephedra tweedieana C. A. Mey	SAm	1						2

Geographies, GEO: EAs, Eurasia; NAm, North America; SAm, South America.

Table A2. **Uncorrelated CHELSA climatology variables (Karger et al., 2017) and plant-available water (PAWM), used to fit and build the *Ephedra* diversity models** (twelve NAm *Ephedra* species; L4 and L5 data).



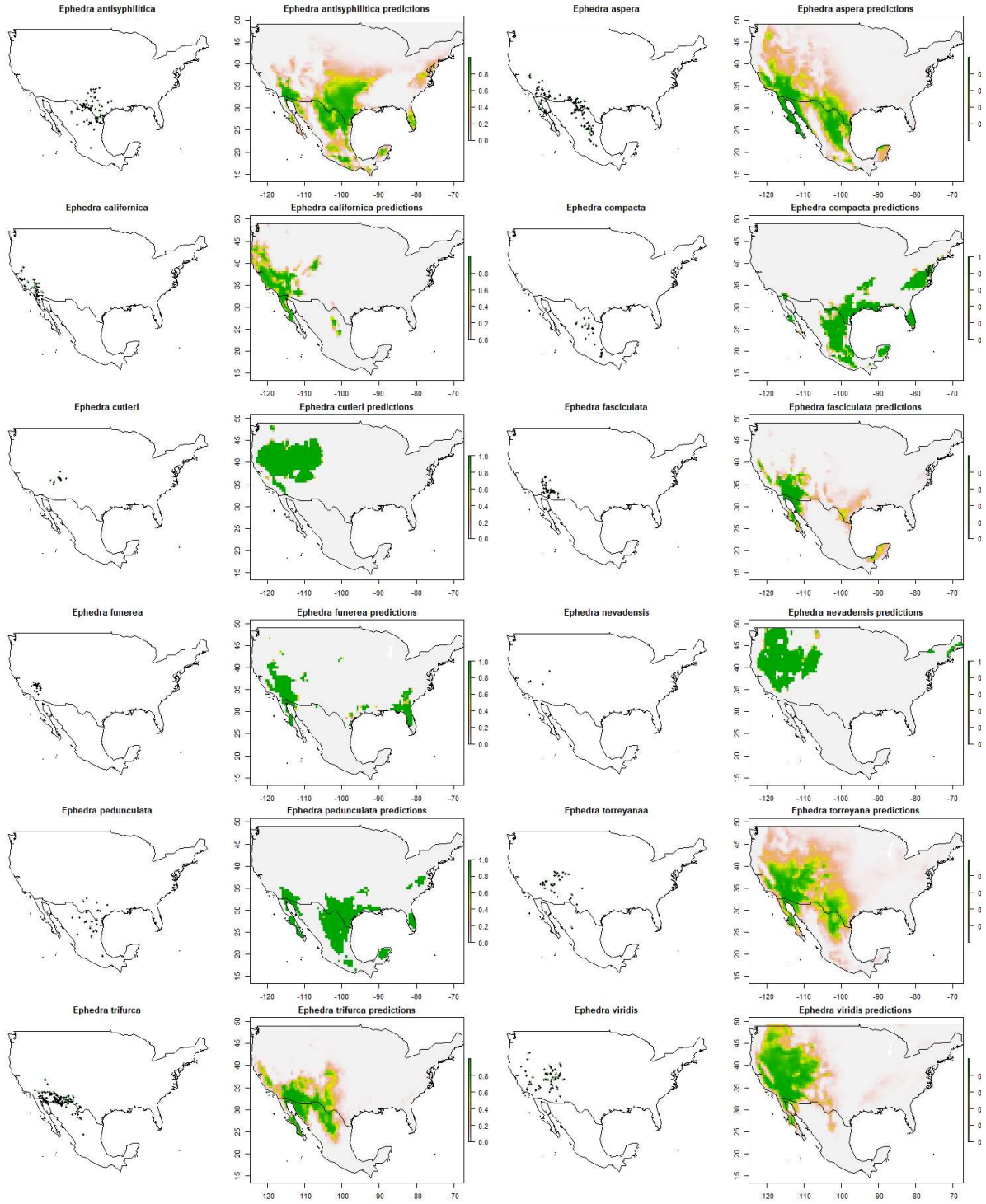


Figure A3. **Observed occurrences and predicted ranges of each North American *Ephedra* species, from L5 expert data**

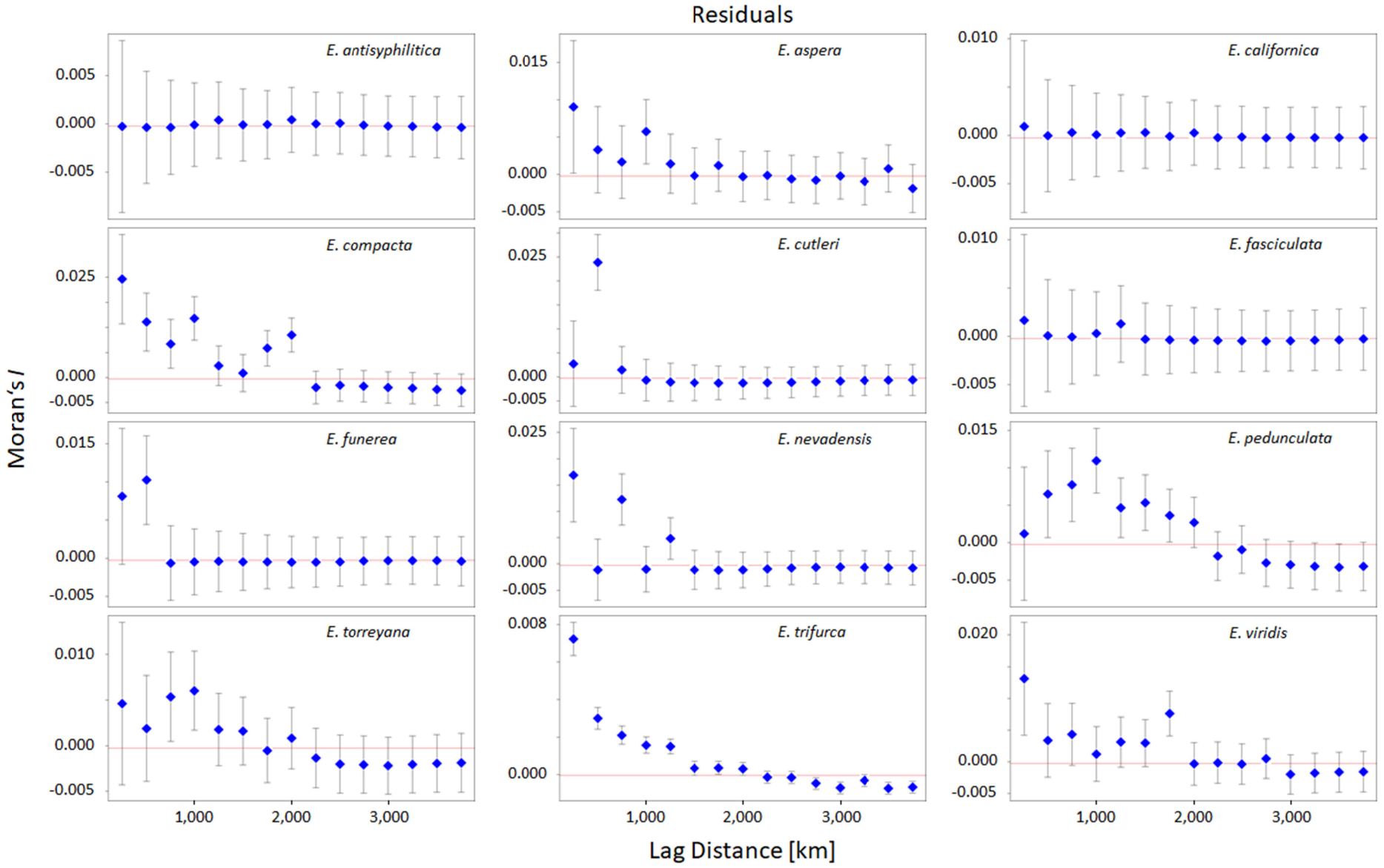


Figure A4. **Spatial correlograms using Moran's *I* of raw species occurrence and residual variation after fitting the examined environmental variables at a grain size of 0.5, example from L5 expert data (*R* package *spdep*, Bivand et al., 2015).** All correlograms are significant ($p<0.0005$), with the exception of *E. antisyphilitica* ($p=0.20$).

B

Supporting information to Chapter 3

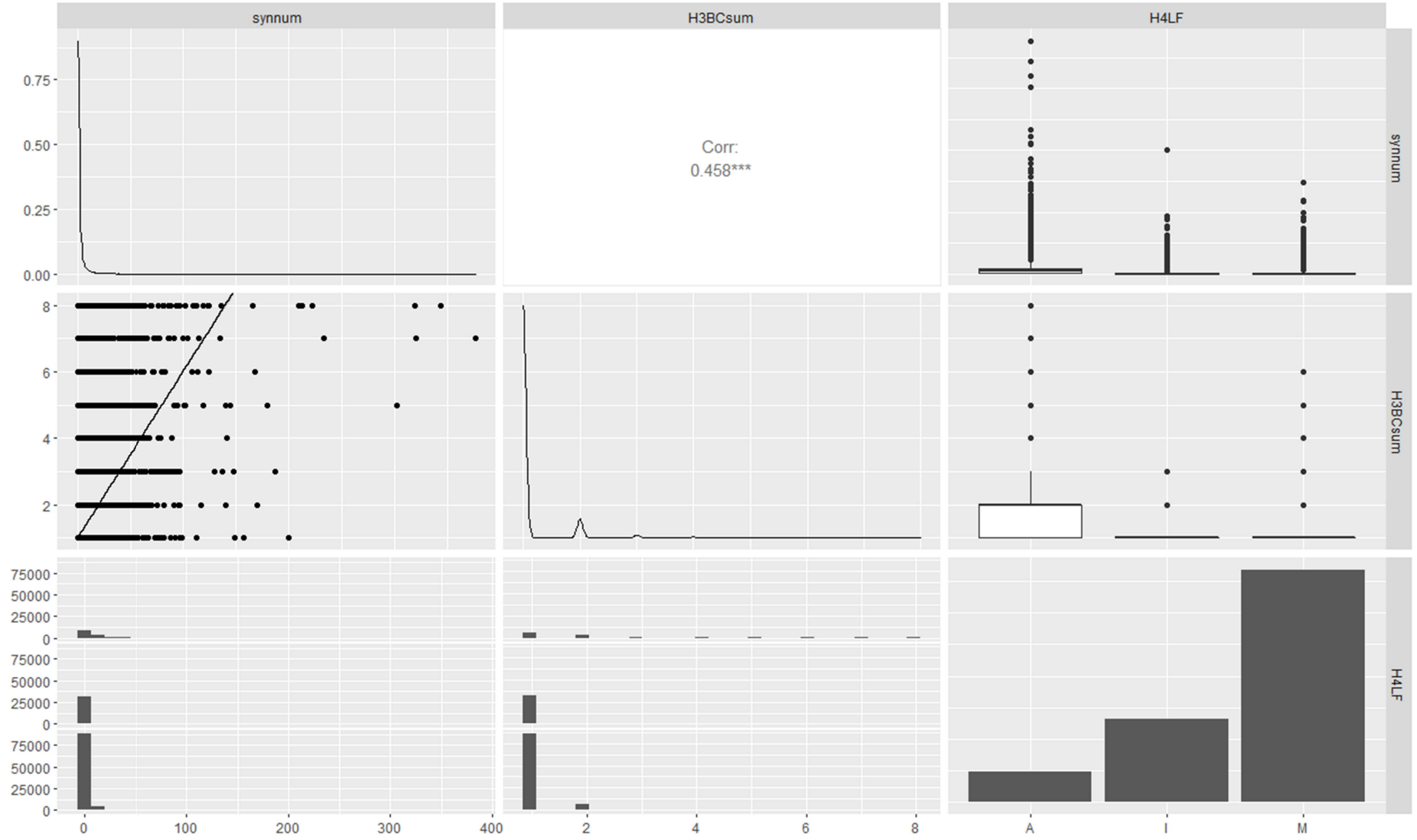


Figure B1(a). **Correlation test of predictor pair "Number of botanical continents present" (H3BCsum) and "Insularity" (H4LF).** The correlation coefficient shows that the predictors are likely uncorrelated (0.458). (Plot: GGally package, ggpairs function, Schloerke et al., 2018).

