What remains to be discovered: A global assessment of tree inventory completeness

Jariya Chanachai

School of Natural Sciences

Faculty of Science and Engineering

Macquarie University, Sydney, NSW 2109

September 9, 2022

A thesis submitted in partial fulfilment of the requirements for the degree of Master of Research

DECLARATION

I declare that this thesis, as a whole or in parts, has not been submitted for a higher degree to any other university or institution. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made in the thesis itself.

I wish to acknowledge the following assistance with the research detailed in this thesis:

Linda Beaumont, Maina Mbui and Ernest Asamoah conceived and designed this study and critically commented on earlier versions of this manuscript. Ernest Asamoah and Peter Wilson provided assistance in using R. Peter Wilson developed R script for removing non-native species records. Ernest Asamoah developed R script for data cleaning and bivariate mapping, provided data for forest landscape integrity and assisted in the analysis. David Nipperess guided the experimental design and provided expert opinions. I undertook all analyses and wrote the manuscript.

No ethics approval was required for this project.

09 September 2022

Jariya/Chanachai Date

This thesis is formatted as a manuscript for submission to the ***Journal of Diversity and Distributions***, with some exceptions to meet the requirements of the Macquarie University. This includes the requirement of an abstract of 200 words, 2cm margins, 1.5x line spacing, figures and tables embedded within the text.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

I would like to acknowledge the valuable support provided by my primary supervisor, Associate Professor Linda Beaumont and my associate supervisor, Dr Maina Mbui, for guiding this study and reviewing previous drafts of this manuscript. Also, I express my gratitude to Dr Ernest Asamoah for assisting with the analysis, providing the forest landscape integrity data, and commenting on early versions of this paper. Thanks to Dr Peter Wilson and Dr David Nipperess for providing expert opinions on the research topic, and to Dr Manuel Esperon-Rodriguez and Mr Sourav (Rony) Das for earlier conversations about coding. This project was funded by the Macquarie University International Research Training Program (Master of Research) scholarship of Macquarie University, Sydney, Australia.

Candidate's statement about the impact of COVID-19 changes on the thesis

Dear Examiner,

Many of our HDR candidates have had to make changes to their research due to the impact of COVID-19. Below you will find a statement from the candidate, approved by their Supervisory Panel, that indicates how their original research plan has been affected by COVID-19 restrictions. Relevant ongoing restrictions in place caused by COVID-19 will also be detailed by the candidate.

Candidate's Statement

Since the start of the project, I experienced difficulties with accessing a sufficiently powerful computer for my research and analysis (data-intensive in nature). Between July and September when I was required to work from home due to COVID-19 lockdown restrictions, I was unable to work effectively and experienced a considerable loss in productivity due to unsuitable working space, poor computing facilities, and limited internet access and connectivity. Thus, I decided to take a COVID-19 leave of absence and delayed data access and analysis. Due to COVID-19 interruptions, I was unable to complete my thesis plan, and thereby excluded the development of multi-variate models to predict tree species richness in areas with low inventory completeness.

Abstract

Aim: Recent unprecedented efforts to digitise and mobilise biodiversity data have resulted in the generation of "biodiversity big data", enabling ecological research at scales previously not possible. Here, we evaluated the completeness of digitised tree records globally, and identified where future surveys should focus to maximise regional inventories.

Location: Global.

Methods: We analysed spatial patterns in tree records from the Global Biodiversity and Information Facility and assessed global tree inventory completeness at a 100 km resolution and the ecoregional scale. We also identified priority areas for future exploration by examining the spatial covariation of completeness and natural habitat and forest conditions.

Results: Spatial patterns of sampling effort and tree inventory completeness are unevenly distributed around the world, with most well-known sites being concentrated in the Global North, whereas large areas in species-rich tropical regions remain poorly documented. Moreover, many sites with low inventory completeness coincided with areas of rapid natural habitat loss and low forest integrity.

Main Conclusions: Digitised biodiversity data has great potential to help address ecology questions and inform conservation actions if their biases and uncertainties are understood. Here we illustrated how such data can be used to improve existing knowledge and identify priority areas for future surveys.

Keywords

Chao1 estimator, data bias, Global Biodiversity Information Facility, global tree inventories, inventory completeness, priority for survey and conservation, Wallacean shortfall.

Introduction

One of the central objectives of ecology is to understand the diversity and composition of species assemblages over space and time by investigating the relationship between species and their environment (Devictor & Bensaude-Vincent, 2016). Thus, the recording of information about a species at a specific place and time is perhaps the most foundational practice in ecology. Since the 19th century, when biological exploration began in earnest, billions of biological samples have been collected and stored in natural history museums and herbaria (hereafter, "museums") across the world (Brummitt et al., 2020). Museums harbour a vast collection of biological specimens, ranging from frozen DNA and microbes to dried plants pressed, butterfly collections and whale skeletons. These institutions serve as invaluable resources for interdisciplinary scientific research and education (Bakker et al., 2020; Pyke & Ehrlich, 2010). However, natural history collections are largely located in developed parts of the world, making their access limited to only a small number of people (e.g., researchers, botanists and scientists) with the means to travel to the institution where the specimen is housed (Brummitt et al., 2020; Edwards, 2004; Edwards et al., 2000).

In the last 20 years, unprecedented international efforts to digitise, store and mobilise biodiversity data, including natural history collections, on online databases have proliferated, resulting in the generation of "biodiversity big data" which has introduced new concepts of global biodiversity, such as biomes, ecoregions and hotspots of diversity, to science (Bisby, 2000; Devictor & Bensaude-Vincent, 2016). Many museums now store information from their specimen collections on electronic databases, such as the Natural History Museum in London (https://www.nhm.ac.uk/), the National Museum of Natural History of the Smithsonian Institution in Washington D.C. (https://naturalhistory.si.edu/research), and the Muséum National d'Histoire Naturelle in Paris (https://www.mnhn.fr/en). Online databases have facilitated access to large quantities of digitised biodiversity data, mainly in the form of occurrence records (i.e., an observation of a species at a particular place and time), enabling the study of biodiversity at taxonomic, spatial and temporal scales previously not possible (Heberling et al., 2021).

The Global Biodiversity Information Facility (GBIF; https://www.gbif.org/) is the world's largest biodiversity data portal. This intergovernmental research and data infrastructure provides anyone worldwide open access to biodiversity data via the Internet. Initially aimed at digitising natural history collections into a single portal (Edwards et al., 2000), GBIF now aggregates biodiversity information from various data sources (Devictor & Bensaude-Vincent, 2016; Heberling et al., 2021; Telenius, 2011) and provides more than 2.2 billion records from over 1,890 institutions (accessed on 4th September 2022). Potential applications of digitised biodiversity data in ecology

are manifold, ranging from estimates of species richness and patterns of diversity (Ballesteros-Mejia et al., 2013; Brummitt et al., 2020; Pelayo-Villamil et al., 2018), to extrapolation of species' potential geographic distributions (Cunze et al., 2020; Elith & Leathwick, 2009; Franklin & Miller, 2009), bio-invasion studies (Beaumont et al., 2014; Cao et al., 2021; O'Donnell et al., 2012), and biodiversity assessments (Guisan et al., 2013; Salinas-Rodríguez et al., 2018).

However, as more digitised biodiversity data become available, concerns remain over data quality because gaps, uncertainties, inaccuracies, and biases are known to exist (Maldonado et al., 2015; Meyer et al., 2016; Soberón & Peterson, 2004). Thus, the question arises: are these data fit for purpose? For scientific research to be reliable and valid, gaps and biases in the underlying data must be understood and addressed.

Gaps and Biases in Digitised Biodiversity Data and Their Impact on Scientific Studies

Although the mobilisation of digitised biodiversity data has advanced research in ecology, and the availability of over two billion records (e.g., through GBIF) may improve understanding of global biodiversity, various geographical regions remain data deficient and many taxonomic groups are under-represented (Meyer et al., 2015; Yesson et al., 2007). Generally, errors and biases in digitised biodiversity data can be classified into three dimensions: temporal, spatial and taxonomic – a set of problems described as "biodiversity knowledge shortfalls" (Hortal et al., 2015; Meyer et al., 2016). Biases in data coverage (e.g., how extensive, continuous and well documented is the existing species in different assemblages?) and information uncertainties (e.g., how precise, accurate and reliable are the existing data?) across any or all three dimensions can lead to biased ecological inferences and inefficient conservation (Meyer et al., 2016).

For instance, if species information is biased towards older records, it is uncertain whether species collected decades ago are still present at those locations (Girardello et al., 2019; Stropp et al., 2016). Spatial biases in the documentation of the species' distribution range can result in some areas being under-sampled, while imprecise or incorrect recording of their locations can lead to incomplete knowledge about the species' distribution and habitat (Rocchini et al., 2011). Spatio-temporal uncertainties and biases towards areas of particular interest can result in incomplete knowledge about the species' distribution, a phenomenon known as the "Wallacean shortfall", which can cause many observed biodiversity maps to resemble maps of sampling effort (Hortal et al., 2015; Lomolino, 2004). Additionally, species misidentification or ambiguous scientific names can lead to uncertainties in their taxonomic identification (Meyer et al., 2016), while biases in taxonomic coverage (i.e., preference for the collection of particular taxa over another) can produce

an under or over-estimation of species diversity, which can result in the discrepancy between the number of described species and the actual species richness, a situation referred to as the "Linnean shortfall" (Lomolino, 2004). While data that is imperfect may be preferable to none at all, the value of information can be enhanced if the uncertainties associated with the data are understood.