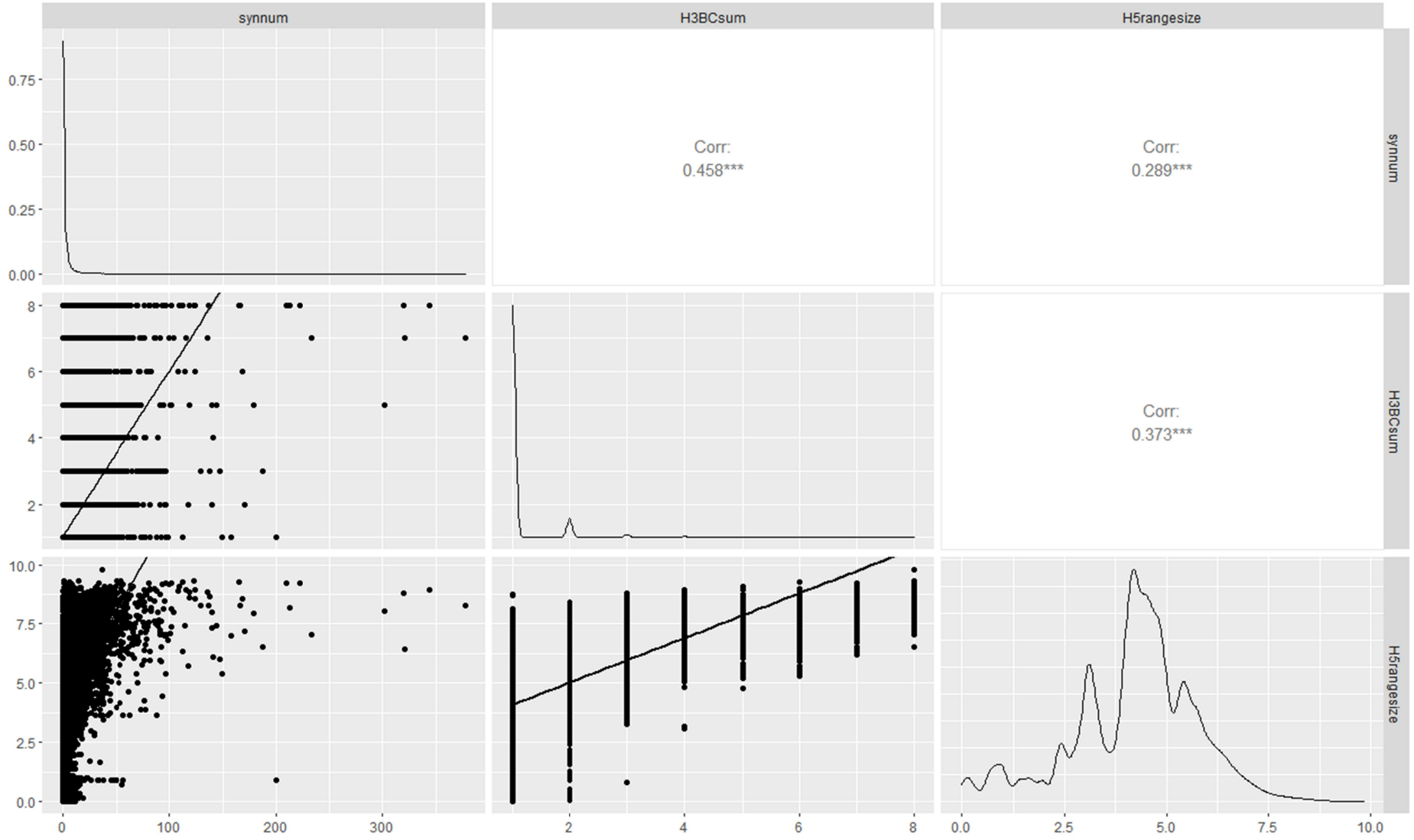


Figure B1(b). **Correlation test of predictor pair "Number of botanical continents present" (H3BCsum) and "Range size" (H5Rangesize).** The correlation coefficients show that the predictors are likely uncorrelated.

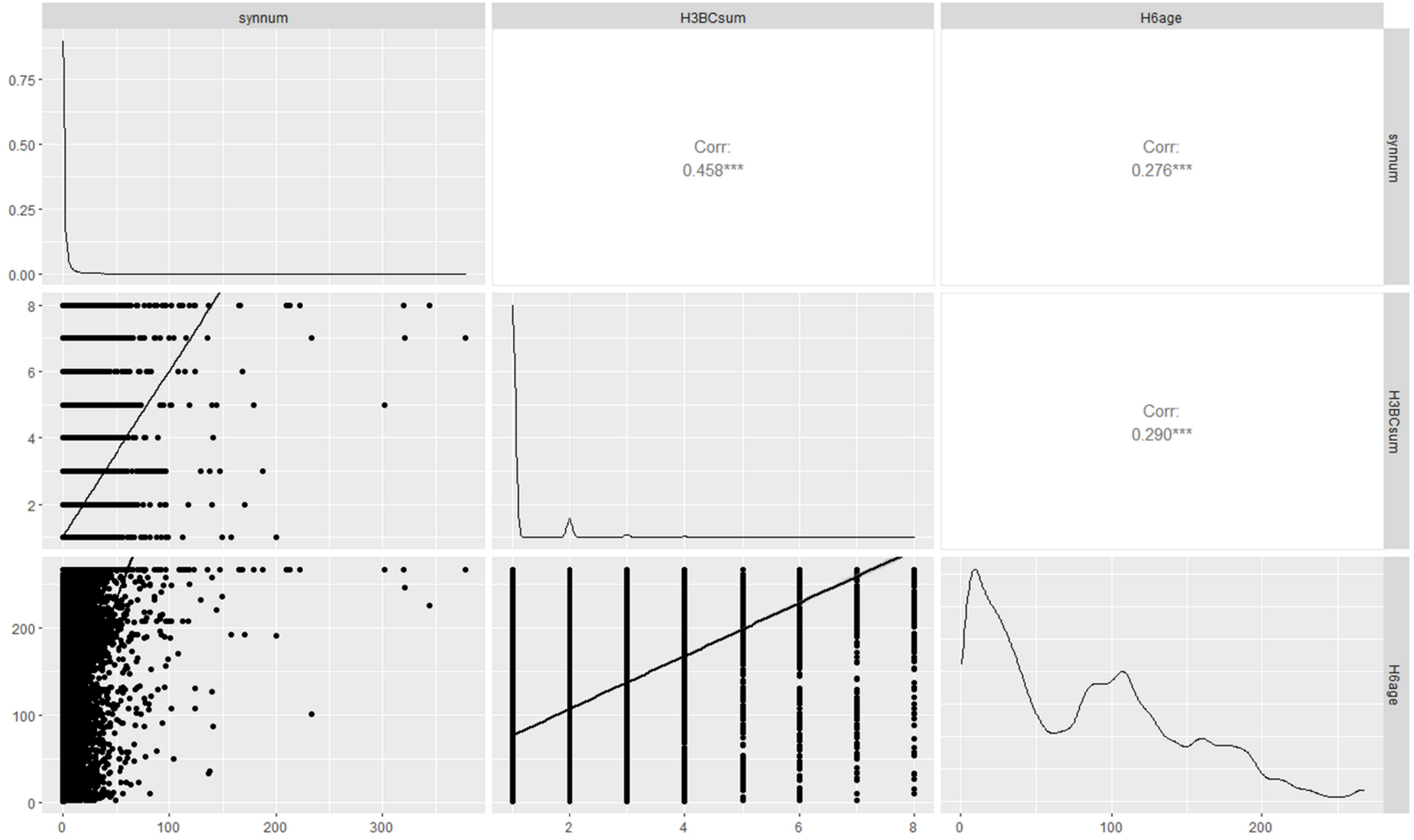


Figure B1(c). **Correlation test of predictor pair "Number of botanical continents present" (H3BCsum) and "Age of a species' name" (H6age, as the proxy for the time passed since the publication of the accepted name).** The correlation coefficients show that the predictors are likely uncorrelated.

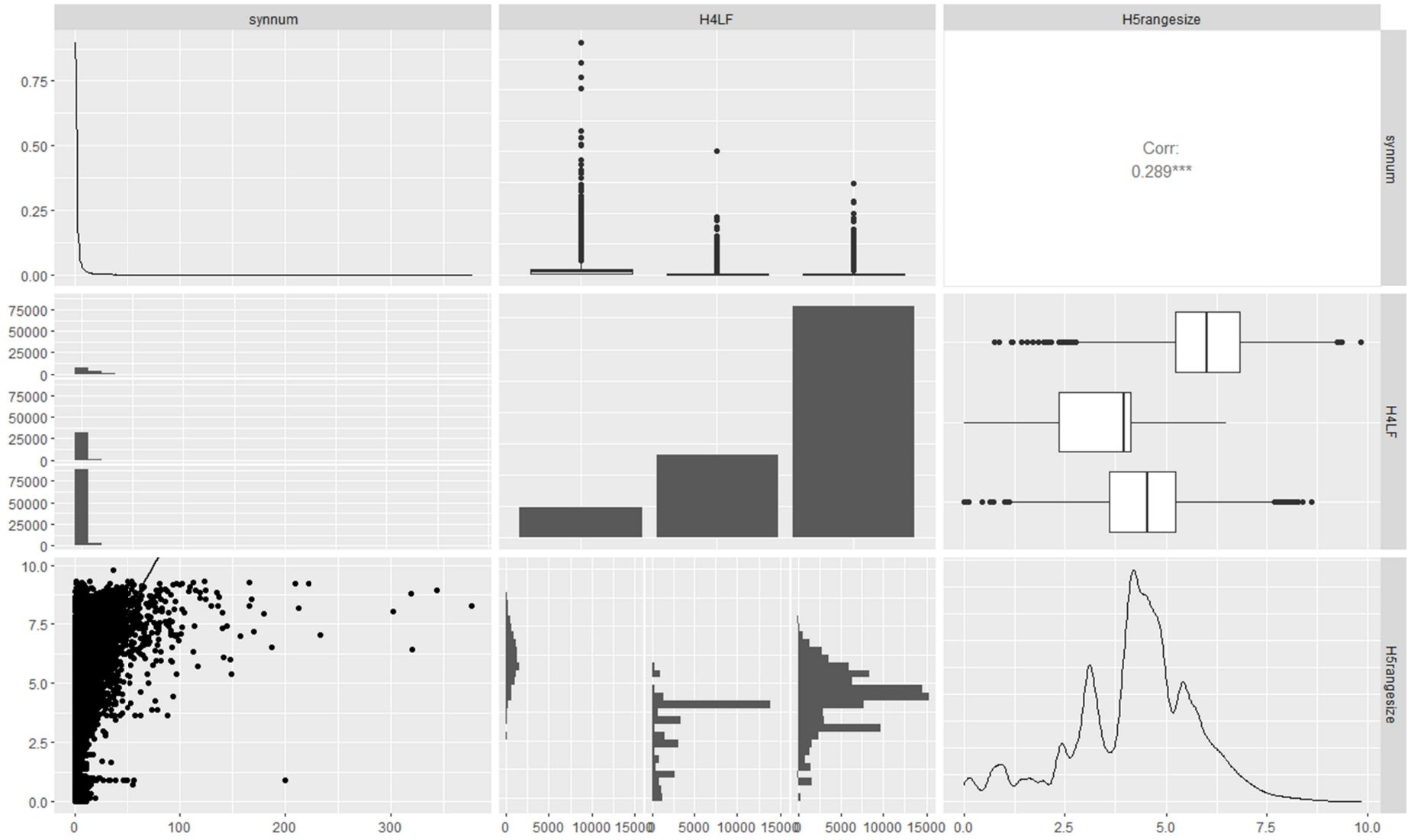


Figure B1(d). **Correlation test of predictor pair "Insularity" (H4LF) and "Range size" (H5rangesize).** The correlation coefficient shows that the predictors are likely uncorrelated.

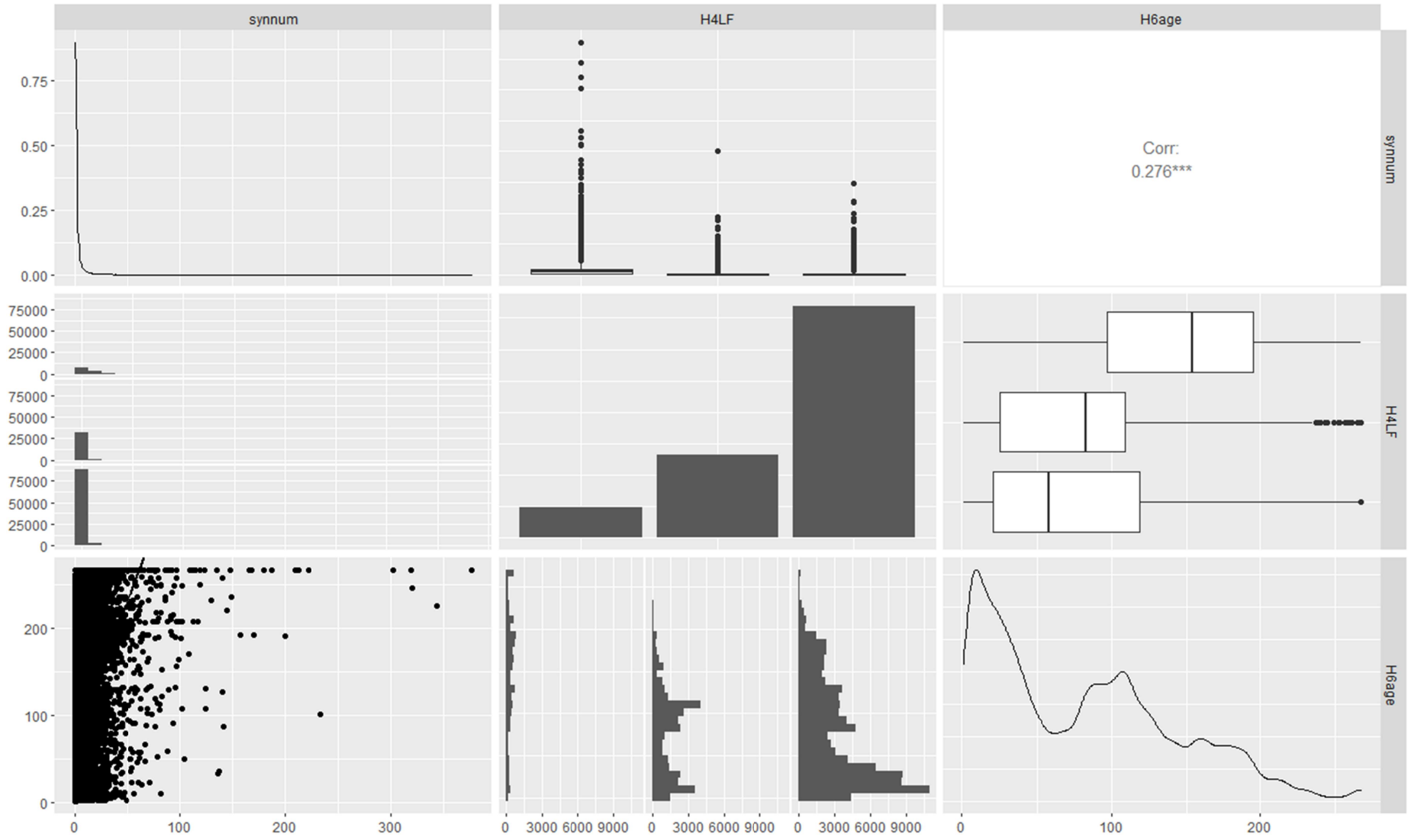


Figure B1(e). **Correlation test of predictor pair "Insularity" (H4LF) and "Age of a species' name" (H6age, as the proxy for the time passed since the publication of the accepted name).** The correlation coefficient shows that the predictors are likely uncorrelated.