The effectiveness of conservation actions will depend on our knowledge and understanding of diversity and distribution patterns, which are typically derived from taxonomic inventories (Mora et al., 2008). Biases and uncertainties in species occurrence information and incompleteness in taxonomic inventories will influence analytical outcomes, resulting in distorted views of true diversity patterns: this can create concerns over the reliability of ecological studies and the effectiveness of conservation strategies (Hughes et al., 2021). If biodiversity data for a certain taxonomic group is limited to only a few species and locations on Earth, it would be difficult to precisely map any of the taxa at large scales, much less for global biodiversity. Thus, quantifying biases and uncertainties in digitised biodiversity data is crucial for the validity of biodiversity studies and for prioritising future sampling efforts to improve our current knowledge.

Global Tree Diversity

Trees provide the basic structure for some of the most diverse ecosystems on land such as forests and woodlands, and play a crucial role in supporting a wealth of other species due to their position at the base of the trophic level (Rivers, 2017). They also regulate the ecosystem through many ecological processes including carbon sequestration, soil stabilisation, and nutrient cycling (Rivers, 2017). Tree species have been used to develop a model system to study patterns in the variation of species richness across broad geographical scales, and correlations between species and their environment, together with the effect of evolutionary history on community assembly (Qian et al., 2013).

However, there is no consensus on how many tree species there are currently on Earth. Estimates have ranged from 45,000 to 100,000 species (Beech et al., 2017). Qian et al. (2019) estimated there are ~61,000 tree species globally, while a more recent study estimated ~73,000 (Gatti et al., 2022). Other reports have suggested that there are 21,000 species in temperate regions (Hunt, 1996) and 40,000–53,000 in the tropics (Slik et al., 2015). In 2017, the world's first list of global tree species and their country-level distribution called 'GlobalTreeSearch' (hereafter, "GTS") was published (Beech et al., 2017). According to the Botanic Gardens Conservation International (BGCI, 2021b), globally 58,497 tree species are currently known to science and

published in the GTS database, 30% of which have been identified as threatened, with forest clearance and habitat loss being the greatest threats to tree species.

Currently, forests cover about 30% (4.06 billion hectares) of the global land surface, though unevenly distributed across the world with more than 50% of this area occurring in the United States of America (USA), Canada, Brazil, China and the Russian Federation (FAO & UNEP, 2020). Between 1990 and 2020, there was a net loss of 178 million hectares of forest and the average rate of net forest loss has declined by roughly 40% between 1990–2000 and 2010–2022 (FAO & UNEP, 2020). However, the rate of forest loss/gain is unevenly distributed. For example, for the period 2000–2012, biodiversity-rich tropical forests had the highest ratio of loss to gain of any forested biome, with 32% of global forest cover loss occurring in tropical rainforests, nearly half of which occurred in South American rainforests (Hansen et al., 2013).

Though the rate of net loss of the world's forests has decreased, especially in the last decade, over 1,400 tree species have been assessed as critically endangered, 21% of the currently known tree species are either data deficient or have not been evaluated, while many more species may remain to be discovered (FAO & UNEP, 2020). Ongoing threats to tree species will adversely impact ecosystem functions and services as well as tree-dependent species. Thus, effective conservation actions are crucial for recovering endangered tree populations and maintaining their associated ecological process. Moreover, tree conservation has been recognised as a key action that will help address climate change (Poorter et al., 2015) and at the 2021 United Nations Climate Change Conference, COP 26, over 100 world leaders pledged to halt and reverse forest loss by 2030 (UNCC, 2021). Additionally, the United Nations Environmental Programme, in partnership with the Food and Agriculture Organisation, have established this decade as the UN Decade on Ecosystem Restoration with the aim to halt and reverse the degradation of ecosystems globally (https://www.decadeonrestoration.org/). Area-based conservation measures targeting 30% of land and sea areas by 2030 and the "Nature Needs Half" approach to biodiversity protection have been proposed by the Convention on Biological Diversity and some conservationists (Dinerstein et al., 2017; Wilson, 2016), and forest protection and restoration have also been recognised as one of the nature-based solutions that can help achieve these goals.

While tree species are being threatened by habitat loss and land clearance, knowledge gaps in global tree diversity and distributions still exist. Bridging these gaps is crucial for effective conversation, and the availability of biodiversity big data enables us to assess the current state of knowledge of tree diversity and distribution at a global scale. One study has evaluated the quality and coverage of global tree species using five major online databases, including GBIF (Serra-Diaz et al., 2018). They found considerable spatial coverage of tree occurrence data in Australia, Europe

and North America, while many biodiverse regions such as southeast Asia, and the Amazon lacked open-access biodiversity data. However, to date, there has not been a comprehensive study on the completeness of global tree inventories which can help us identify knowledge gaps and possible biases in the current description of tree diversity that can be influenced by uneven sampling efforts and taxonomic biases (Sousa-Baena et al., 2014). This study aims to utilise tree inventory big data available through GBIF to (a) assess the quality of digitised tree data and its limitations, (b) estimate tree inventory completeness throughout the world and (c) identify priority areas for future botanical survey efforts to improve existing knowledge of tree diversity and to provide recommendations for conservation.

Methods

Dataset

We obtained a list of 58,496 tree species from GTS (BGCI, 2021a; accessed on 29 October 2021), the world's most comprehensive global checklist of tree species (Beech et al., 2017). GTS uses the tree definition agreed on by the IUCN's Global Tree Specialist Group (GTSG): "a woody plant with usually a single stem growing to a height of at least two metres, or if multi-stemmed, then at least one vertical stem five centimetres in diameter at breast height". The validity of tree species names was assessed and standardised using the Taxonomic Name Resolution Service online tool (Boyle et al., 2013). Then, using the 'rgbif' package version 3.6.0 (Chamberlain et al., 2022) in R (R Core Team, 2021), we downloaded tree occurrence records from the GBIF database (GBIF.org, 02 March 2022; DOI:10.15468/dl.aptyh5) that had geographic coordinates, contained no documented geospatial issues, had taxonomic status as either 'accepted' or 'synonym', and had the basis of record as *'preserved_specimen'*, *'living_specimen'*, *'human_observation'*, *'observation'*, *'machine_observation'* or *'occurrence'*. This resulted in 42,431,811 records for 52,065 species from 3,796 datasets.

The downloaded records (hereafter, "raw data") were then filtered using the following multi-step procedure: (i) we removed records with occurrence status as "absent"; (ii) used the clean_coordinates function in the 'CoordinateCleaner' package (Zizka et al., 2019) with default options to filter and remove potentially erroneous coordinates, which included those assigned to country centroids, GBIF headquarters or biodiversity institutions, had equal longitude and latitude, fell into the ocean or had coordinates containing only zeros; (iii) excluded records with no decimals in longitude or latitude; (iv) selected records that were identified to species level or lower and were not representing hybrid/cultivated species; (v) removed duplicated records (defined as two or more records with the same combination of species name, collection date and location).

In addition to the initial data cleaning, we interrogated checklists of native administrative boundary codes (three-letter code for "Botanical Country" Level 3 of the Biodiversity Information Standards, formerly known as the Taxonomic Databases Working Group (TDWG); www.tdwg.org/standards/wgsrpd/) for tree species from Kew Garden's Plants of the World Online database (POWO, 2022) using the 'taxize' package version 0.9.98 (Chamberlain et al., 2020) in R (R Core Team, 2021). Then we validated the remaining occurrence records against the acquired checklists and eliminated species records which fell outside their native "Botanical Country" boundaries.

The above filtering processes yielded 26,423,277 records (hereafter, “cleaned data”), representing 50,290 species from 284 plant families. Out of the cleaned data, 4,405,806 (16.6%) records lacked a full collection date (day, month, year). For our study, all cleaned data were analysed.

Data Analysis

We calculated global tree inventory completeness at two scales: (i) 100 km x 100 km grid cells (18,115 sampling units; hereafter, “SUs”) and (ii) ecoregional level (Dinerstein et al., 2017). While many studies have analysed biodiversity inventory completeness at various grid cell resolutions (Ballesteros-Mejia et al., 2013; Haque et al., 2018; Meyer et al., 2016; Stropp et al., 2016), we considered the ecoregional scale in our analysis as ecoregions are biogeographic units representing distinct assemblages of biodiversity in particular regions. They also provide a useful base map for conservation planning because they draw on natural boundaries, rather than political ones (Olson et al., 2001). We obtained the ecoregion shapefile from RESOLVE Ecoregions 2017 which consisted of 846 terrestrial ecoregions, grouped into 14 biomes and 8 realms (Dinerstein et al., 2017). The occurrence records, SU and ecoregion shapefiles were projected from latitude and longitude (WGS 1984) to Mollweide (equal-area) projection.

Distribution of Tree Occurrence Records

To determine the trend and spatial patterns in data collection, we calculated the number of records per species, spatially intersected the occurrence records with both the SUs and ecoregion layers and calculated the number of occurrence records in each SU/ecoregion. We analysed the temporal pattern of data accumulation by totalling the number of records sampled in each year. We also calculated sampling effort as the number of records per km^2 and plotted maps to visualise the spatial distribution of records (see Appendix S1). We excluded Antarctic and the “rock and ice” ecoregion, leaving 845 ecoregions and 13,259 SUs for analysis.

Inventory Completeness

Inventory completeness, defined as the ratio of observed to the estimated number of species in a given area, is a useful way to determine how complete and representative sampling has been

(Soberón et al., 2007). We calculated the inventory completeness index (*C*) at the SU and ecoregional scales as follows:

$$C_i = {S_{obs(i)}}/{S_{est(i)}} \quad \text{(equation 1)}$$

where $S_{obs(i)}$ is the observed species richness and $S_{est(i)}$ is the estimated species richness for SU/ecoregion *i*, with the value for *C* ranging from zero to one. High *C* values suggest high degrees of completeness, while low *C* suggests low inventories. This information can be used to describe the state of current knowledge about species richness in a particular region and identify areas where floristic knowledge is lacking (Soberón et al., 2007; Sousa-Baena et al., 2014).