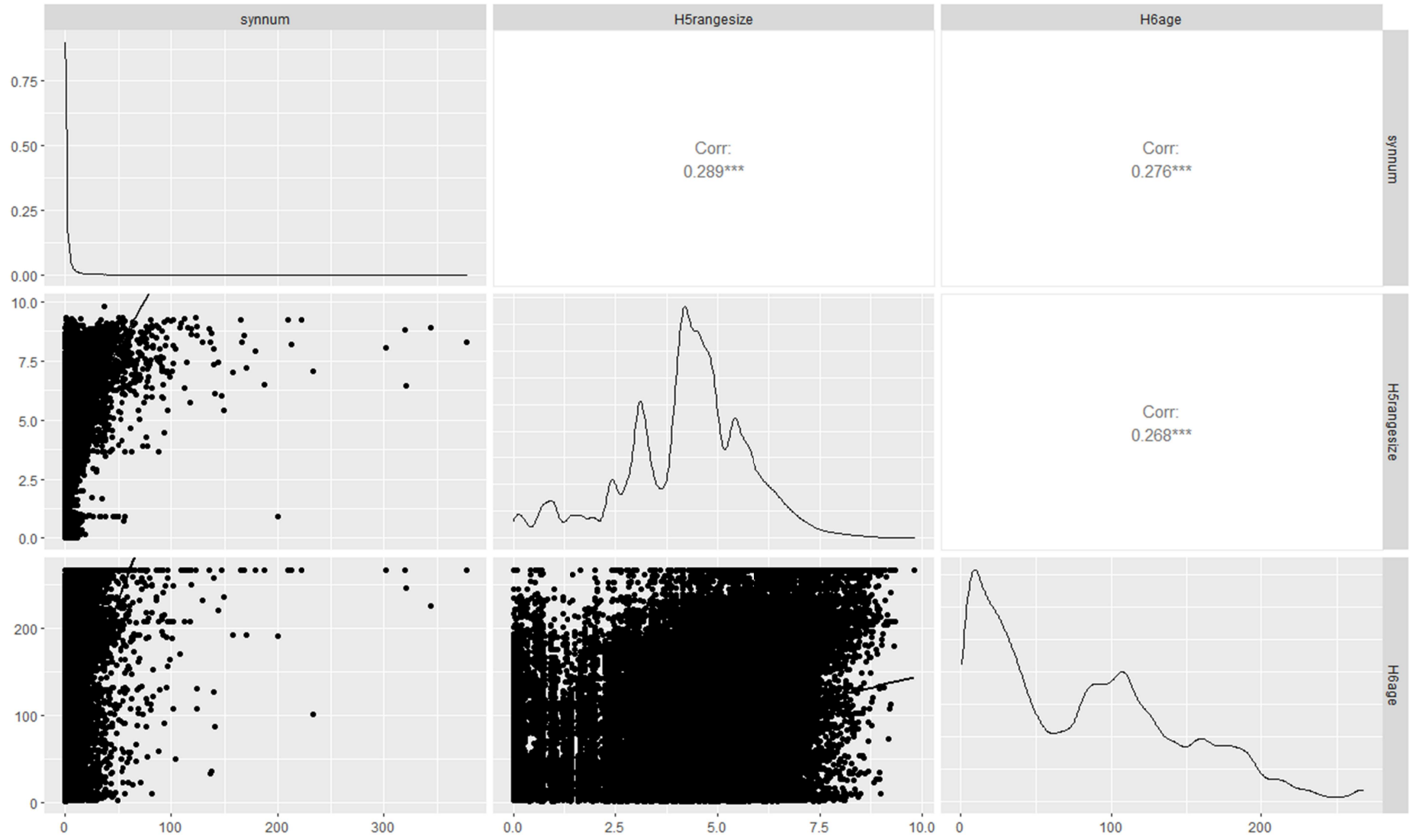
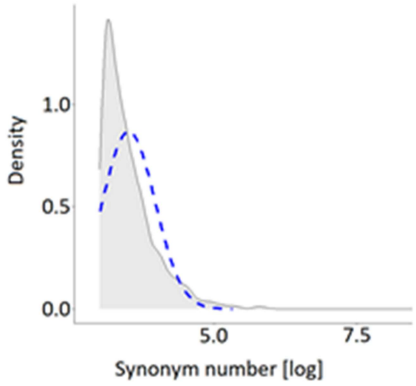


Figure B1(f). **Correlation test of predictor pair "Range size" (H5rangesize) and "Age of a species' name" (H6age, as the proxy for the time passed since the publication of the accepted name).** The correlation coefficients show that the predictors are likely uncorrelated.

Table B2. **(a) Density distribution of the log-transformed synonym number (*synNum*).** Observed distribution: grey; superimposed normal distribution: dashed blue. The *synNum* showed non-normal distribution with positive (right) skewness (skewness coefficient: 1.33, kurtosis: 4.68), indicating zero-inflation in the count data (Density plot: *ggdensity*, Kassambara, 2020a). **(b) Analysis of potential count data issues and solutions.** Frequent issues to be handled in count data are zero-inflation and overdispersion. Table: Comparison of three suitable model-fitting methods to best handle the count data issues, using the final model parameters (*lme4*: Bates et al., 2015). The Poisson distribution, that included zero-inflation (*glmmTMB* package, Bolker, 2016), showed the optimal model-fitting results in the *DHARMa* diagnostic tests (Hartig, 2020).

(a)



(b)

GLMM	**Poisson**	**Poisson/zi**	**negative binomial**
***R* package**	*lme4*::glmer	***glmmTMB***	*lme4*::glmnb
R^2 cond	0.939	**0.958**	0.705
R^2 marg (fixed)	0.398	**0.414**	0.361
AIC (E+05)	5.032	**4.880**	4.022
RMSE	3.948	**4.005**	4.882
***DHARMa* diagnostics**			
KS test: deviation	p = 0, sign.	**p = 0, sign.**	p = 0, sign.
Dispersion: dev.	p = 0.032, sign.	**p = 0.504, n. sign.**	p = 0, sign.
Outlier test: dev.	p = 0.004, n. sign.	**p = 0.454, n. sign.**	p = 0, n. sign.

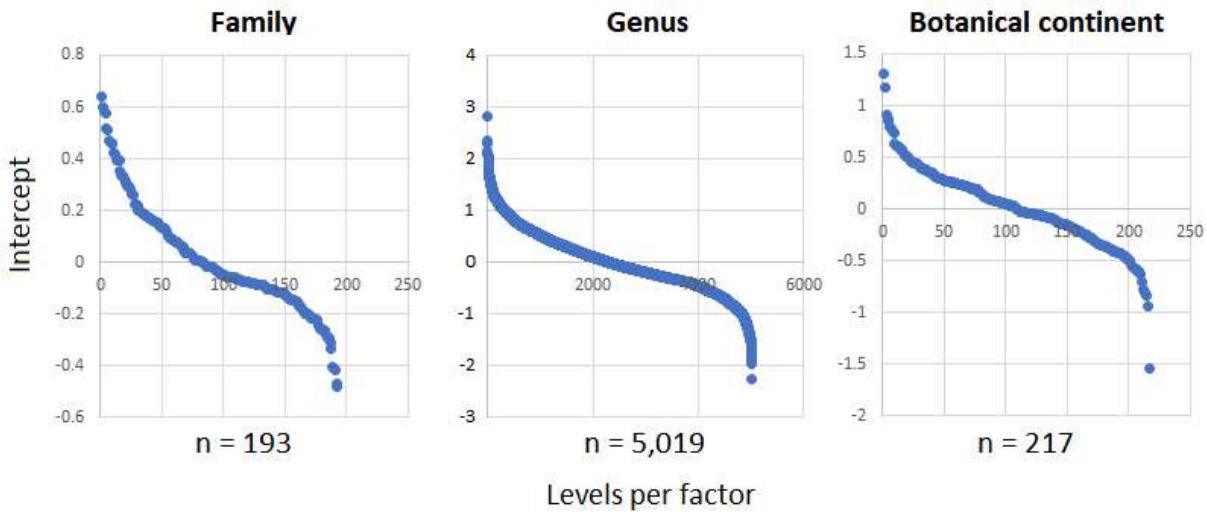


Figure B3. **Random Factor selection: Taxonomic Family and Genus, and Botanical Continents, a Species is Present.** Each grouping factor per random factor has its own random intercept. We selected these predictor variables with high level numbers as random factor (McGill, 2015). The three variables also exhibited low standard errors and very low *p*-values (< 0.001), suggesting that they were predictive.

Tables B4. **Four selected models of global angiosperm synonymy: Performance diagnostics and model evaluation. (a)** Determining details of the model, by the explanatory variables, we found that model 4 minimized the Akaike information criterion, increased the pseudo-R^2s (Diagnostics: *jtools*, Long, 2017, *sjPlot*, Lüdecke, 2021), and reduced overdispersion and zero inflation below a significant threshold, as als shown in the final *DHARMa* diagnostic tests (Appendix, Figure A5). **(b)** Four global models of angiosperm synonymy. Selection conditions of the models were: (1) AIC at a stable minimum, and (2) a maximized pseudo-R^2. Result of GLMM of a combined nine-predictor model, by random factors and fixed factors. H1 to H5: Hypotheses (see: Table 2). ***, $p < 0.001$.

(a)	GLMM	Model 1	Model 2	Model 3	Model 4
	Performance parameters:				
	R^2 cond	0.964	0.964	0.958	**0.958**
	R^2 marg (fixed factors)	0.421	0.440	0.396	**0.414**
	Random factor share	0.543	0.524	0.562	**0.544**
	AIC (E+05)	4.882	4.882	4.880	**4.880**
	RMSE	4.005	4.005	4.005	**4.005**
	***DHARMa* residual diagnostics:**				
	KS test: deviation	p = 0, sign.	p = 0, sign.	p = 0, sign.	**p = 0, sign.**
	Dispersion: dev.	p = 0.008, sign.	p = 0.016, sign.	p = 0.252, n. sign.	**p = 0.504, n. sign.**
	Outlier test: dev.	p = 0, sign.	p = 0.230, n. sign.	p = 0, sign.	**p = 0.454, n. sign.**

(b)	Factor	Model 1	Variation	Model 2	Variation	Model 3	Variation	Model 4	Variation
	Random Factor R^2 share	**0.543**		**0.524**		**0.562**		**0.544**	
	Botanical continents	0.312	31.2%	0.301	30.1%	0.311	31.1%	0.302	30.2%
	Genus of species	0.231	23.1%	0.223	22.3%	0.231	23.1%	0.223	22.3%
	Family of species	-	0.0%	-	0.0%	0.020	2.0%	0.019	1.9%
	Fixed Factor R^2 (Marg.)	**0.421**		**0.440**		**0.396**		**0.414**	
	Range size	0.214	21.4%	0.215	21.5%	0.202	20.2%	0.201	20.1%
	Age of accepted name	0.118	11.8%	0.118	11.8%	0.111	11.1%	0.111	11.1%
	Insularity	0.058	5.8%	0.058	5.8%	0.054	5.4%	0.054	5.4%
	No. inhab. continents	0.031	3.1%	0.031	3.1%	0.029	2.9%	0.029	2.9%
	Range size * Insularity	-	0.0%	0.018	1.8%	-	0.0%	0.019	1.9%
	Total R^2 (Cond.)	**0.964**		**0.964**		**0.958**		**0.958**	

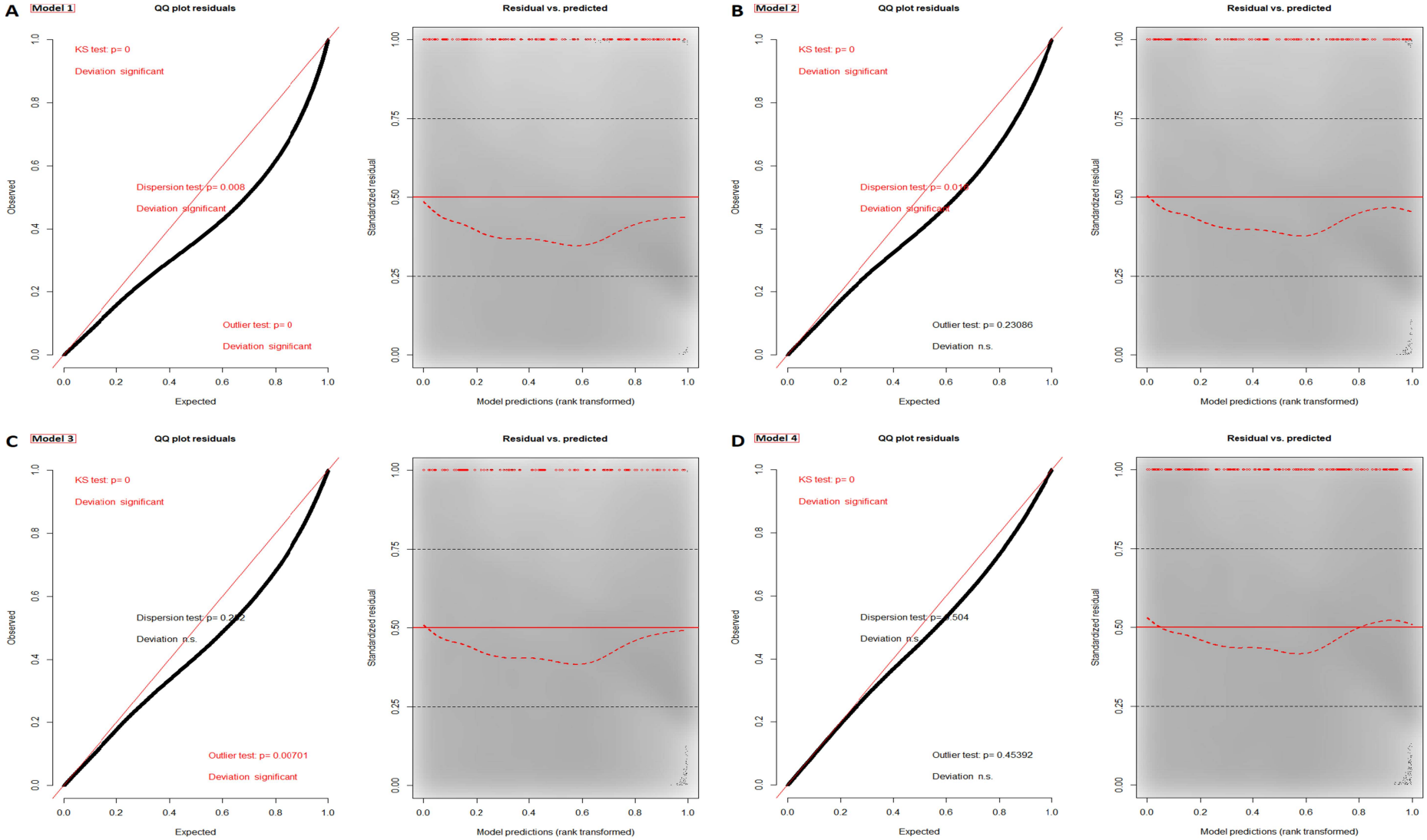


***Figure B5. DHARMa* diagnostic protocols (Hartig, 2020) for the four best-performing models 1 to 4.** See also: Table 3 and Appendix, Tables A4, (a) and (b) for further details.

***Table B6(a).* 25 angiosperm families with the highest *synRates*.** Listed are families with more than 10 accepted names. Large families such as Orchidaceae (28,899 accepted names, rank 99), Rubiaceae (10,796 accepted names, rank 94), and Myrtaceae (5,778 accepted names, rank 96) rank in the middle, due to their *synRate*. **B6(b). 25 angiosperm genera with the highest *synRates*.** Notably, many of the genera in the table have only one or a few accepted species (names). Images: Angiosperm families (Table A6(a): A. *Canna generalis,* Pos. 1; B: *Potamogeton gramineus*, Pos.2) and genera (Table A6(b): C. *Ricinus communis,* Pos. 1; D: *Phillyrea angustifolia,* Pos. 2). Abbreviations: *accNum*: Number of accepted species. *synNum*: Number of synonyms. *synRate*: synonymy rate.

Table B6(a).

Pos	family	accNum	synNum	synRate%	Ratio: synnum/accnum
1	Cannaceae	12	238	95.20%	19.8
3	Potamogetonaceae	106	790	88.20%	7.5
5	Ruppiaceae	11	69	86.30%	6.3
7	Irvingiaceae	12	67	84.80%	5.6
8	Paeoniaceae	36	187	83.90%	5.2
10	Betulaceae	172	846	83.10%	4.9
11	Stilbaceae	21	100	82.60%	4.8
12	Juncaginaceae	22	99	81.80%	4.5
13	Cornaceae	103	463	81.80%	4.5
14	Typhaceae	62	272	81.40%	4.4
15	Pontederiaceae	45	193	81.10%	4.3
17	Alismataceae	138	583	80.90%	4.2
19	Poaceae	11540	47443	80.40%	4.1
20	Tofieldiaceae	28	114	80.30%	4.1
24	Basellaceae	19	76	80.00%	4.0
26	Fagaceae	958	3757	79.70%	3.9
28	Oleaceae	619	2162	77.70%	3.5
29	Plantaginaceae	57	198	77.60%	3.5
31	Cymodoceaceae	18	61	77.20%	3.4
36	Altingiaceae	15	48	76.20%	3.2
37	Pandaceae	17	54	76.10%	3.2
39	Juncaceae	470	1460	75.60%	3.1
40	Bignoniaceae	874	2710	75.60%	3.1
41	Melanthiaceae	184	554	75.10%	3.0
46	Nothofagaceae	38	113	74.80%	3.0

Table B6(b).

Pos	H1family	genus	recnum	synNum	synRate
1	Euphorbiaceae	*Ricinus* L.	1	212	99.5%
2	Oleaceae	*Phillyrea* L.	2	247	99.2%
3	Poaceae	*Avenula* (Dumort.) Dumort.	1	83	98.8%
4	Arecaceae	*Cocos* L.	1	56	98.2%
5	Apocynaceae	*Nerium* L.	1	45	97.8%
6	Campanulaceae	*Platycodon* A. DC.	1	41	97.6%
7	Poaceae	*Arctophila* Rupr. ex Andersson	1	41	97.6%
8	Poaceae	*Molinia* Schrank	2	77	97.5%
9	Lamiaceae	*Mentha* L.	24	889	97.4%
10	Poaceae	*Apluda* L.	1	36	97.3%
11	Poaceae	*Taeniatherum* Nevski	1	34	97.1%
12	Poaceae	*Vulpiella* (Batt. & Trab.) Burollet	1	34	97.1%
13	Poaceae	*Sasaella* Makino	11	341	96.9%
14	Poaceae	*Oplismenus* P. Beauv.	7	212	96.8%
15	Myrtaceae	*Blepharocalyx* O. Berg	4	120	96.8%
16	Araceae	*Pistia* L.	1	29	96.7%
17	Poaceae	*Vahlodea* Fr.	1	27	96.4%
18	Potamogetonaceae	*Stuckenia* Börner	7	176	96.2%
19	Hydrocharitaceae	*Hydrilla* Rich.	1	25	96.2%
20	Asparagaceae	*Eustrephus* R.Br.	1	25	96.2%
21	Potamogetonaceae	*Groenlandia* J. Gay	1	25	96.2%
22	Apocynaceae	*Apocynum* L.	4	97	96.0%
23	Poaceae	*Trachypogon* Nees	4	97	96.0%
24	Poaceae	*Dupontia* R. Br.	1	24	96.0%
25	Poaceae	*Ampelodesmos* Link	1	24	96.0%

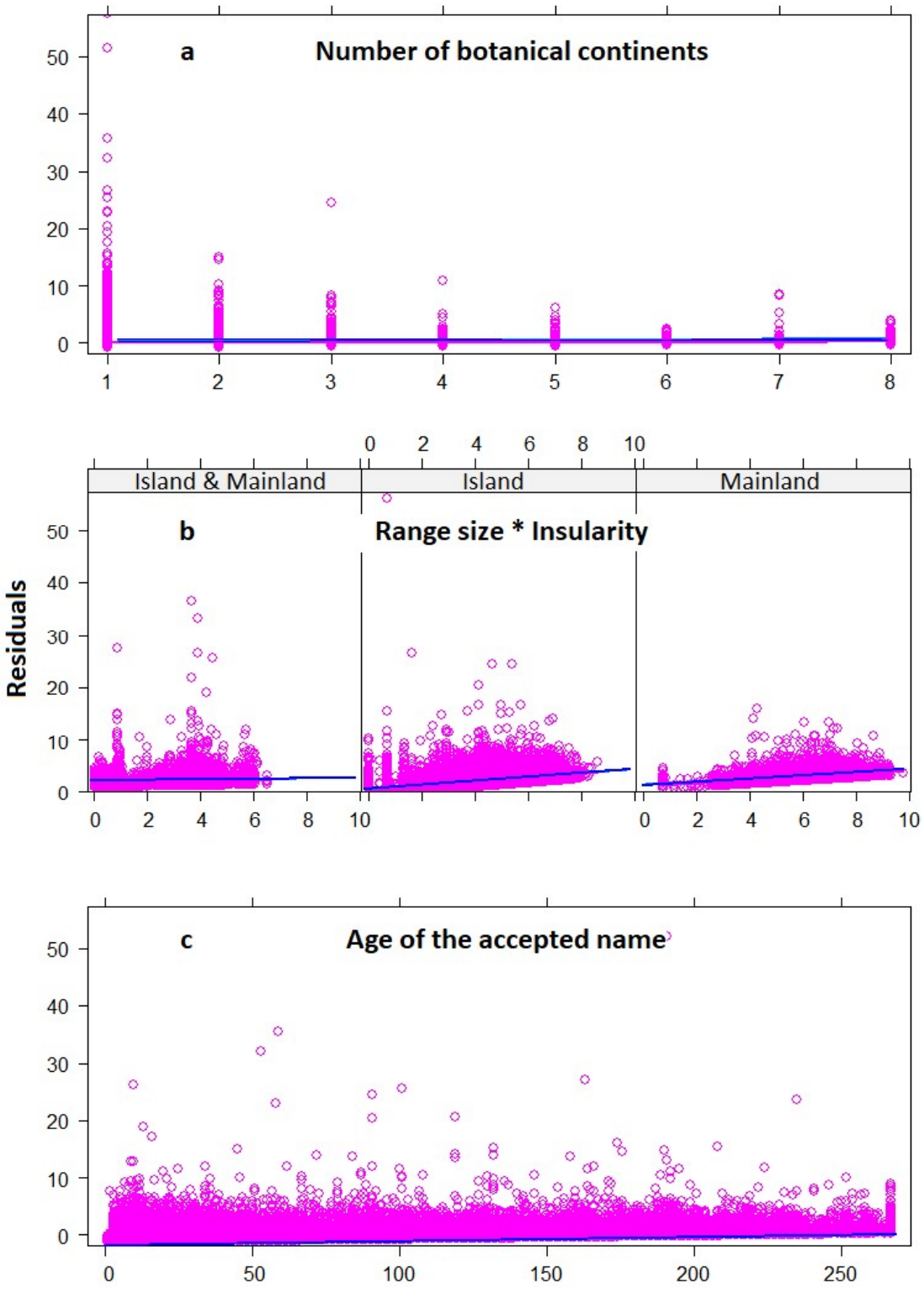


***Figure A7*. Working residuals of the global model of angiosperm synonymy**. (a) The working residuals vary with the *number of botanical continents, a species is present* (variable *BCNum*; (b), the working residuals vary within the different *insularities* and the *range size*, respectively, (c) The working residuals vary with the *age of an accepted name*. Details regarding the predictor rankings, see Table 3). All residual plots support the confidence intervals of the predicted regression lines. The plots were prepared, using the *effects* package (Fox et al. 2016).

References

Ackery, P. R., & Vane-Wright, R. I. (1984). Milkweed butterflies - their cladistics and biology, being an account of the natural history of the Danainae, a subfamily of the Lepidoptera, Nymphalidae. London: British Museum (Natural History).

Ahrends, A., Rahbek, C., Bulling, M. T., Burgess, N. D., Platts, P.J., Lovett, J.C., Kindemba, V. W., Owen, N., Sallu, A. N., Marshall, A. R., Mhoro, B.E., Fanning & E., Marchant, R. (2011). Conservation and the botanist effect. *Biological Conservation* **144**: 131–140. DOI: https://doi.org/10.1016/j.biocon.2010.08.008.

Alroy, J. (2002). How many named species are valid? *Proceedings of the National Academy of Sciences* **99**: 3706–3711. DOI: https://doi.org/10.1073/pnas.062691099.

Anderson, R. P., Araujo, M., Guisan, A., Lobo, J. M., Martinez-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).

Araújo, M. B., & Guisan, A. (2006). Five (or so) challenges for species distribution modelling. *Journal of Biogeography* **33**: 1677–1688. DOI: https://doi.org/10.1111/j.1365-2699.2006.01584.x.

Araújo, M. B., Anderson, R. P., Márcia Barbosa, A., Beale, C. M., Dormann, C. F., Early, R., Garcia, R. A., Guisan, A., Maiorano, L., Naimi, B., O'Hara, R. B., Zimmermann, N. E. & Rahbek, C. (2019). Standards for distribution models in biodiversity assessments. *Science Advances* **5**: eaat4858. DOI: https://doi.org/10.1126/sciadv.aat4858.

Bakshi, K. (2012). *Considerations for big data: Architecture and approach.* In: 2012 IEEE aerospace conference. [Accessed 2021 April 08]. Available from: https://ieeexplore.ieee.org/abstract/document/ 6187357.

Baselga, A., Lobo, J. M., Hortal, J., Jiménez-Valverde, A., & Gómez, J. F. (2010). Assessing alpha and beta taxonomy in eupelmid wasps: determinants of the probability of describing good species and synonyms. *Journal of Zoological Systematics* **48**: 40–49. DOI: https://doi.org/ 10.1111/j.1439-0469.2009.00523.x.