We used the *Chao1* non-parametric estimator to calculate expected species richness, $S_{est(i)}$. *Chao1* estimates the true species richness likely to be present in a region based upon the number of rare species observed in a sample. It can be calculated as:

$$S_{est(i)} = S_{obs(i)} + \frac{f_1^2}{2f_2} \quad \text{(equation 2)}$$

where f_1 and f_2 refer to rare species represented only once (singleton) and twice (doubleton) in a sample, respectively. The *Chao1* estimator is based on the proposition that additional species are less likely to be found when all the species in the samples are represented by at least two individuals (Gotelli & Colwell, 2011). Species richness was estimated using the ChaoRichness function in the 'iNext' package (Hsieh et al., 2016). The output yielded observed and estimated species richness, estimated standard error and the 95% confidence interval for each SU and ecoregion. Inventory completeness was calculated using equation 1, and the number of unrecorded species was calculated as the difference between estimated and observed richness. Moreover, for inventory completeness at the ecoregional scale, we calculated the mean *C* value for 14 terrestrial biomes. In SUs and ecoregions where the number of records is low, the estimates may be unstable and yield false/artifactual *C* values which are unreliable (Sousa-Baena et al., 2014). Thus, for both spatial scales, we used a minimum *C* value and the sample size distribution as criteria to define "well-known" sites (see Appendix S2).

Priority Areas

To identify priority areas for future sampling efforts, we examined the spatial covariation of *C* and the remaining natural habitat and protected area (for ecoregions) and forest integrity (for SUs). For ecoregions, we mapped completeness against four Nature Needs Half (NNH) categories. The NNH framework explores ecoregional options to protect 50% of land to achieve

comprehensive biodiversity conservation and maintain ecosystem functions for human benefits (Dinerstein et al., 2017). To achieve this, Dinerstein et al. (2017) estimated how much of the natural habitat had remained and how much land is protected in each ecoregion. Based on the concept of NNH, they sorted ecoregions into four categories, which include: (1) “Half Protected” (more than 50% of the ecoregion area is protected), (2) “Nature Could Reach Half” (although the amount of protected area is less than 50%, the total amount of protected land and unprotected natural habitat remaining is more than 50%), (3) “Nature Could Recover” (the total area of remaining natural habitat and protected land is less than 50% but more than 20%) and (4) “Nature Imperiled” (the total area of remaining natural habitat and protected land is less than 20%). Thus, we overlaid a map of tree inventory completeness onto NNH data for 462 ecoregions located in forested biomes (see Appendix S3 for names of forested biomes).

At the SU scale, we mapped completeness against the Forest Landscape Integrity Index (FLII) (Grantham et al., 2020) which describes the degree of forest modification by anthropogenic activities for 2019. FLII was calculated based on observed and inferred human pressures, and loss of forest connectivity, producing an index ranging from zero to ten, with ten representing the highest forest integrity (least impacted by human activities), and zero the lowest integrity (see detailed method in Grantham et al. (2020)). Using the original classifications of FLII, we calculated the proportion of forests having medium to low integrity in each SU (i.e., FLII <9.6, Grantham et al., 2020). We excluded SUs where the amount of forest area present within the grid cell is less than 10%, leaving 7,168 SUs for analysis. SUs and ecoregions with decreasing sampling opportunities have low *C* and high losses of natural habitat, while areas of future sampling opportunities have low *C* but still retain much of the native vegetation and forest integrity.

Results

Trends in Data Collection

Our data filtering led to the removal of 16,008,534 (37.7%) records from the raw data and 1,775 (3.4 %) species from the initial taxon names retrieved from GBIF. The initial data cleaning process (as described in the multi-step procedure) excluded 11,088,234 (26.1%) records, most of which were either duplicates or records with erroneous coordinates. The removal of non-native species occurrences led to an additional exclusion of 4,920,300 (11.6%) records.

Most of the cleaned data (78.9%, N = 20,841,932) were obtained from human observations, with only 18% (N = 4,701,716) being preserved specimens. The number of records averaged 525 per species (± 10,946 SD). Generally, most species have a low number of records (Fig. 1). About half (53.2%) of the tree species had ≥ 20 records, and 35.1% had ≥ 50 records. Above this threshold (50 records), the number of species declines sharply with the increasing number of records (Fig. 1). Moreover, only 1,860 (3.7%) species had ≥ 1,000 records, yet these species contributed 85.7% of the records to the dataset.

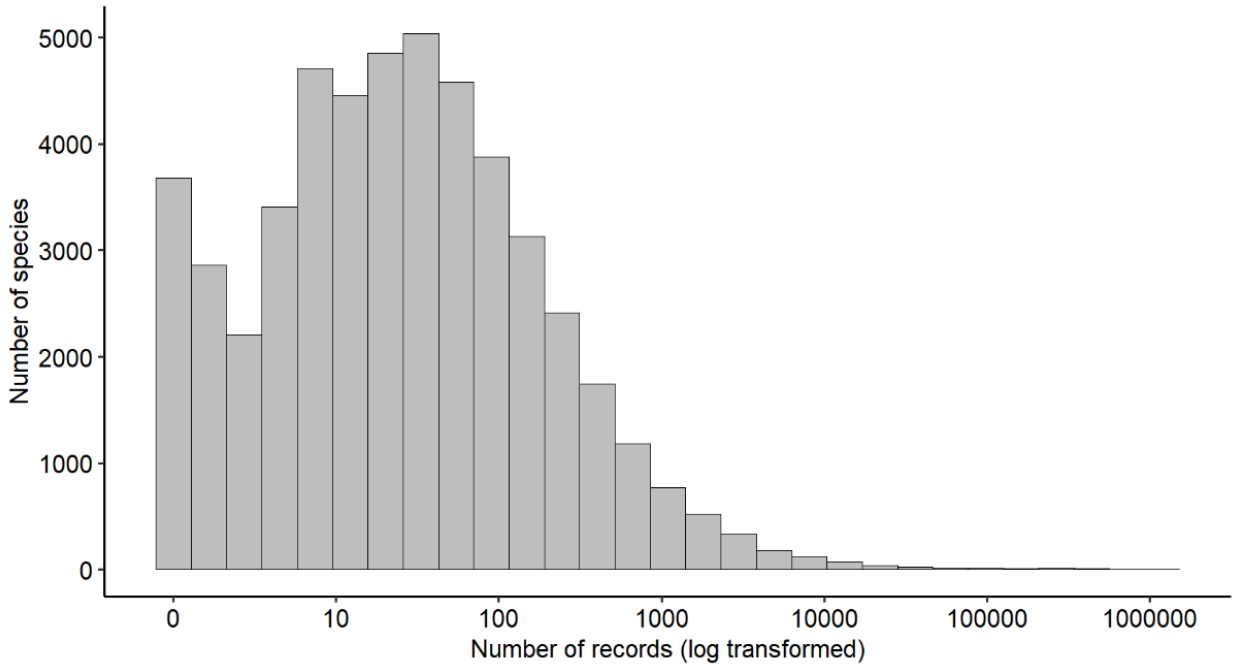

Figure 1 Frequency histogram of the number of occurrence records per species (the x-axis is log transformed).

The cleaned data were collected between 1600 and early 2022, with < 1% collected before 1900 (Fig. 2). Globally, the number of records increased exponentially from the late 1900s to the 21st century, and intensive data collection took place between the 1960s and 2005 (Fig. 2). The highest amount of data collected in a single year occurred in 2019 (1,359,037 records). Noteworthy, half of our cleaned data were collected from 2000 onwards, coinciding with the intensive data digitisation and mobilisation period (Devictor & Bensaude-Vincent, 2016).

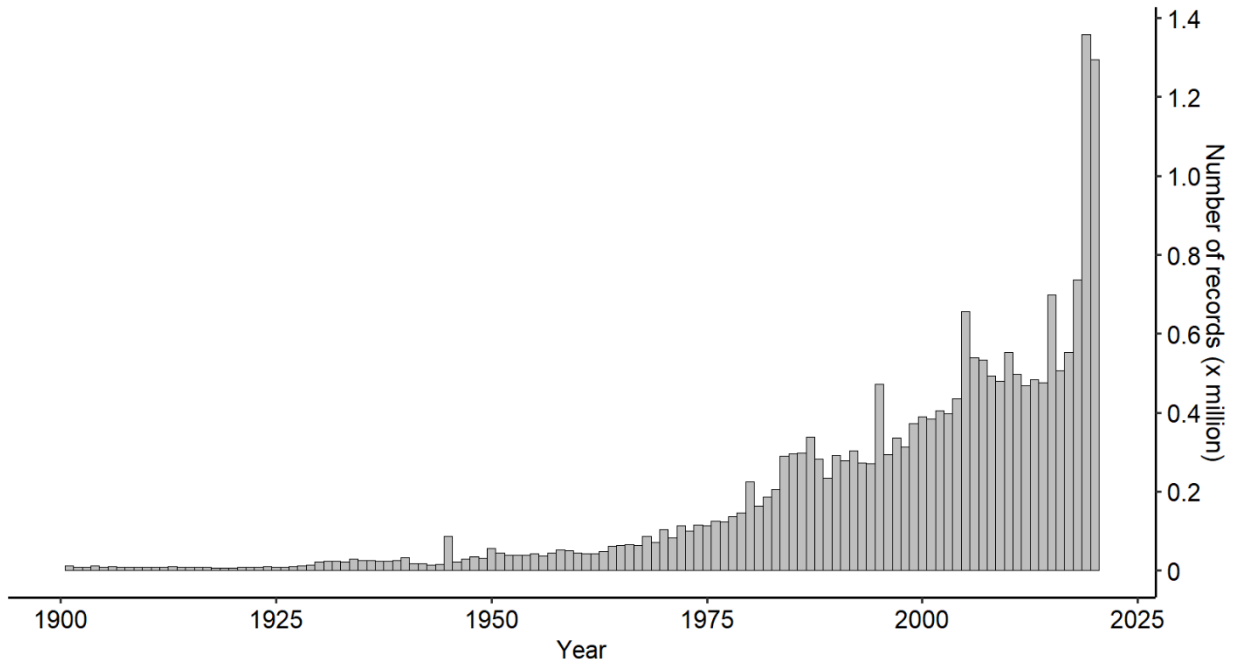

Figure 2 Trend in the number of digitised tree records collected per year from 1900 to 2021. Note that 0.7% (152,923) of records within our cleaned dataset were collected prior to 1900, and only 0.2% (39,587) were collected in early 2022.