Baskauf, S. J., Wieczorek, J., Deck, J., & Webb, C. O. (2016). Lessons learned from adapting the Darwin Core vocabulary standard for use in RDF. *Semantic Web* **7**: 617–627. DOI: https://doi.org/10.3233/SW-150199.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using *lme4*. *Journal of Statistical Software* **67**: 1–48. DOI:10.18637/jss.v067.i01.

Beck, J., Ballesteros-Mejia, L., Nagel, P., & Kitching, I. J. (2013). Online solutions and the 'Wallacean shortfall': what does GBIF contribute to our knowledge of species' ranges? *Diversity and Distributions* **19**: 1043–1050. DOI: https://doi.org/10.1111/ ddi.12083.

Belbin, L., Daly, J., Hirsch, T., Hobern, D., & La Salle, J. (2013). A specialist's audit of aggregated occurrence records: An 'aggregator's' perspective. *ZooKeys* **305**: 67–76. DOI: https://doi.org/10.3897%2Fzookeys.305.5438.

Bell, A., Fairbrother, M., & Jones, K. (2019). Fixed and random effects models: making an informed choice. *Quality & Quantity* **53**: 1051–1074. DOI: https://doi.org/10.1007/s11135-018-0802-x.

Bell, M. L., & Grunwald, G. K. (2011). Small sample estimation properties of longitudinal count models. *Journal of Statistical Computation and Simulation* **81**: 1067–1079. DOI: https://doi.org/10.1080/00949651003674144.

Biber, M. F., Voskamp, A., Niamir, A., Hickler, T., & Hof, C. (2020). A comparison of macroecological and stacked species distribution models to predict future global terrestrial vertebrate richness. *Journal of Biogeography* **47**: 114–129. DOI: https://doi.org/10.1111/ jbi.13696.

Bisang, I., & Urmi, E. (1994). Studies on the status of rare and endangered bryophytes in Switzerland. *Biological Conservation* **70**: 109–116. DOI: https://doi.org/10.1016/0006-3207(94)90278-X.

Bivand, R., Altman, M., Anselin, L., Assunção, R., Berke, O., Bernat, A., & Blanchet, G. (2015). Package *spdep*. *The Comprehensive R Archive Network*. [Accessed 2020 November 08]. Available from: https://www.yumpu.com/en/document/view/9283478/package-spdep-the-comprehensive-r-archive-network.

Bivand, R., Gómez-Rubio, V., & Rue, H. (2013). *Applied spatial data analysis with R, 2nd edition*. Springer, New York. [Accessed 2020 November 08]. Available from: https://asdar-book.org/

Bivand, R., Lewin-Koh, N., Pebesma, E., Archer, E., Baddeley, A., Bearman, N., & Golicher, D. (2021). *R* Package '*maptools*'. [Accessed 2022 May 25]. Available from: https://r-forge.r-project.org/projects/maptools/

Bivand, R., Ono, H., Dunlap, R., & Stigler, M. (2015). *R* Package '*classInt*'. [Accessed 2021 March 08]. Available from: https://github.com/r-spatial/classInt/

Bolker, B. (2016). *Getting started with the* glmmTMB *package*. *R* Foundation for Statistical Computing, Vienna, Austria. [Accessed 2020 December 02]. Available from: cran.uni-muenster.de.

Bougeard, S., & Dray, S. (2018). Supervised multiblock analysis in *R* with the ade4 package. *Journal of statistical software* **86**: 1–17. DOI: https://doi.org/10.18637/jss.v086.i01.

Brummitt, R.K. (2001). *World Geographical Scheme for Recording Plant Distributions*. Plant Taxonomic Database Standards No. 2, Edition 2. Hunt Institute for Botanical Documentation, Carnegie Mellon University, Pittsburgh. [Accessed 2017 October 10]. Available from: biofund.org.mz.

Bruun A. F. (1950). The Systema Naturae of the twentieth century. *Science* **112**: 342–343. DOI: https://doi.org/10.1126/science.112.2908.342.b.

Burnham, K. P., &Anderson, D. R. (2004). Multimodel inference: understanding AIC and BIC in model selection. *Sociological Methods & Research* **33**: 261–304. DOI: https://doi.org/10.1177%2F004912410426 8644.

Calabrese, J. M., Certain, G., Kraan, C., & Dormann, C. F. (2014). Stacking species distribution models and adjusting bias by linking them to macroecological models. *Global Ecology and Biogeography* **23**: 99–112. DOI: https://doi.org/10.1111/ geb.12102.

Caveney, S., Charlet, D. A., Freitag, H., Maier-Stolte, M., & Starratt, A. N. (2001). New observations on the secondary chemistry of world *Ephedra* (Ephedraceae). *American Journal of Botany* **88**: 1199–1208. DOI: https://doi.org/10.2307/3558330.

CBD. (2009). *The Convention on Biological Diversity Plant Conservation Report*: A Review of Progress in Implementing the Global Strategy of Plant Conservation. Montreal: Secretariat of the Convention on Biological Diversity. [Accessed 2021 July 28].

Chamberlain, S. (2020). *rgbif*: Interface to the Global Biodiversity Information Facility API ver. 3.2.0. [Accessed 2021 January 03]. Available from: https://docs.ropensci.org/rgbif/

Chandra, A., & Idrisova, A. (2011). Convention on Biological Diversity: a review of national challenges and opportunities for implementation. *Biodiversity and Conservation* **20**: 3295–3316. DOI: https://doi.org /10.1007/s10531-011-0141-x.

Chapman, A. D. (2005). *Principles and Methods of Data Cleaning – Primary Species and Species-occurrence Data, version 1.0*. Report for the Global Biodiversity Information Facility, Copenhagen.

Chapman, A. D., Belbin, L., Zermoglio, P. F., Wieczorek, J., Morris, P. J., Nicholls, M., Rees, E. R., Veiga, A. K., Thompson, A., Saraiva, A. M., James, S. A., Gendreau, C., Benson, A. & Schigel, D. (2020). *Developing Standards for Improved Data Quality and for Selecting Fit for Use* Biodiversity *Data*. Biodiversity Information Science and Standards 4: e50889. [Accessed 2020 November 13]. Available from: https://doi.org/10.3897/biss.4.50889.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. *SPSS inc* **9**: 13. [Accessed 2019 October 27].

Christenhusz, M. J., Vorontsova, M. S., Fay, M. F. & Chase, M. W. (2015). Results from an online survey of family delimitation in angiosperms and ferns: recommendations to the Angiosperm Phylogeny Group for thorny problems in plant classification. *Botanical Journal of the Linnean Society* **178**: 501–528. DOI: https://doi.org/10.1111/boj.12285.

Claridge, M. F., Dawah, H. A., & Wilson, M. R. (1997). *Species: the units of biodiversity*. Chapman and Hall Ltd.

Collen, B., Ram, M., Zamin, T., & McRae, L. (2008). The tropical biodiversity data gap: addressing disparity in global monitoring. *Tropical Conservation Science* **1**: 75–88. DOI: https://doi.org/10.1177%2F194008290800100202.

Costello, M. J., Michener, W. K., Gahegan, M., Zhang, Z. Q., & Bourne, P. E. (2013). Biodiversity data should be published, cited, and peer reviewed. *Trends in Ecology and Evolution* **28**: 454-461. DOI: https://doi.org/10.1016/j.tree.2013.05.002.

Crane, P. R. (1985). Phylogenetic analysis of seed plants and the origin of angiosperms. *Annals of the Missouri Botanical Garden* **72**: 716–793. DOI: https://doi.org/10.2307/2399221.

Cutler, H.C. (1939). Monograph of the North American species of the genus *Ephedra*. *Annals of the Missouri Botanical Garden* **26**: 373–428. DOI: https://doi.org/10.2307/2394299.

De Wet, J. M. J., & Harlan, J. R. (1972). The origin and domestication of *Sorghum bicolor*. *Economic Botany* **25**: 128–135. [Accessed 2020 December 15]. Available from: https://www.jstor.org/stable/4253238.

Demchenko, Y., Grosso, P., De Laat, C., & Membrey, P. (2013). *Addressing big data issues in scientific data infrastructure*. International conference on collaboration technologies and systems (CTS). IEEE, 2013. [Accessed 2017 October 10]. Available from: https://ieeexplore.ieee.org/abstract/document/6567203.

Despot-Belmonte, K., Neßhöver, C., Saarenmaa, H., Regan, E., Meyer, C., Martins, E., Groom, Q., Hoffmann, A., Caine, A., Bowles-Newark, N., Bae, H., Lange Canhos, D. A., Stenzel, S., Bowler, D., Schneider, A., Weatherdon, L. V. & Martin, C. S. (2017). Biodiversity data provision and decision-making-addressing the challenges. *Research Ideas and Outcomes* **3**: 1-11e12165. [Accessed 2021 December 21]. Available from: https:// doi.org/10.3897/rio.3.e12165.

Di Cola, V., Broennimann, O., Petitpierre, B., Breiner, F. T., d'Amen, M., Randin, C., Engler, R. Pottier, J., Pio, D., Dubuis, A., Pellissier, L., Mateo, R. G., Hordijk, W., Salamin, N., & Guisan, A. (2017). ecospat: an *R* package to support spatial analyses and modeling of species niches and distributions. *Ecography* **40**: 774–787. DOI: https://doi.org/10.1111/ecog.02671.

Dial, H.L. (2012). Plant guide for *sorghum* (*Sorghum bicolor* L.). USDA-Natural Resources Conservation Service, Tucson Plant Materials Center, Tucson, AZ. Accessed on: 2022-07-11. Available at: https://plants.usda.gov/pdf/pg_sobi_2.

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., García Marquéz, J. R., Gruber, B., Lafourcade, B., Leitão, P. J., Münkemüller, T., McClean, C., Osborne, P. E., Reineking, B., Schröder, B., Skidmore, A. K., Zurell, D., & Lautenbach, S. (2013). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography* **36**: 27–46. DOI: https://doi.org/10.1111/j.1600-0587.2012.07348.x.

Doyle, J. A., & Donoghue, M. J. (1986). Seed plant phylogeny and the origin of angiosperms: an experimental cladistic approach. *The Botanical Review* **52**: 321–431. DOI: https://doi.org/10.1007/BF02861082.

Doyle, J. A., & Donoghue, M. J. (1992). Fossils and seed plant phylogeny reanalyzed. *Brittonia* **44**: 89–106. DOI: https://doi.org/10.2307/2806826.

Dubois, A. (2008). A partial but radical solution to the problem of nomenclatural taxonomic inflation and synonymy load. *Biological Journal of the Linnean Society* **93**: 857–863. DOI: https://doi.org/10.1111/j.1095-8312.2007.00900.x.

Endress, P. K. (1997). Relationships between floral organization, architecture, and pollination mode in Dillenia (Dilleniaceae). *Plant Systematics and Evolution* **206**: 99–118. DOI: https://doi.org/10.1007/BF00987943.

Evenhuis, N. L. (2008). The "Mihi itch" – a brief history. *Zootaxa* **1890**: 59–68. DOI: https://doi.org/10.11646/zootaxa.1890.1.3.

Fenneman, J. (2017). *Synonyms Explained: Why Plants Sometimes Have Other Scientific Names*. Electronic Atlas of the Flora of British Columbia (eflora.bc.ca). [Accessed 2019 June 13]. Available from: http://ibis.geog.ubc.ca/biodiversity/eflora/VascularPlantSynonymy.html.

Franz, N. M., Peet, R. K., & Weakley, A. S. (2008). *On the use of taxonomic concepts in support of biodiversity research and taxonomy*. The New Taxonomy. CRC Press.

Freitag, H, & Maier-Stolte, M. (1994). *Ephedraceae*. In: Browicz K. ed., Poznan: Chorology of trees and shrubs in Southwest Asia and adjacent regions. Polish Scientific Publishers.

Freitag, H., & Maier-Stolte, M. (2003). The genus *Ephedra* in NE tropical Africa. *Kew Bulletin* **58**: 415–426. DOI: https://doi.org/10.2307/4120624.

Fu L.G., Yu, Y. F. & Riedl, H. (1999). *Ephedraceae*. In: Wu Z. Y., Raven P. H. eds. Flora of China. Beijing: Science Press.

Garson, G. D. (2013). *Hierarchical linear modeling: Guide and applications*. Sage.

Gaston, K. J., & Mound, L. A. (1993). Taxonomy, hypothesis testing and the biodiversity crisis. *Proceedings of the Royal Society of London. Series B: Biological Sciences* **251**: 139–142. DOI: https://doi.org/10.1098/rspb.1993.0020.

GBIF. (2020a). GBIF Occurrence Download. [Accessed 2020 November 18]. Available from https://doi.org/10.15468/dl.2eg5ab.

GBIF. (2020b). GBIF Occurrence Download. [Accessed 2020 November 17]. Available from: https://doi.org/10.15468/dl.r2cg62.

GBIF. (2021c). Terms of use: Data agreement. [Accessed 2019 June 13]. Available from: https://www.gbif.org/terms.

GBIF.org. (2020). GBIF Home Page. [Accessed 2020 March 14]. Available from: https://www.gbif. org.

GBIF.org. (2022), GBIF Home Page. [Accessed 2022 January 07]. Available from: https:// www.gbif.org.

GMTED. (2020). GMTED Digital elevation data, elevation above sea level (mn30_grd.zip). [Accessed 2020 April 24]. Available from: https://topotools.cr.usgs.gov/gmted_viewer/viewer.htm.

Goodwin, Z. A., Harris, D. J., Filer, D., Wood, J. R., & Scotland, R. W. (2015). Widespread mistaken identity in tropical plant collections. *Current biology* **25**: R1066–R1067. DOI: https://doi.org/10.1016/j.cub.2015.10.002.

Gotelli, N. J. (2004). A taxonomic wish–list for community ecology. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* **359**: 585–597. DOI: https://doi.org/10.1098/rstb.2003.1443.

Graham, C. H., Ferrier, S., Huettman, F., Moritz, C., & Peterson, A. T. (2004). New developments in museum-based informatics and applications in biodiversity analysis. *Trends in Ecology and Evolution* **19**: 497–503. DOI: https://doi.org/10.1016/j.tree.2004.07.006.

Gueta, T., Carmel, Y. (2016). Quantifying the value of user-level data cleaning for big data: A case study using mammal distribution models. *Ecological Informatics* **34**: 139–145. DOI: https://doi.org/10.1111/j.1600-0587.2012.07348.x.

Guisan, A., Thuiller, W., & Zimmermann, N. E. (2017). *Habitat suitability and distribution models: with applications in R*. Cambridge University Press.

Guralnick, R. P., Hill, A. W., & Lane, M. (2007). Towards a collaborative, global infrastructure for biodiversity assessment. *Ecology Letters* **10**: 663–672. DOI: https://doi.org/10.1111/j.1461-0248.2007.01063.x.

Guralnick, R., Walls, R., & Jetz, W. (2018). Humboldt Core–toward a standardized capture of biological inventories for biodiversity monitoring, modeling and assessment. *Ecography* **41**: 713–725. DOI: https://doi.org/10.1111/ ecog. 02942.

Hamilton, A. J. (2005). Species diversity or biodiversity? *Journal of Environmental Management* **75**: 89–92. DOI: https://doi.org/10.1016/j.jenvman.2004.11.012.

Hardin, J. W., & Hilbe, J. (2007). *Generalized linear models and extensions*. Stata press.

Hartig, F. (2019). GLMM for unbalanced zero inflated data. [Accessed 2021 December 08]. Available from:https://stats.stackexchange.com/questions/396336/r-glmm-for-unbalanced-zero-inflated-data-glmmtmb.

Hartig, F. (2020). *DHARMa*: residual diagnostics for hierarchical (multi-level/mixed) regression models. *R* package version 0.3.3. [Accessed 2020 December 02]. Available from: https://cran.r-project.org/web/packages/DHARMa/index.html.

Haston, E., Richardson, J. E., Stevens, P. F., Chase, M. W., & Harris, D. J. (2009). The Linear Angiosperm Phylogeny Group (LAPG) III: a linear sequence of the families in APG III. *Biological Journal of the Linnean Society* **161**: 128–131. DOI: https://doi.org/10.1111/j.1095-8339.2009.01000.x.

Henrich, J., & Gil-White, F. (2001). The evolution of prestige: Freely conferred deference as a mechanism for enhancing the benefits of cultural transmission. *Evolution and Human Behaviour* **22**: 165–196. DOI: https://doi.org/10.1016/S1090-5138(00)00071-4.

Heß, D. (1990). *Die Blüte*. Stuttgart: Ulmer.

Hijmans, R. J., & Elith, J. (2019). Spatial Distribution Models. [Accessed 2019 December 08]. Available from: https://www.rspatial.org/sdm/SDM.pdf.

Hijmans, R. J., & Elith, J. (2020). *dismo*: Species Distribution Modeling. [Accessed 2020 December 08]. Available from: https://rspatial.org/raster/sdm/.

Hijmans, R. J., Phillips, S., Leathwick, & J., Elith, J. (2017). Package '*dismo*'. *Circles* **9**: 1–68. [Accessed 2019 December 08]. Available from: https://cran.microsoft.com/snapshot/2018-04-14/web/packages/dismo/dismo.pdf.

Hijmans, R.J., & van Etten, J. (2021). *raster*: Geographic data analysis and modeling. *R* package version 3.4-10. [Accessed 2021 January 21]. Available from: http://CRAN.R-project.org/package=raster.

Hobern, D., Baptiste, B., Copas, K., Guralnick, R., Hahn, A., van Huis, E., Kim, E. S., McGeoch, M., Naicker, I., Navarro, I., Noesgaard, D., Price, M., Rodrigues, A., Schigel, D., Sheffield, C. A., & Wieczorek, J. (2019). Connecting data and expertise: a new alliance for biodiversity knowledge. *Biodiversity data journal* **7**. [Accessed 2021 April 08]. Available from: https://www.ncbi.nlm.nih.gov/ pmc/articles/ PMC6420472/.

Hollander, J. L., & Vander Wall, S. B. (2009). Dispersal syndromes in North American *Ephedra*. *International Journal of Plant Sciences* **170**: 323–330. DOI: https://doi.org/10.1086/596334.

Holman, E. W. (1987). Recognizability of sexual and asexual species of rotifers. *Systematic Zoology* **36**: 381–386. DOI: https://doi.org/10.2307/2413402.

Hortal, J., de Bello, F., Diniz-Filho, J. A. F., Lewinsohn, T. M., Lobo, J. M., & Ladle, R. J. (2015). Seven shortfalls that beset large-scale knowledge of biodiversity. *Annual Review of Ecology, Evolution, and Systematics* **46**: 523–549. [Accessed 2017 October 27]. Available from: https://www. academia.edu/download/45132164/a6.pdf.

Hortal, J., Lobo, J. M., & Jiménez-Valverde, A. (2007). Limitations of biodiversity databases: case study on seed-plant diversity in Tenerife, Canary Islands. *Conservation Biology* **21**: 853–863. DOI: https://doi.org/10.1111/j.1523-1739.2007.00686.x.

Hox, J. J., Moerbeek, M., & Van de Schoot, R. (2017). *Multilevel analysis: Techniques and applications,* 3rd Ed. Routledge, New York. DOI: https://doi.org/10.4324/9781315650982.

Huang, J., Giannasi, D. E., & Price, R. A. (2005). Phylogenetic relationships in *Ephedra* (Ephedraceae) inferred from chloroplast and nuclear DNA sequences. *Molecular Phylogenetics and Evolution* **35**: 48–59. DOI: https://doi.org/10.1016/j.ympev. 2004.12.020.

Hunziker, J. H. (1949). *Sinopsis de las especies argentinas del género* "*Ephedra*". Ministerio de Agricultura y Ganaderia de la Nación, Dirección General de Investigaciones Agrícolas, Instituto de Botánica.

Ickert-Bond, S. M. (2003). Systematics of New World *Ephedra* L. (Ephedraceae): Integrating of morphological and molecular data. Ph.D. Dissertation. Tempe: Arizona State University.

Ickert-Bond, S. M. (2005). *Ephedraceae*. In: Harling G, Anderson L. eds: Flora of Ecuador. Göteborg University: Elanders Berlings. 75: 3–10.

Ickert-Bond, S. M., & Renner, S. S. (2016). The Gnetales: recent insights on their morphology, reproductive biology, chromosome numbers, biogeography, and divergence times. *Journal of Systematics and Evolution* **54**: 1–16. DOI: https://doi.org/10.1111/jse.12190.

Ickert-Bond, S. M., & Wojciechowski, M. F. (2004). Phylogenetic relationships in *Ephedra* (Gnetales): evidence from nuclear and chloroplast DNA sequence data. *Systematic Botany* **29**: 834–849. DOI: https://doi.org/10.1600/0363644042451143.

Ickert-Bond, S. M., Murray, D., Oliver, M. G., Berrios, H. K., & Webb, C. O. (2019). The *Claytonia arctica* complex in Alaska—Analyzing a Beringian taxonomic puzzle using taxonomic concepts. *Annals of the Missouri Botanical Garden* **104**: 478–494. DOI: https://doi.org/10.3417/2019491.

Isaac, N. J., Mallet, J., & Mace, G. M. (2004). Taxonomic inflation: its influence on macroecology and conservation. *Trends in Ecology & Evolution* **19**: 464–469. DOI: https://doi.org/10.1016/j.tree.2004.06.004.

Jansen, F., & Dengler, J. (2010). Plant names in vegetation databases – a neglected source of bias. *Journal of Vegetation Science* **21**: 1179–1186. DOI: https://doi.org/10.1111/j.1654-1103.2010.01209.x.

Jensen, S. (2019). *Ausländerstudium in Deutschland: die Attraktivität deutscher Hochschulen für ausländische Studierende.* Springer-Verlag.

Johnson, J. B., & Omland, K. S. (2004). Model selection in ecology and evolution. *Trends in ecology & evolution* **19**: 101–108. DOI: https://doi.org/10.1016/j.tree.2003.10.013.

Johnson, P. C. D. (2014). Extension of Nakagawa & Schielzeth's *R* (2) GLMM to random slopes models. *Methods in Ecology and Evolution* **5**: 944–946. DOI: https://doi.org/10.1111/2041-210X.12225.

Joppa, L. N., Roberts, D. L., & Pimm, S. L. (2010). How many species of flowering plants are there? *Proceedings of the Royal Society B: Biological Sciences* **278**: 554–559. DOI: https://doi.org/10.1098/rspb.2010.1004.

Kadmon, R., Farber, O., & Danin, A. (2004). Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. *Ecological Applications* **14**: 401–413. DOI: https://doi.org/10.1890/02–5364.

Karger, D. N., Conrad, O., Böhner, J., Kawohl, T., Kreft, H., Soria-Auza, R. W., Niklaus E. Zimmermann, N. E., Linder, H. P., & Kessler, M. (2017). Climatologies at high resolution for the earth's land surface areas. *Scientific data* **4**: 1–20. DOI: https://doi.org/10.1038/sdata.2017.122.

Kassambara, A. (2020). *rstatix*: Pipe-friendly framework for basic statistical tests. *R* package version 0.6.0. [Accessed 2021 April 19]. Available from: https://rpkgs.datanovia.com/rstatix/.

Kassambara, A. (2020a). Package *ggpubr*. *R* package version 0.1, 6. [Accessed 2021 April 19]. Available from: cran.microsoft.com.

Kier, G., Kreft, H., Lee, T. M., Jetz, W., Ibisch, P. L., Nowicki, C., Mutke, J. & Barthlott, W. (2009). A global assessment of endemism and species richness across island and mainland regions. *Proceedings of the National Academy of Sciences* **106**: 9322–9327. DOI: https://doi.org/10.1073/ pnas.0810306106.

Kim, S. (2015). *ppcor*: an *R* package for a fast calculation to semi-partial correlation coefficients. *Communications for Statistical Applications and Methods* **22**: 665. DOI: https://doi.org/10.5351%2FCSAM. 2015.22.6.665.

Koh, L. P., Dunn, R. R., Sodhi, N. S., Colwell, R. K., Proctor, H. C., & Smith, V. S. (2004). Species coextinctions and the biodiversity crisis. *Science* **305**: 1632–1634. DOI: https:// doi.org/10.1126/science.1101101.

Komsta, L., & Novomestky, L. (2015). *moments*, cumulants, skewness, kurtosis and related tests. *R* package version 0.14. [Accessed 2021 April 19]. Available from: http://cran.r-project.org/package=moments.

Kozhamzharova, L. S., Sarsenbaev, K. N., Bekbayeva, L. K., Misni, S., & Tuymebayeva, B. E. (2013). Genetic Signs of Interspecific Polymorphism of L. Species in Kazakhstan Flora. *World Applied Sciences Journal* **21**: 428–432. [Accessed 2017 October 10]. Available from: http:// citeseerx.ist.psu.edu/viewdoc/download?doi= 10.1.1.388.5974&rep=rep1&type=pdf.

Kutsch, E., & Hall, M. (2010). Deliberate ignorance in project risk management. *International Journal of Project Management* **28**: 245–255. DOI: https://doi.org/10.1016/j.ijproman.2009.05.003.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. (2017). *lmerTest* package: tests in linear mixed effects models. *Journal of Statistical Software* **82**: 1–26. DOI: https://doi.org/10.18637/jss.v082.i13.

Linnaeus, C. (1753). *Species Plantarum. Vol. 1*. London.

Loconte, H., & Stevenson, D. W. (1990). Cladistics of the Spermatophyta. *Brittonia* **42**: 197–211. DOI: https://doi.org/10.2307/2807216.

Loera, I., Ickert-Bond, S. M., & Sosa, V. (2015). Ecological consequences of contrasting dispersal syndromes in New World *Ephedra*: higher rates of niche evolution related to dispersal ability. *Ecography* **38**: 1187–1199. DOI: https://doi.org/ 10.1111/ecog.01264.

Loera, I., Ickert-Bond, S. M., & Sosa, V. (2017). Pleistocene refugia in the Chihuahuan Desert: the phylogeographic and demographic history of the gymnosperm *Ephedra compacta*. *Journal of Biogeography* **44**: 2706–2716. DOI: https://doi.org/ 10.1111/jbi.13064.

Lomolino M. (2004). *Conservation Biogeography*. In: M. V. Lomolino and L. R. Heaney (eds): Frontiers of Biogeography: new directions in the geography of nature. Sinauer Associates, Sunderland, MA, 293–296.

Long, J. A. (2017). Package *jtools*. [Accessed 2021 April 17]. Available from: cran.microsoft.com.

Lüdecke, D. (2021). Package *sjPlot*. [Accessed 2021 June 07]. Available from: mran.revolutionanalytics.com.

Lüdecke, D., Ben-Shachar, M. S., Patil, I., Waggoner, P., & Makowski, D. (2021). *performance*: An *R* package for assessment, comparison and testing of statistical models. *Journal of Open Source Software* **6**. [pdf]. DOI: https://doi.org/10.21105/joss.03139.