At both spatial scales analysed, sampling effort (number of records/km^2) is unevenly distributed across the world (Fig. 3), with 95% (801) of ecoregions and 84% (11,095) of SUs having samples within our cleaned dataset and 83% (702) of ecoregions and 40% (5,316) of SUs contained at least 100 occurrence records. The mean density of records is 0.42 (± 1.33) and 0.21 (± 1.30) records/km^2 per ecoregion and SU, respectively. Sampling effort is more scattered at the SU scale, but generally, the highest sampling effort is concentrated in Europe, Japan, Central America, south-eastern Australia and New Zealand, and parts of east and west North America. In contrast, very few regions in Africa and Asia have high sampling effort, with major parts of central Asia, northwest of China, the Middle East and North Africa coinciding with deserts or mountainous regions lacking tree occurrence records.

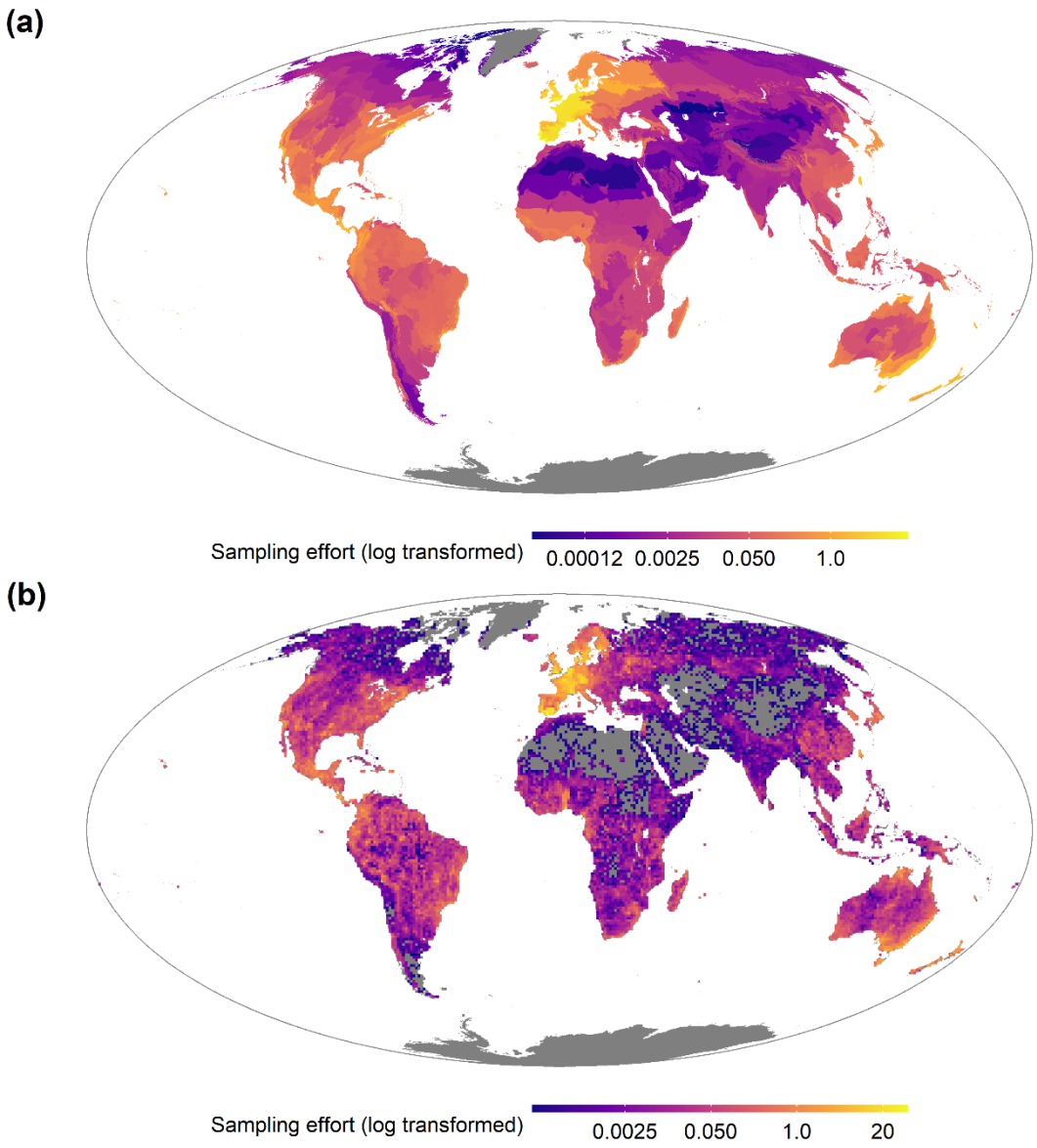

Figure 3 Spatial distribution of sampling effort for tree inventories across (a) 801 terrestrial ecoregions and (b) 11,095 SUs. Sampling effort values are stretched on a natural logarithm scale. Grey regions either contained no records or were not considered in the analysis.

The ecoregions with the highest sampling effort were the southwest Iberian Mediterranean sclerophyllous and mixed forests (16.64 records/km^2), Azores temperate mixed forests (14.90 records/km^2) and English lowlands beech forests (11.81 records/km^2), all located within Europe. In contrast, 47% (398) of ecoregions and 62% (8,249) of SUs have record density lower than 0.050 records/km^2 (see Appendix S4). Across the biomes, the mean sampling effort was highest in the Mediterranean forests, woodlands and scrub (1.51 ± 3.19 records/km^2), temperate broadleaf and mixed forests (1.40 ± 2.72 records/km^2) and tropical and subtropical coniferous forests (0.38 ± 0.33 records/km^2) (see Appendix S3).

Inventory Completeness

As would be expected, tree inventory completeness is unevenly distributed across the world at both spatial scales analysed (Fig. 4). At the ecoregional scale, the mean *C* value is 0.76 (± 0.18 SD) (see Appendix S5). At the SU scale, *C* values become more scattered, with the mean *C* reducing to 0.69 (± 0.24). Generally, sites with *C* values ≥ 0.80 are regarded as having more complete inventories or being "well-known" (at least 80% of the species have been sampled) (Haque et al., 2017; Mora et al., 2008; Soberón et al., 2007). However, estimates based on a small number of records may result in artifactual and unreliable *C* values. Therefore, we used a combination of the *C* value and the median of the sample size distribution as criteria to define well-known sites. At the ecoregional scale, the median number of records is 2,000 (rounded to the nearest thousand), and at the SU scale, it is 100 records (rounded to the nearest hundred) (see Appendix S2). Hence, we restrict well-known sites to ecoregions and SUs with $C \geq 0.80$, and ≥ 2,000 records and ≥ 100 records, respectively.

There are 5,316 SUs with ≥ 100 records and 417 ecoregions with ≥ 2,000 records (Appendix S5). As such, 37% (293) of ecoregions and 25% (2,754) of SUs analysed can be considered well-known (Fig. 4). At the ecoregional scale, well-known sites are concentrated around Europe, the USA, parts of Brazil, western Africa, Australia and New Zealand. Notwithstanding, most ecoregions remain under-inventoried, including those located in most parts of Africa, the Middle East, and Asia. At the SU scale, tree inventories are more complete in eastern Europe, eastern and western USA, Australasia (except New Guinea), Japan and New Zealand. In contrast, most sites in Asia and Africa have low inventory completeness.

Across forested biomes, tree inventory completeness is highest in boreal forests/taiga (mean $C = 0.90 \pm 0.10$), followed by temperate broadleaf and mixed forests (0.88 ± 0.13) and temperate conifer forests (0.83 ± 0.13), and lowest in mangroves (0.60 ± 0.17) (see Appendix S3). For non-

forested biomes, inventory completeness is highest in Mediterranean forests, woodlands and scrubs (0.87 ± 0.12), and tundra (0.86 ± 0.14), and lowest in flooded grasslands and savannas (0.56 ± 0.23).

Generally, the spatial pattern of observed versus estimated tree species richness is similar at both scales (Fig. 5a,b,d,e). Highest observed and estimated species richness is concentrated in the tropics, particularly South America, southeast Asia, and eastern China. The spatial distribution of unrecorded species revealed that most are located in the tropics, particularly Borneo (Fig 5. c,f).

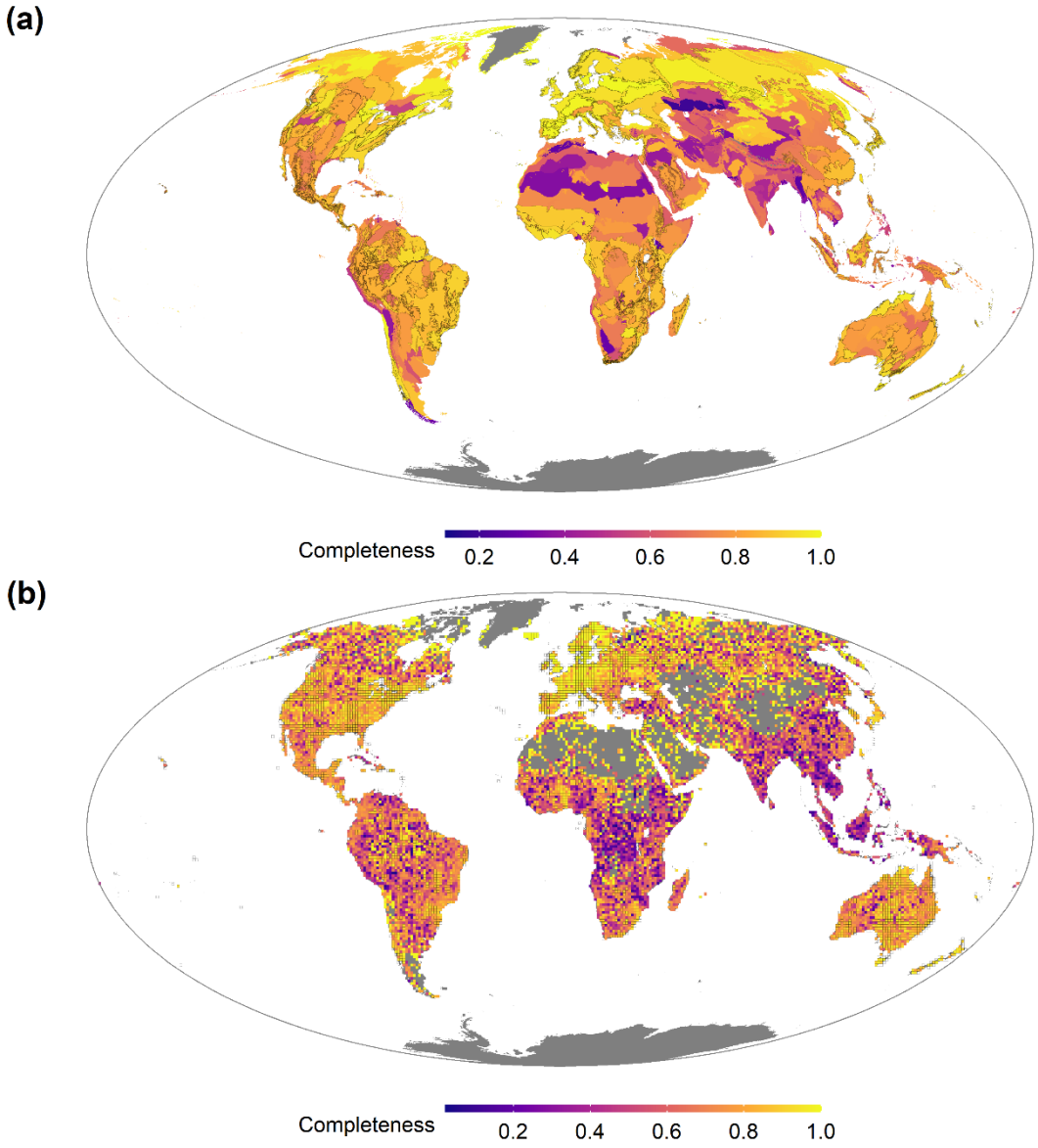

Figure 4 Spatial distribution of tree inventory completeness (a) at the ecoregional scale and (b) at a 100 x 100 km SU scale. Ecoregions and SUs outlined in grey represent sites we considered to be well-known, i.e., $C \geq 0.80$, and ≥ 2000 records and ≥ 100 records, respectively.

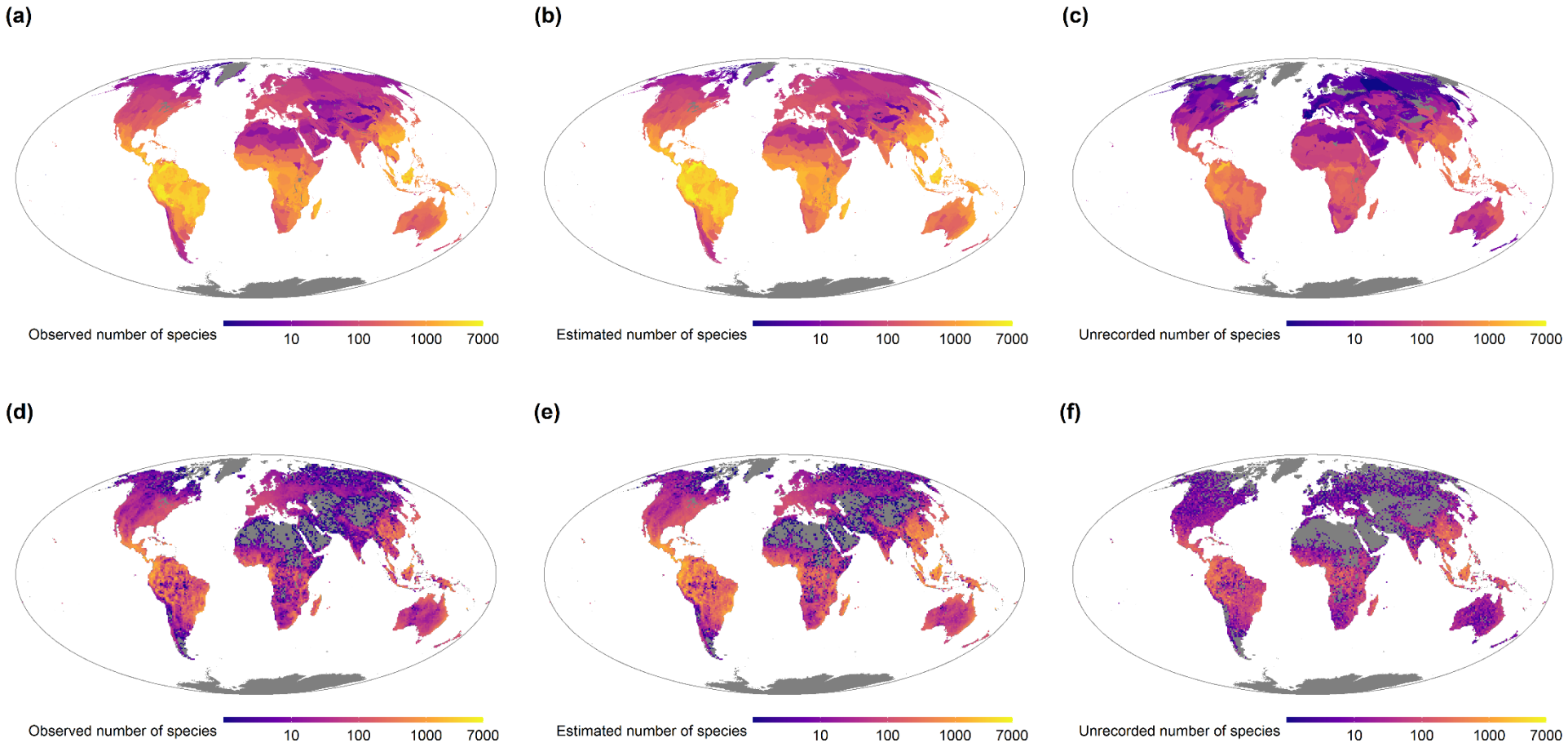

Figure 5 Spatial distribution of observed, estimated and unrecorded number of species at (a-c) ecoregional and (d-f) SU scales.

Sampling Effort and Inventory Completeness

At both ecoregional and SU scales, inventory completeness is weakly predicted by sampling effort (correlation of determination [r^2] = 0.14 and 0.026, respectively; Fig. 6). This means that completeness is partially influenced by sampling effort and higher sampling efforts tend to result in higher *C* (Fig. 6). However, there are exceptions with some sites having relatively low sampling effort but higher *C* (e.g., most ecoregions in the boreal forests/taiga and tundra biomes).