Lughadha, E. N., Govaerts, R., Belyaeva, I., Black, N., Lindon, H., Allkin, R., McGill, R. E. & Nicolson, N. (2016). Counting counts: revised estimates of numbers of accepted species of flowering plants, seed plants, vascular plants and land plants with a review of other recent estimates. *Phytotaxa* **272**: 82–88. DOI: https://doi.org/10.11646/phytotaxa.272.1.5.

McGill, B. (2015). Is it a fixed or random effect? [Accessed 2021 October 11]. Available from: https:// dynamicecology.wordpress.com/2015/11/04/is-it-a-fixed-or-random-effect/.

McGill, B. J., Dornelas, M., Gotelli, N. J., & Magurran, A. E. (2015). Fifteen forms of biodiversity trend in the Anthropocene. *Trends in Ecology and Evolution* **30**: 104–113. DOI: https://doi.org/10.1016/j.tree.2014. 11.006.

Meier, R., & Dikow, T. (2004). Significance of specimen databases from taxonomic revisions for estimating and mapping the global species diversity of invertebrates and repatriating reliable specimen data. *Conservation Biology* **18**: 478–488. DOI: https://doi.org/10.1111/j.1523-1739.2004.00233.x.

Mesibov R. (2013). A specialist's audit of aggregated occurrence records. *ZooKeys* **293**: 1–18. DOI: https://doi.org/10.3897%2Fzookeys.293.5111.

Meyer, C., Weigelt, P., & Kreft, H. (2016). Multidimensional biases, gaps and uncertainties in global plant occurrence information. *Ecology letters* **19**: 992–1006. DOI: https://doi.org/10.1111/ele.12624.

Meyer, J. (2021). Overdispersion in Count Models: Fit the Model to the Data, Don't Fit the Data to the Model. https://www.theanalysisfactor.com/overdispersion-in-count-models-fit-the-model-to-the-data-dont-fit-the-data-to-the-model/.

Mitchell, P. J., Monk, J., & Laurenson, L. (2017). Sensitivity of fine-scale species distribution models to locational uncertainty in occurrence data across multiple sample sizes. *Methods in Ecology and Evolution* **8**: 12–21. DOI: https://doi.org/10.1111/2041-210X.12645.

Mori, S. A. (2013). *Plant talk – The Shifting Science of Botanical Nomenclature, I and II*. [Accessed 2022 January 10]. Available from: https://www.nybg.org/blogs/plant-talk/tag/scott-mori/

Morton, S. R., & Hill, R. (2014). What is biodiversity, and why is it important? In: Morton SR, Sheppard AW & Lonsdale WM (eds): Biodiversity: science and solutions for Australia, CSIRO Publishing, Collingwood, Melbourne, 1–12.

Murphey, P. C., Guralnick, R. P., Glaubitz, R., Neufeld, D., & Ryan, J. A. (2004). Georeferencing of museum collections: a review of problems and automated tools, and the methodology developed by the Mountain Informatics Initiative (Mapstedi). *PhyloInformatics* **21**: 1–29. [Accessed 2021 November 21]. Available from: https://paleosolutions.com/publications/Georeferencing_ of_ Museum_Collections.pdf.

Myers, N., & Swanson, T. M. (1995). Tropical deforestation: population, poverty and biodiversity. *The economics and ecology of biodiversity decline: the forces driving global change*: 111–22. Cambridge University Press, Cambridge.

Myung, I. J. (2000). The importance of complexity in model selection [Special issue]. *Journal of Mathematical Psychology* **44**: 37. DOI: https://doi.org/10.1006/jmps.1999.1283.

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution* **4**: 133–142. DOI: https://doi.org/10.1111/j.2041-210x.2012.00261.x.

Neuwirth, E., & Brewer, R. C. (2014). The ColorBrewer palettes. *R* package version, 1. [Accessed 2021 January 20]. Available from: https://www.r-graph-gallery.com/38-rcolorbrewers-palettes.html.

Nicolson, D. H. (1991). A history of botanical nomenclature. *Annals of the Missouri Botanical Garden* **78**: 33–56. DOI: https://doi.org/10.2307/2399589.

Nicolson, N. (2019). Automating the construction of higher order data representations from heterogeneous biodiversity datasets. Dissertation. Brunel University London.

Nixon, C. G. (2016). *How Valuable is that Plant Species: Application of a Method for Enumerating the Contribution of Selected Plant Species to New Zealand's GDP*. Ministry for Primary Industries.

Otegui, J., Ariño, A. H., Encinas, M. A., & Pando, F. (2013). Assessing the primary data hosted by the Spanish node of the Global Biodiversity Information Facility (GBIF). *PloS One* **8**: e55144. DOI: https://doi.org/10.1371/journal.pone.0055144.

Pebesma, E., & Bivand, R. S. (2005). S classes and methods for spatial data: the *sp* package. *R news* **5**: 9–13. [Accessed 2021 January 19]. Available from: https://CRAN.R-project.org/doc/Rnews/.

Peinado, M., Macías, M. Á., Delgadillo, J., & Aguirre, J. L. (2006). Major plant communities of North America's most arid region: the San Felipe Desert, Baja California, Mexico. *Plant Biosystems* **140**: 280–296. DOI: https://doi.org/10.1080/11263500600947715.

Pillon, Y., & Chase, M. W. (2007). Taxonomic exaggeration and its effects on orchid conservation. *Conservation Biology* **21**: 263–265. DOI: https://doi.org/10.1111/j.1523-1739.2006.00573.x.

Pimm, S.L, & Joppa, L.N. (2015). How Many Plant Species are There, Where are They, and at What Rate are They Going Extinct? *Annals of the Missouri Botanical Garden* **100**: 170–176. DOI: https://doi.org/10.3417/2012018.

Powers, R. P., & Jetz, W. (2019). Global habitat loss and extinction risk of terrestrial vertebrates under future land-use-change scenarios. *Nature Climate Change* **9**: 323–329. DOI: https://doi.org/10.1038/s41558-019-0406-z.

Prance G. T. (1977). Floristic Inventory of the Tropics: Where do we stand? *Annals of the Missouri Botanical Garden* **64**: 659–684. DOI: https://doi.org/10.2307/2395293.

Pyke G. H., & Ehrlich P. (2010). Biological collections and ecological/environmental research: a review, some observations and a look to the future. *Biological reviews of the Cambridge Philosophical Society* **85**: 247–66. DOI: https://doi.org/10.1111/j.1469-185X.2009.00098.x.

Qiang Y., Weigelt, P., Fristoe, T. S., Zhang, Z., Kreft, H., Stein, A., Seebens, H., Dawson, W., Essl, F., König, C., Lenzner, B., Pergl, J., Pouteau, R., Pyšek, P., Winter, M., Ebel, A. L., Fuentes, N., Giehl, E. L., Kartesz, J., Krestov, P., Kukk, T., Nishino, M., Kupriyanov, A., Villaseñor, J. L., Wieringa, J. J., Zeddam, A., Zykova, E.,& van Kleunen, M. (2021). The global loss of floristic uniqueness. *Nature Communication* **12**: 1–10. DOI: https://doi.org/ 10.1038/s41467-021-27603-y

R Core Team. (2013). *R*: A language and environment for statistical computing. *R* Foundation for Statistical Computing, Vienna, Austria. Available from: http://www.R-project.org/

Raes, N., & Aguirre-Gutierrez, J. (2018). Modeling framework to estimate and project species distributions space and time. *Mountains, climate and biodiversity*: 309. John Wiley & Sons.

Rao, M. V. (2004). The importance of botanical nomenclature and synonymy in taxonomy and biodiversity. *Current Science* **87**: 602–606. [Accessed 2020 March 12]. Available from: https://www.jstor.org/stable/24109039.

Reddy, S., & Dávalos, L.M. (2003). Geographical sampling bias and its implications for conservation priorities in Africa. *Journal of Biogeography* **30**: 1719–1727. DOI: https://doi.org/10.1046/j.1365-2699.2003.00946.x.

Reydon, T. A., & Kunz, W. (2019). Species as natural entities, instrumental units and ranked taxa: new perspectives on the grouping and ranking problems. *Biological Journal of the Linnean Society* **126**: 623–636. DOI: https://doi.org/10.1093/biolinnean/blz013.

Rodriguez-Perez, J., Larrinaga, A. R., & Santamaria, L. (2012). Effects of frugivore preferences and habitat heterogeneity on seed rain: a multi-scale analysis. *PloS One* **7**: e33246. DOI: https://doi.org/10.1371/journal.pone.0033246.

Roy, D., Alderman, D., Anastasiu, P., Arianoutsou, M., Augustin, S., Bacher, S., Başnou, C., Beisel, J., Bertolino, S., Bonesi, L., Bretagnolle, F., Chapuis, J. L., Chauvel, B., Chiron, F., Clergeau, P., Cooper, J., Cunha, T., Delipetrou, P., Desprez-Loustau, M., Détaint, M., Devin, S., Didžiulis, V., Essl, F., Galil, B. S., Genovesi, P., Gherardi, F., Gollasch, S., Hejda, M., Hulme, P. E., Josefsson, M., Kark, S., Kauhala, K., Kenis, M., Klotz, S., Kobelt, M., Kühn, I., Lambdon, P. W., Larsson, T., Lopez-Vaamonde, C., Lorvelec, O., Marchante, H., Minchin, D., Nentwig, W., Occhipinti-Ambrogi, A., Olenin, S., Olenina, I., Ovcharenko, I., Panov, V. E., Pascal, M., Pergl, J., Perglová, I., Pino, J., Pyšek, P., Rabitsch, W., Rasplus, J., Rathod, B., Roques, A., Roy, H., Sauvard, D., Scalera, R., Shiganova, T. A., Shirley, S., Shwartz, A., Solarz, W., Vilà, M., Winter, M., Yésou, P., Zaiko, A., Adriaens, T., Desmet, P., & Reyserhove, (2020). *DAISIE - Inventory of alien invasive species in Europe. Version 1.7*. Research Institute for Nature and Forest (INBO). Checklist dataset. [Accessed 2022 March 15]. Available from: https:// doi.org/10.15468/ybwd3x, accessed via GBIF.org.

Ruhfel, B. R., Gitzendanner, M. A., Soltis, P. S., Soltis, D. E., & Burleigh, J. G. (2014). From algae to angiosperms–inferring the phylogeny of green plants (Viridiplantae) from 360 plastid genomes. *BMC Ecology and Evolution* **14**: 1–27. [Accessed 2020 March 12]. Available from: https://bmcecolevol. biomedcentral.com/track/pdf/10.1186/1471-2148-14-23.pdf.

Rydin, C. (2018). The Gnetales-a small window onto a lost world. *Svensk Botanisk Tidskrift* **112**: 4–21. [Accessed 2019 January 24]. Available from: https://www.cabdirect.org/cabdirect/abstract/20183299636.

Rydin, C., & Bolinder, K. (2015). Moonlight pollination in the gymnosperm *Ephedra* (Gnetales). *Biology Letters* **11**: 20140993. DOI: https://doi.org/10.1098/rsbl.2014.0993.

Schielzeth, H., Dingemanse, N. J., Nakagawa, S., Westneat, D. F., Allegue, H., Teplitsky, C., Réale, D., Dochtermann, N. A., Garamszegi, L.Z., & Araya-Ajoy, Y. G. (2020). Robustness of linear mixed-effects models to violations of distributional assumptions. *Methods in Ecology and Evolution.* **11**: 1141–1152. DOI: https://doi.org/10.1111/2041-210X.13434

Schloerke, B., Crowley, J., & Cook, D. (2018). Package '*GGally*'. *Extension to 'ggplot2.* [Accessed 2022 January 12]. Available from: cran.microsoft.com

Schmidt-Lebuhn, A. N., Knerr, N. J., & Kessler, M. (2013). Non-geographic collecting biases in herbarium specimens of Australian daisies (Asteraceae). *Biodiversity and Conservation* **22**: 905–919. DOI: https://doi.org/ 10.1007/s10531-013-0457-9.

Scholes, R. J., Montanarella, L., Brainich, E., Barger, N., Ten Brink, B., Cantele, M., Erasmus, B., Fisher, J., Gardner, J., Holland, T. G., Kohler, F., Kotiaho, S., von Maltitz, G., Nangendo, G., Pandit, R., Parrotta, J., Potts, M. D., Prince, S., Sankaran, M., & Willemen, L. (2018). *IPBES 2018: Summary for policymakers of the assessment report on land degradation and restoration of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services*. IPBES secretariat, Bonn, Germany.

Scott, W. A., & Hallam, C. J. (2002). Assessing species misidentification rates through quality assurance of vegetation monitoring. *Plant Ecology* **165**: 101–115. DOI: https://doi.org/10.1023/A:1021441331839.

Sing, T., Sander, O., Beerenwinkel, N., & Lengauer, T. (2015). Package *ROCR*. Visualizing the performance of scoring classifiers. *Bioinformatics*. DOI: https://doi.org/10.1093/bioinformatics/bti623.

Soberón, J., & Peterson, T. (2004). Biodiversity informatics: managing and applying primary biodiversity data. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* **359**: 689–698. DOI: https://doi.org/10.1098/rstb.2003.1439.

Sousa-Baena, M. S., Garcia, L. C. & Peterson, A. T. (2014). Completeness of digital accessible knowledge of the plants of Brazil and priorities for survey and inventory. *Diversity and Distribution* **20**: 369–381. DOI: https://doi.org/10.1111/ddi.12136.

Stapf, O. (1889). Die Arten der Gattung *Ephedra*. KK Hof-und Staatsdruckerei, in Commission bei F. Tempsky.

Steinbart, P. J., & Nath, R. (1992). Problems and issues in the management of international data communications networks: the experiences of American companies. *MIS quarterly*: 55–76. DOI: https://doi.org/10.2307/ 249701.