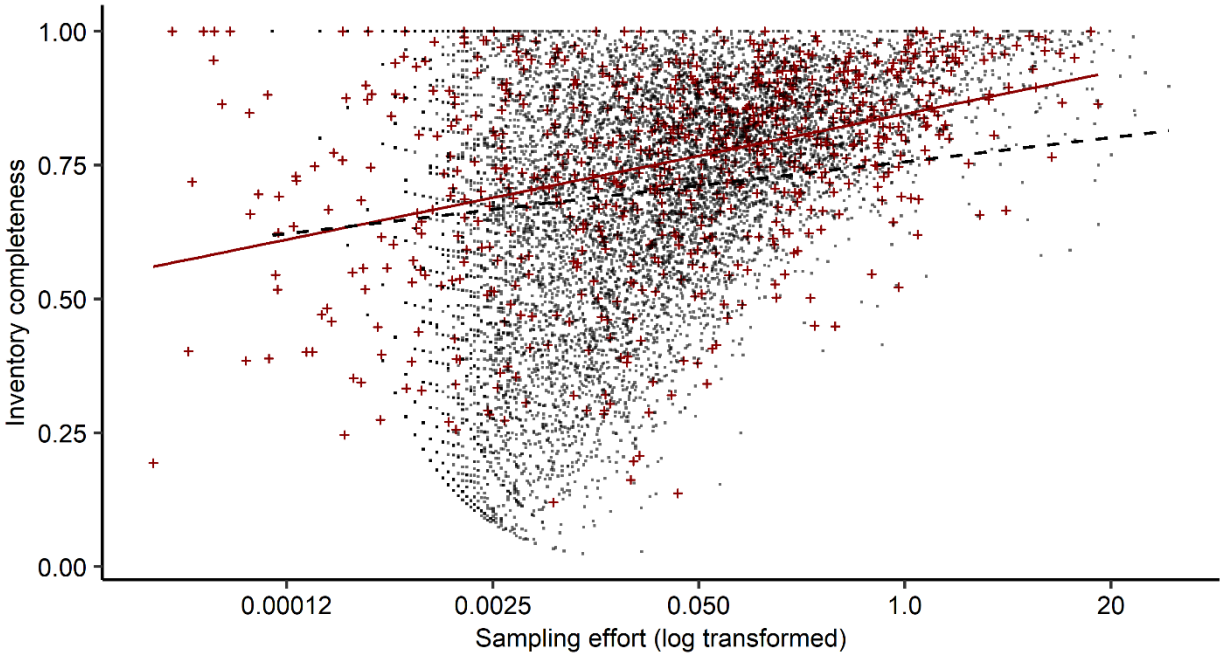

Figure 6 Plot of the relationship between inventory completeness and sampling effort (number of records/km^2). Red crosses represent ecoregions and black squares represent SUs. The solid and dashed lines of linear regressions are added to assist in visualising trends for ecoregions (adjusted r^2 = 0.14; $p < 0.001$) and SUs (adjusted r^2 = 0.026; $p < 0.001$), respectively. Note that the x-axis has been stretched on a natural logarithm scale.

Global Priority Areas for Sampling Tree Diversity

We plotted bivariate maps of estimates of *C* against NNH (for ecoregions) and FLII (for SUs) to visualise the spatial distribution of priority areas for future sampling (Fig. 7). At the ecoregional scale, we found sites having low inventory completeness and retaining < 50% of their natural habitat (NNH 3–4) concentrated in Asia, including most ecoregions in India (e.g., the east Deccan moist deciduous forests, lower Gangetic plains moist deciduous forests, and Narmada Valley dry deciduous forests), northeast and south China (e.g., the northeast China plain deciduous forests and Yunnan Plateau subtropical evergreen forests) and southeast Asia (e.g., the Irrawaddy

moist deciduous forests and dry forests, Tonle Sap-Mekong peat swamp forests, Chao Phraya freshwater swamp forests and Sumatran freshwater swamp forests). When natural habitats within these ecoregions are highly degraded and tree inventory completeness is low, sampling opportunities and the likelihood of obtaining a complete tree inventory is considerably diminished.

Ecoregions that still retain ≥ 50% of their natural habitat (NNH 1–2) but have incomplete inventories include the Juruá-Purus moist forests and Xingu-Tocantins-Araguaia moist forests in Brazil, the southern New Guinea freshwater swamp forests and southern New Guinea lowland rain forests in Papua New Guinea, the Mizoram-Manipur-Kachin rain forests located between Bangladesh and Myanmar, the central Congolian lowland forests in central Africa and the eastern Canadian Forest-Boreal transition in North America. These sites correspond to areas of opportunities where additional sampling effort can most likely yield species previously unrecorded for these areas.

Ecoregions that retain < 50% of their natural habitat (NNH 3–4) but contain high inventory completeness are predominately located in the global North, including Europe (e.g., the western European broadleaf forests, Scandinavian and Russian taiga and Celtic broadleaf forests), east coast of Australia (e.g., the southeast Australia temperate forests and eastern Australian temperate forests), New Zealand (e.g., the northland temperate kauri forests and New Zealand north island temperate forests), and east and west USA (e.g., the New England-Acadian forests and Ozark mountain forests). The western Guinean lowland forests and eastern Guinean forests in West Africa, the central Korean deciduous forests in Korea, the Caatinga and Araucaria moist forests in South America also fall within category. Most of the forested ecoregions in Russia (e.g., the east Siberian taiga and Okhotsk-Manchurian taiga), Canada (e.g., the northern and eastern Canadian Shield taigas), central Africa (e.g., the northeast and northwest Congolian lowland forests), and central and north Brazil (e.g., the Guianan lowland moist forests and Purus-Madeira moist forests) contain ≥ 50% of their natural habitat and have relatively high inventory completeness at the ecoregional scale.

At the SU scale, sites with low inventory completeness and low forest integrity were concentrated in Asia and scattered across several sites in east Brazil and west Africa. These represent areas of reduced sampling opportunities due to high anthropogenic impacts. Sites with low inventory completeness and high forest integrity were predominately located in central Brazil, central Africa, Papua New Guinea and scattered in Canada and Russia, corresponding to high sampling opportunities. Sites with high inventory completeness and low forest integrity were mostly located in the global North, particularly, west and east of the USA, Europe, Australia and

New Zealand. Lastly, sites with both high inventory completeness and forest integrity were scattered in northern North America, central Brazil and Russia.

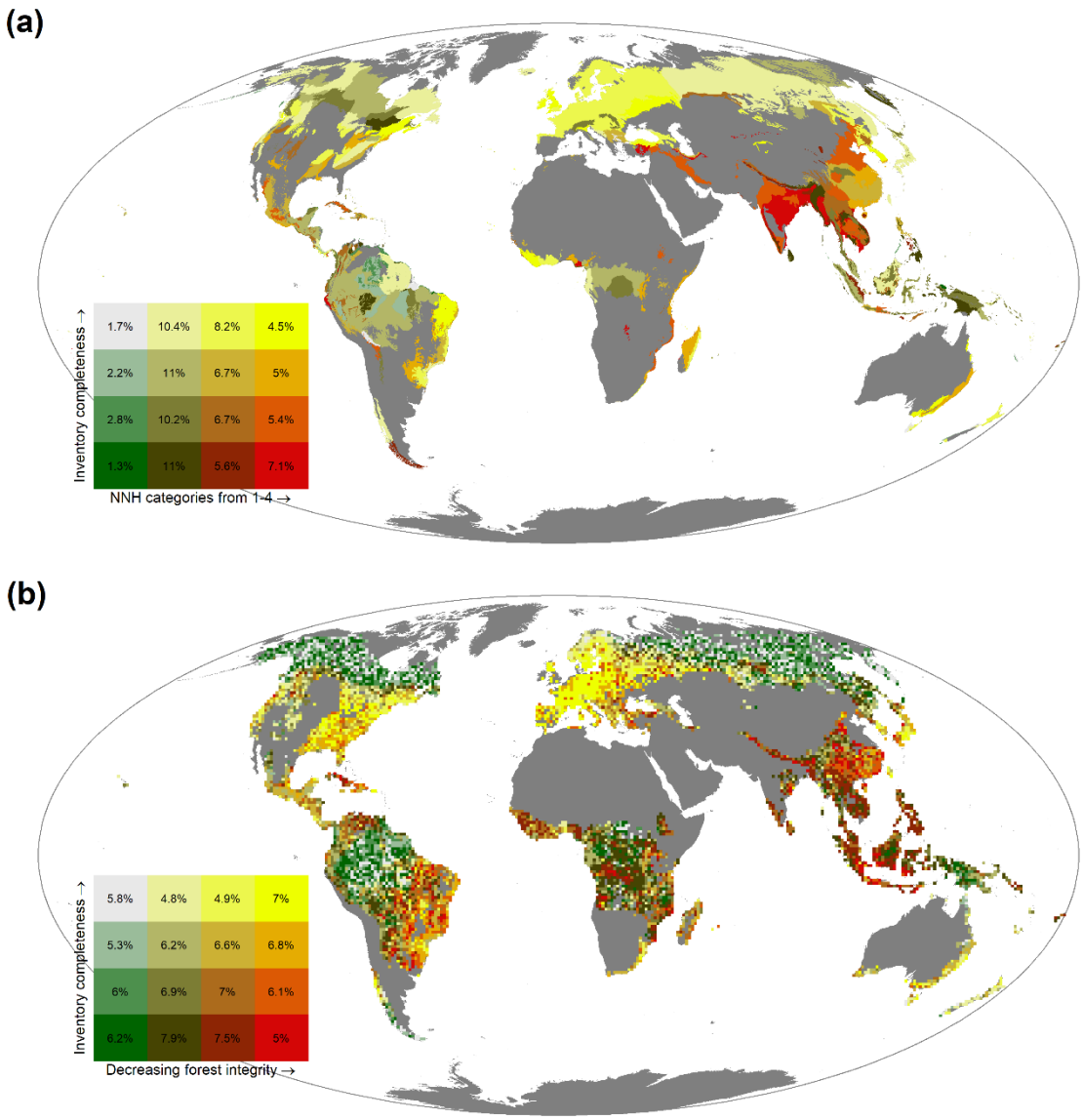

Figure 7 Bivariate maps of tree inventory completeness (a) versus Nature Needs Half (NNH) categories for 462 forested ecoregions and (b) versus Forest Landscape Integrity Index (FLII). The NNH categories are numbered 1 to 4, representing ecoregions that are described by Dinerstein et al. (2017) as (1) “Half Protected”, (2) “Nature Could Reach Half”, (3) “Nature Could Recover”, and (4) “Nature Imperiled”, respectively. The Forest Landscape Integrity Index (FLII) ranges from zero to ten with ten representing the highest forest integrity (Grantham et al., 2020).

Discussion

Our knowledge and understanding of species diversity and distribution patterns is typically derived from biodiversity taxonomic inventories, and to date countless studies have utilised species occurrence data to explore these patterns (Brummitt et al., 2020; Kier et al., 2005; Kreft & Jetz, 2007; Lovett et al., 2000). However, biases and uncertainties in species occurrence data can result in distorted patterns of diversity and distribution. Thus, it is essential to determine the likely completeness of species inventories in order to evaluate the level of confidence in our estimations and predictions of richness (Sousa-Baena et al., 2014). In this study, we analysed digitised records of the world's tree species and explored the trend in tree data collection across species, locations, and time. We then characterised spatial patterns in sampling effort and tree inventory completeness across the world at ecoregional and 100 x 100 km SU scales. We illustrated how sampling effort and the completeness of tree inventories are unevenly distributed, with these two variables being positively, albeit weakly, correlated. We extended our research to identify knowledge gaps (i.e., regions with low levels of completeness) and highlight areas of lost and future sampling opportunities to help prioritise and maximise the efficiency of future botanical surveys.

Trends in Data Collection

Since the mid-1900s, there has been an increase in the annual accumulation of tree records, highlighting a growing interest in tree inventories, and improved technology for accessing and sampling biodiversity, among others. The intensity in data collection, especially in the last 30 years (Fig. 2), corresponds to the period of strenuous efforts to digitise and mobilise biodiversity data (Heberling et al., 2021). Exponential growth in digitised records over recent years has also been observed in global studies for other taxonomic groups, including butterflies and freshwater fish (Girardello et al., 2019; Mora et al., 2008). Although our study did not analyse the temporal trend in tree data accumulation per country/ecoregion, it is highly likely that trends in data accumulation over the last century differ throughout the world, as observed by Girardello et al.'s (2019) global analysis of butterfly inventory completeness.

Despite the increasing number of digitised tree records available in the GBIF repository, taxonomic and spatial biases exist. At both spatial scales included in this study, high sampling efforts are clearly concentrated in the USA, Europe and Australia, a pattern also identified for birds, mammals, amphibians and vascular plants (Meyer et al., 2015; Meyer et al., 2016). Although our global sample coverage is relatively high for both ecoregional (95%) and SUs (84%) scales, it is

noteworthy that sampling efforts within many of those sites remain very sparse and this will be evident at finer resolutions.

There is notable taxonomic bias in digitised tree records. That over 85% of tree occurrence records included in our study came from less than 4% of tree species is unsurprising. Tobler, Honorio, Janovec, and Reynel's (2007) study of biodiversity knowledge in Peru revealed that most species were represented by only a small number of specimens collected from a few sites, while a few broadly distributed species tend to dominate collections. Previous assessment of tree occurrence data from five major online databases also found a similar trend in taxonomic bias, with only 26% of ~ 49,000 tree species having at least 20 records of high quality (Serra-Diaz et al., 2018). In addition, a recent global analysis of over 370,000 species across terrestrial and marine realms revealed that more than 50% of those species' records account for < 2% of the studied species, with birds being largely overrepresented (Hughes et al., 2021). Thus, efforts must continue to collect, digitise and mobilise biodiversity data strategically in order to address the Wallacean shortfall and achieve a better understanding of tree diversity and distribution.

How Complete are Global Tree Inventories?

Inventory completeness provides a useful index for accessing the degree of species inventory completeness and the state of biodiversity knowledge about a region (Soberón et al., 2007). Different parametric and non-parametric models have been developed to estimate species richness and the results of such models usually vary considerably depending on the attributes of the data, thus yielding different degrees of biases, precisions, and efficiencies (Colwell & Coddington, 1994; Gotelli & Colwell, 2011). Although different studies may recommend alternative models, the *Chao1* estimator is perhaps one of the most common non-parametric methods used to estimate species richness because it is easy to calculate and has been shown to perform well in landscapes with different bioclimatic conditions (Haque et al., 2017; Sousa-Baena et al., 2014; Stropp et al., 2016). While this method is conservative and estimates the lower bound of species richness, Chao (1984) found that it performed well on test data sets and it has since been widely used in many studies estimating richness using presence-only records (Ballesteros-Mejia et al., 2013).

To address the concern over unreliable estimates of inventory completeness (C) due to the small sample size, we restricted well-known sites to ecoregions and SUs with $C \geq 0.80$, and $\geq 2{,}000$ records and ≥ 100 records, respectively. According to this threshold, our global tree inventory assessment revealed that 37% of ecoregions and 25% of SUs are well-known. This finding indicates that taxonomic inventories are highly scale-dependent, and inventory completeness is lower at finer

spatial scales. We note that our criteria for well-known sites, especially at the ecoregional scale, are stricter than several other studies (Sousa-Baena et al., 2014; Stropp et al., 2016), but this gives us more confidence in our estimates and helps to identify sites where biodiversity knowledge is truly reliable versus those where information is incomplete or clustered in a few well-sampled regions.

Ecoregions with high tree inventory completeness are concentrated in Europe, the USA and Central America, parts of South America and eastern Africa, Australasia and a few countries in east Asia. A similar pattern was observed for inventory completeness of freshwater fish species at a country level (Pelayo-Villamil et al., 2018) and butterfly inventories at a coarse spatial resolution of 880 km (Girardello et al., 2019). A similar pattern was also observed at the SU scale: SUs having high inventory completeness are restricted to Europe, the USA, Australasia (except New Guinea), Japan and Korea, while low inventory completeness was largely concentrated in the tropics, especially in southeast Asia, central Africa and South America, and scattered throughout Canada and east Asia. This pattern is also common in other taxa including butterflies, vertebrates and vascular plants (Girardello et al., 2019; Meyer et al., 2015; Meyer et al., 2016).

We found that mean species inventory completeness also varies greatly among the world's biomes, with the Mediterranean forests, woodlands and scrub, and the temperate forests having higher completeness, while tropical and subtropical forests have lower completeness. However, while we also found tree inventories in tundra and boreal forests to have high average completeness, this contrasts to vertebrates (Meyer et al., 2015) and butterflies (Shirey et al., 2021), which were found to be vastly under-inventoried in this biome. This highlights a strong geographic difference in completeness among different taxa. Consequently, a single-taxon inventory completeness pattern may be a poor predictor for un-assessed taxa, and taxon-specific information maps should be independently identified (Meyer et al., 2015).

Within our study, sampling effort partially explained inventory completeness at both spatial scales analysed. Generally, well-known regions across Europe, North America and Australia have the highest number of records per unit area (Fig. 3 & Fig. 4). However, at the SU scale, this is less the case for the species-rich regions in South America, and the boreal forests. It should also be noted that incompleteness may result from a sampling strategy in which sites with high number of poorly planned sampling points can yield biased estimates of true richness, and thus coherent and systematic surveys are required to capture true richness (Haque et al., 2017; Soberón et al., 2007).

Drivers of Inventory Completeness

Generally, sampling effort and inventory completeness are positively correlated, although the relationship changes with resolution (Fig. 6). Sampling effort can partially explain spatial biases in inventory completeness, and it can be driven by several variables, including geographic and socio-economic factors (Hughes et al., 2021; Meyer et al., 2015; Yang et al., 2014). Firstly, the observed high tree inventory completeness in the Global North (Europe, North America, and Australia) is likely to be positively correlated to the country's gross domestic product (GDP)/capita (Hughes et al., 2021) as has been shown elsewhere (e.g., Meyer et al., 2015). While research, financial and institutional resources are available in developed countries, research funding and infrastructure are limited in countries with lower GDP, thus high sampling effort and coverage is almost exclusive to the Global North (Hughes et al., 2021). Meyer et al. (2015) also found that a country's commitment to GBIF data sharing emerged as a strong factor determining completeness of digitised biodiversity data. However, it should be noted that even within a country inventory completeness may be unevenly distributed and regions with herbaria and research infrastructure may have higher sampling effort (Pelayo-Villamil et al., 2018; Yang et al., 2014). Also, socio-political conflicts within the country will negatively influence botanical interest, and war-torn regions are less likely to have high biodiversity inventory completeness because collectors and researchers will restrict sampling to areas that are more politically stable, lacking armed conflicts (Meyer et al., 2015).

Moreover, human population density has shown to be positively correlated with inventory completeness, and species inventories in densely populated areas are likely to be more complete (Yang et al., 2014). Accessibility has also been reported to have a strong effect on sampling effort and the resulting completeness (Ballesteros-Mejia et al., 2013; Tobler et al., 2007; Stropp et al., 2016). Collectors and researchers usually choose areas that are close to roads, airports, and infrastructure which are easy to access, and these locations are often associated with urbanised areas and major cities (Hughes et al., 2021; Meyer et al., 2015). Citizen science observations are also closely linked to accessibility, which might limit our knowledge of species distributions to populated areas (Hughes et al., 2021). Inventory completeness can be influenced by certain appeal, such as endemism, biodiversity hotspots, protected areas or species 'charisma' (Girardello et al., 2019; Meyer et al., 2015). Studies have found that certain taxa have higher inventory completeness in mountainous areas (Girardello et al., 2019; Meyer et al., 2015; Yang et al., 2014). Collectors can also appear to show preference for reserves or for a particular species or taxa. Thus, citizen science or human observation may contribute to biases in inventory completeness (Hughes et al., 2021). The spatial biases in sampling effort and inventory completeness observed in this study could also

be attributed to the fact that nearly 80% of our cleaned data came from human observations. Although citizen science has greatly improved digitised biodiversity knowledge, efforts should be put into acquiring comparable data from less-accessible areas where community science observations are insufficient (Heberling et al., 2021). Additionally, botanical interest and the availability of research and data mobilisation programmes can result in high sampling coverage (Meyer et al., 2016). For instance, the Missouri Botanical Garden has long focused on the botanical exploration in Madagascar due to the country's high plant diversity and endemism, and it was one of the first institutions to engaged in data mobilisation for Madagascan records (Meyer et al., 2016).