Sterner, B., & Franz, N. M. (2017). Taxonomy for humans or computers? Cognitive pragmatics for big data. *Biological Theory* **12**: 99–111. DOI: https://doi.org/10.1007/s13752-017-0259-5.

Stevens, P. F. (2016). *Angiosperm Phylogeny Website*. Version 13. [Accessed 2017 January 16]. Available from: http://www.mobot.org/MOBOT/research/APweb/.

Stevenson, D. W. (1993). *Flora of North America, volume 2: Ephedraceae*. 428–434. Flora of North America Editorial Committee [eds.]. Oxford University Press, New York, USA. [Accessed 2017 March 17]. Available from: http://www.efloras.org/florataxon.aspx?flora_id=1&taxon_id=10313.

Stoffel, M. A., Nakagawa, S., & Schielzeth, H. (2017). *rptR*: repeatability estimation and variance decomposition by generalized linear mixed-effects models. *Methods in Ecology and Evolution* **8**: 1639. DOI: https://doi.org/10.1111/2041-210X.12797

Stoffel, M. A., Nakagawa, S., & Schielzeth, H. (2021). *partR2*: Partitioning R2 in generalized linear mixed models. *Bioinformatics and Genomics* **9**: e11414. DOI: https://doi.org/10.7717/peerj.11414.

Stropp, J., Ladle, R. J., M. Malhado, A. C., Hortal, J., Gaffuri, J., H. Temperley, W., Skøien, J. O., & Mayaux, P. (2016). Mapping ignorance: 300 years of collecting flowering plants in Africa. *Global Ecology and Biogeography* **25**: 1085–1096. DOI: https://doi.org/10.1111/geb.12468.

TDWG. (2021). *Biodiversity Information Standards (TDWG) - SpatialPolygonsDataFrame*. [Accessed 2020 May 17]. Available from: www.tdwg.org/standards/109/tdwg_lv3.

Ter Steege, H., Haripersaud, P. P., Bánki, O. S., & Schieving, F. (2011). A model of botanical collectors' behavior in the field: never the same species twice. *American Journal of Botany* **98**: 31–37. DOI: https://doi.org/10.3732/ajb.1000215.

Tessarolo, G., Ladle, R., Rangel, T., & Hortal, J. (2017). Temporal degradation of data limits biodiversity research. *Ecology and Evolution* **7**: 6863–6870. DOI: https://doi.org/10.1002/ece3.3259.

Thiers, B. (2022) continuously updated. *Index Herbarium: a global directory of public herbaria and associated staff.* New York Botanical Garden's Virtual Herbarium. [Internet]. [Accessed 5 July 2022]. Available from: http://sweetgum.nybg.org/science/ih.

Thomson, S. A., Pyle, R. L., Ahyong, S. T., Alonso-Zarazaga, M., Ammirati, J., Araya, J. F., Ascher, J. S., Audisio, T. S., Azevedo-Santos, V. M., Bailly, N., J. Baker, W. J., Balke, M., Barclay, M. V. L., Barrett, R. L., Benine, R. C., Bickerstaff, J. R. M., Bouchard, P., Bour, R., Bourgoin, T., … & Zhou, H. Z. (2018). Taxonomy based on science is necessary for global conservation. *PLoS One* **16**: e2005075. DOI: https://doi.org/10.1371/ journal.pbio.2005075.

Thuiller, W., & Lafourcade, B. (2019). *biomod2* package, pseudo.abs function. [Accessed 2020 June 27]. Available from: https://rdrr.io/rforge/BIOMOD/man/pseudo.abs.html.

Thuiller, W., Georges, D., Engler, R., & Breiner, F. (2016). Package '*biomod2*'. Species distribution modeling within an ensemble forecasting framework. [Accessed 2020 June 27]. Available from: https://CRAN. *R-proje ct. org/packa ge= biomod2*.

Tlhaloganyang, B. P., & Sakia, R. M. (2020). Zero inflated Poisson distribution in equidispersed data with excessive zeros. *Research Journal of Mathematics and Statistics* **8**: 31–34. [Accessed 2022 February 26]. Available from: www.iscamaths.com.

Töpel, M., Zizka, A., Calió, M. F., Scharn, R., Silvestro, D., & Antonelli, A. (2017). *SpeciesGeoCoder*: fast categorization of species occurrences for analyses of biodiversity, biogeography, ecology, and evolution. *Systematic Biology* **66**: 145–151. DOI: https://doi.org/10.1093/ sysbio/syw064.

Turland, N. J., Wiersema, J. H., Barrie, F. R., Greuter, W., Hawksworth, D. L., Herendeen, P. S., Knapp, S., Kusber, W. H., Li, D. Z., Marhold, K., May, T.W., McNeill, J., Monro, A. M., Prado, J., Price, M. J., & Smith, G. (2018). *International Code of Nomenclature for algae, fungi, and plants (Shenzhen Code) adopted by the Nineteenth International Botanical Congress Shenzhen, China, July 2017*. Koeltz Botanical Books, 2018. [Accessed 2021 October 03]. Available from: https://www.iaptglobal.org/the-code-pdf.

Valdecasas, A. G., Williams, D., & Wheeler, Q. D. (2008). Integrative taxonomy then and now: a response to Dayrat (2005). *Biological Journal of the Linnean Society* **93**: 211–216. DOI: https://doi.org/10.1111/j.1095-8312.2007.00919.x.

Vetter, M. 1990. *Aufbau betrieblicher Informationssysteme mittels konzeptioneller Datenmodellierung*. 5. Auflage. Springer Verlag, Wiesbaden.

WCSP. (2020). *World Checklist of Selected Plant Families*. Facilitated by the Royal Botanic Gardens, Kew. [Accessed 2017 August 7]. Available from: http://wcsp.science. kew.org/.

WCSP. (2020a). *World Checklist of Selected Plant Families*. Facilitated by the Royal Botanic Gardens, Kew. [Accessed 2017 August 7]. Available from: http://wcsp.science. kew.org/.

WCSP. 2018. World Checklist of Selected Plant Families. Facilitated by the Royal Botanic Gardens, Kew. [Accessed 2017 August 7]. Available from: http://wcsp.science. kew.org/.

Wearn, J. A., Chase, M. W., Mabberley, D. J., & Couch, C. (2013). Utilizing a phylogenetic plant classification for systematic arrangements in botanic gardens and herbaria. *Botanical Journal of the Linnean Society* **172**: 127–141. DOI: https://doi.org/10.1111/ boj.12031.

Wei, T., & Simko, V. (2021). *R* Package '*corrplot*': Visualization of a correlation matrix. (Version 0.90). *Statistician* **56**. [Accessed 2021 May 22]. Available from: https://github.com/taiyun/corrplot.

Wells, A., Johanson, K. A., & Dostine, P. (2019). Why are so many species based on a single specimen? *Zoosymposia* **14**: 32–38. DOI: https://doi.org/10.11646/zoosymposia.14.1.5.

White, L. M., Gardner, S. F., Gurley, B. J., Marx, M. A., Wang, P. L., & Estes, M. (1997). Pharmacokinetics and cardiovascular effects of ma-huang (*Ephedra sinica*) in normotensive adults. *Journal of Clinical Pharmacology* **37**: 116–122. DOI: https://doi.org/10.1002/j.1552-4604.1997.tb04769.x.

Whittaker, R. J., Araújo, M. B., Jepson, P., Ladle, R. J., Watson, J. E., & Willis, K. J. (2005). Conservation biogeography: assessment and prospect. *Diversity and Distribution* **11**: 3–23. DOI: https://doi.org/10.1111/ j.1366-9516.2005.00143.x.

Wickett, N. J., Mirarab, S., Nguyen, N., Warnow, T., Carpenter, E., Matasci, N., Ayyampalayam, S., Barker, M. S., Burleigh, J. G., Gitzendanner, M. A., Ruhfel, B. R., Wafula, E., Der, J. P., Graham, S. W., Mathews, S., Melkonian, M., Soltis, D. E., Soltis, P. S., Miles, N. W., Rothfels, C. J., Pokorny, L., Shaw, A. J., DeGironimo, L., Stevenson, D. W., Surek, B., Villarreal, J. C., Roure, B., Philippe, H., dePamphilis, C. W., Chen, T., Deyholos, M. K., Baucom, R. S., Kutchan, T. M., Augustin, M. M., Wang, J., Zhang, Y., Tian, Z., Yan, Z., Wu, X., Sun, X., Wong, G. K., Wickett, N., & Leebens-Mack, J. (2014). Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proceedings of the National Academy of Sciences* **111**: E4859–E4868. DOI: https://doi.org/10.1073/pnas. 1323926111.

Wickham, H. (2016). ggplot2: *Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D. A., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson5, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., & Yutani, H. (2019). Welcome to the Tidyverse. *Journal of open source software* **4**: 1686. DOI: https://doi.org /10.21105/joss.01686.

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**: e29715. DOI: https://doi.org/10.1371/journal.pone.0029715.

Wilson, E. O. (2004). Taxonomy as a fundamental discipline. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* **359**, 739–739. DOI: https://doi.org/10.1098/rstb.2003.1440.

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*. Vol. 1: 29–40. [Accessed 2017 October 10]. Available from: http://www.cs.unibo.it/~danilo.montesi/CBD/ Beatriz/10.1.1. 198.5133.pdf.

Wortley, A. H., & Scotland, R. W. (2004). Synonymy, sampling and seed plant numbers. *Taxon* **53**: 478–480. DOI: https://doi.org/10.2307/4135625.

Yesson, C., Brewer, P. W., Sutton, T., Caithness, N., Pahwa, J. S., Burgess, M., Gray, W. A., White, R. J., Jones, A. C., Bisby, F. A., & Culham, A. (2007). How global is the global biodiversity information facility? *PLoS One* **2**: e1124 DOI: https://doi: 10.1371/journal.pone.0001124.

Zermoglio, P. F., Guralnick, R. P., & Wieczorek, J. R. (2016). A standardized reference data set for vertebrate taxon name resolution. *PLoS One* **11**: e0146894. DOI: https://doi.org/10.1371/journal.pone.0146894.

Zhang, Y., Schaap, M. G., & Zha, Y. (2018). A high-resolution global map of soil hydraulic properties produced by a hierarchical parameterization of a physically based water retention model. *Water Resources Research* **54**: 9774–9790. DOI: https://doi.org/10.1029/2018WR023539.

Zhu, Y. P. (1998). Chinese materia medica: chemistry, pharmacology and applications. CRC press.

Zizka, A. (2019). Cleaning GBIF data for the use in biogeography (Tutorial). [Accessed 2020 November 4]. Available from: https://ropensci.github.io/CoordinateCleaner/articles/Cleaning_GBIF_data_with_CoordinateCleaner.html.

Zizka, A., Carvalho, F. A., Calvente, A., Baez-Lizarazo, M. R., Cabral, A., Coelho, J. F. R., Colli-Silva, M., Ramos Fantinati, M., Fernandes, M. F., Ferreira-Araújo, T., Lambert Moreira, F. G., Cunha Santos, N. M., Borges Santos, T. A., dos Santos-Costa, R. C., Serrano, F. C., Alves da Silva, A. P., de Souza Soares, A., Cavalcante de Souza, P. G., Tomaz, E. C., Fonseca Vale, V., Vieira, T. L., & Antonelli, A. (2020). No one-size-fits-all solution to clean GBIF. *Biodiversity and Conservation* **8**, e9916 DOI: https://doi.org/10.7717/peerj.9916.

Zizka, A., Silvestro, D., Andermann, T., Azevedo, J., Duarte Ritter, C., 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 in Ecology and Evolution* **10**: 744–751. DOI: https://doi.org/10.1111/2041-210X. 13152.

Zumajo-Cardona, C., & Ambrose, B. A. (2021). Deciphering the evolution of the ovule genetic network through expression analyses in Gnetum gnemon. *Annals of Botany* **128**: 217–230. DOI: https://doi.org/10.1093/aob/mcab059.

Acknowlegments

Many people have helped and supported me during this thesis; this challenge would have been difficult to cope with without them. It is a pleasure to acknowledge their help and various contributions.

First of all, I want to thank my supervisors, Prof. Dr. Holger Kreft and Prof. Dr. Stefanie Ickert-Bond, for giving me the opportunity to conduct this Ph.D. project. Thanks for your encouragement, support, repeated critical reviewing, and many inspiring ideas, for stimulating discussions, and for creating such a friendly and motivating work environment. Also, thanks for supporting me throughout the design, implementation, writing, and publishing. Your expertise was essential for the conceptual design of the research, and your broad taxonomic and biogeographic knowledge was invaluable for its realization.

Special thanks are extended to Dr. Patrick Weigelt, colleague and co-author of chapter 3, for your friendly discussions whenever needed, patience with all my questions, and professional advice. Also, thanks to Prof. Dr. Kerstin Wiegand, who kindly agreed to be a member of my thesis committee, I am grateful for the discussions about my projects.

I am also grateful for the many valuable comments by Pedro Tarroso, and other anonymous referees, which significantly improved the research chapters of this thesis. I also acknowledge all authors whose studies were included in the reviews and those who kindly provided additional information.

I want to thank all my lab-mates in the Macroecology, Biodiversity, and Conservation Biogeography group for lively lab meetings and tea/coffee-break discussions, valuable suggestions, constant helpfulness, exchange of *R* codes and statistical advice, shared lunch, tea, and coffee breaks, and generally for making research life enjoyable. Also, the support from the University of Göttingen is greatly acknowledged.

Finally, thanks to my family and friends, particularly my husband Ralph, for continuous support and encouragement, your kind interest in my work, progress, ups and downs, and many motivational messages and moral support.