Priorities for Sampling Tree Diversity

Digitised biodiversity data has significantly transformed science and research, enabling us to study biodiversity at scales previously not possible (Heberling et al., 2021). However, for many taxa, these data either do not exist or remain unrepresentative of the spatial distribution of the species. In sum, much remains to be discovered. However, additional information must be collected strategically if current gaps and biases are to be addressed. We identified regions where digitised knowledge of tree diversity is lacking (low inventory completeness) at both ecoregional and SU scales and where the greatest gains in inventory completeness could be made given current natural habitat and forest conditions. We also provide recommendation for future sampling and conservation efforts (Box 1).

Ecoregions for which the area occupied by both natural habitat and protected land spans less than 50% (NNH 3–4), and which have low tree inventory completeness, have rapidly diminishing opportunities for completing their tree inventories, because of high anthropogenic impact and low vegetation intactness. It is possible that many unrecorded species may, or have, become extinct even before they are discovered. Immediate action is required to manage and reduce present threats. More broadly, restoration efforts will be required to maintain ecological processes of existing tree populations and botanical surveys will be needed to document what is remaining (Box 1).

Ecoregions with low tree inventory completeness but that also have ≥ 50% of their area containing natural habitat or protected land (NNH 1–2) are regions of exploration opportunities where new biodiversity knowledge can be acquired. Threats in these ecoregions should be limited and establishing new protected areas will be beneficial for achieving comprehensive biodiversity conservation. For those ecoregions where coverage of natural habitat and protected land spans less than 50% of their area (NNH 3–4), but that have high inventory completeness (as is the case for many developed countries), management of present threats and restoration efforts will be required

to ensure ecological functioning of the existing tree populations. Botanical surveys should be carried out periodically to confirm whether species collected previously are still representative of those found on the ground. Lastly, ecoregions with high inventory completeness and ≥ 50% of their natural habitat and protected land remaining (NNH 1–2) should continue to be monitored and managed (establish new protected areas to achieve representativeness). As climate change intensifies, these ecoregions may experience increasing species turnover, hence systematic botanical surveys should be carried out to capture community-level changes.

We highlight some examples where biodiversity conservation goals have been followed by practical and strategic actions (Box 1). Although local conservation actions typically occur at a much smaller spatial scale (Meyer et al., 2015), in this study we provide recommendations for botanical survey and conservation efforts at the ecoregional scale. Ecoregions provide a useful starting point for regional conservation planning and is often used as a biogeographic framework to highlight distinct areas of the world with high representation values (Dinerstein et al., 2017; Olson et al., 2001). The concept of ecoregions is becoming increasingly important as scientists realise that species-specific conservation methods do not allow for the conservation of ecological communities and ecosystems (Shreeve & Dennis, 2011). In fact, many countries have adopted ecoregion strategies for biodiversity management which have included indigenous communities to manage large areas of land and these strategies have proven success even in countries with low GDP (Dinerstein et al., 2017).

Improving biodiversity data

Since digitised biodiversity data are known to suffer from uncertainties and biases, one of the easy way to fix issues related to data quality is to add or correct geographical coordinates, full collection date and standardised scientific names for the record (Sousa-Baena et al., 2014), or improve data digitisation processes to make it easier for citizen science. Systematic approaches to data cleaning and quality assessment will also increase the reliability of digitised biodiversity data, although there is no one-size-fits-all solution to data cleaning (Zizka et al., 2020). Other sources of valuable biodiversity data, many of which are housed in natural history collections waiting to be identified, should be made available in digital format and this effort should be supported sufficiently (Sousa-Baena et al., 2014). Also, a commitment to data sharing and mobilisation should be encouraged and supported at the national level (Meyer et al., 2015). Lastly, strategic sampling and digitisation in under-inventoried regions will help to improve representativeness, reduce biases and increase our knowledge of tree diversity (Sousa-Baena et al., 2014).

Box 1. Recommendation for increasing biodiversity knowledge and conserving ecoregions based on the degree of tree inventory completeness and the amount of remaining natural habitat and protected land (NNH; Dinerstein et al., 2017).

Degree of tree inventory completeness, and remaining natural habitat and protected land	Recommendation for appropriate ecoregional level management and sampling effort	Example of effective implementation guided by key principles of biodiversity conservation
Ecoregions having low degrees of tree inventory completeness, and < 50% of natural habitat and protected land remaining (orange-red in Fig. 7)	Immediate action is required to manage present threats to prevent further loss of natural habitat and forest integrity. Botanical surveys are required to document what remains in order to better prepare for habitat restoration to link connectivity and increase tree population size.	Ecoregion conservation strategies in Nepal involve local communities in managing forested areas. Small but abundant community-managed forest parcels facilitated population recovery of many endangered large mammals including a 61% increase in tigers and a 31% increase in rhinos between 2008 and 2013. In return, communities receive half of the revenue generated by wildlife parks (Wikramanayake et al., 2010).
Ecoregions having low degrees of tree inventory completeness, and ≥ 50% of natural habitat and protected land remaining (dark green in Fig. 7)	Identify and limit present threats to ensure vegetation is intact and forest integrity remains high. Redistribute resources into sampling these ecoregions to increase biodiversity knowledge and explore opportunities for biodiversity conservation measures such as expanding protected areas.	Expansion of protected areas in the Brazilian Amazon between 1997 and 2008 led to a 37% reduction in deforestation rates between 2004 and 2006 in the Brazilian Amazon (Soares-Filho et al., 2010).
Ecoregions having high degrees of tree inventory completeness, and < 50% of natural habitat and protected land remaining (yellow in Fig. 7)	Manage and reduce present threats to maintain functional populations of extant species and ensure forest integrity does not decline further. Habitat restoration is required to link connectivity, increase tree population size and allow tree species to cope with rapid land-use changes. Monitor and survey extant species periodically against present threats.	The north Atlantic coast ecoregions are critically endangered with 40% of their natural habitat lost to urban development and agriculture. Conservation in this region is a collection of efforts from public agencies and private organisations, and about 5% of land secured for conservation is attributed to the Nature Conservancy (The Nature Conservancy, 2006).
Ecoregions having high degrees of tree inventory completeness, and ≥ 50% of natural habitat and protected land remaining (light grey and pale yellow in Fig. 7)	Identify and limit present threats to ensure vegetation is intact and forest integrity remains high. Conserve existing biodiversity by expanding protected areas. Periodically carry out botanical surveys to capture species turnover.	The Yukon Arctic ecoregions in Alaska have been assessed by the Nature Conservancy (a non-profit conservation organisation) using a predictive ecosystem model to identify areas that are representative of the ecoregion's biodiversity for future conservation (The Nature Conservancy, 2005).

Limitation of Study

At coarser spatial resolutions, caution must be taken when interpreting completeness because high *C* may result from sampling artefacts where *C* may be overestimated from a few well-surveyed sites. Although analysis at finer resolutions reduces sampling artefacts because beta-diversity is reduced and sampling is more representative of the full site diversity, it may not always be achievable, especially for large-scale analysis, because of the existing gaps in sampling efforts.

It is important to note that the results reported here are analysed from digitised tree records aggregated in GBIF, which may not be complete. We acknowledge that in many regions of the world, inventory data may be available but not yet digitised and/or mobilised, or specimens have been collected but remain to be identified due to the lack of resources (Girardello et al., 2019; Serra-Diaz et al., 2018). We also considered that tree data may be stored in other major online biodiversity databases such as the Botanical Information and Ecological Network (BIEN; https://bien.nceas.ucsb.edu/bien/) and the Global Forest Biodiversity Initiative (GFBI; https://www.gfbinitiative.org/), and region-specific electronic databases such as the Latin American Seasonally Dry Tropical Forest Floristic Network (DRYFLOR; http://www.dryflor.info/), and the Sub-Saharan tropical Africa database RAINBO (http://rainbio.cesab.org/).

Although beyond the scope of this study, it is also important to examine when a site has achieved high inventory completeness because temporal biases in completeness can result in a distorted view of biodiversity trends over time (Haque et al., 2018). If most of the records collected were from historical periods, in view of species turnover and increasing land-use changes coupled with climate variability, it is unclear whether the sets of species found several decades ago are still representative of what is observed in those regions today (Stropp et al., 2016). This points to the risk of out-of-date knowledge in assuming that a site is well-known, while in reality some species might have been extirpated from that area (Stropp et al., 2016). Thus, ongoing and systematic botanical surveys are crucial to update inventories periodically, even in sites where current knowledge is regarded as complete.

It should also be considered that the *Chao1* estimator is scale-dependent and estimating richness at low resolutions may lead to inflation in *C* values, which may make it difficult to draw meaningful conclusions or adequately answer questions in ecological studies (Meyer et al., 2015; Soberón et al., 2007; Sousa-Baena et al., 2014). However, as previously pointed out, we characterised inventory completeness at the ecoregional scale because ecoregions represent distinct assemblages of biodiversity in particular regions and provide a useful base map for conservation planning as they draw on natural boundaries, rather than political ones (Dinerstein et al., 2017). To

account for the sensitivity of our estimator, we analysed completeness at the SU scale and observed a relatively consistent pattern, although *C* values were more scattered at a finer scale. Even at this resolution, many sites lacked sample coverage and if we were to define well-known sites with stricter criteria, most of the world would be regarded as under-inventoried.

Additionally, because *C* is a ratio, informing us of how well a site has been surveyed, it does not necessarily indicate that knowledge of biodiversity in a site with a high *C* value is also high. For example, an ecoregion in Brazil with 10,000 tree species might have obtained a *C* of 0.8, but this also means that ~ 2000 species are yet to be recorded. In contrast, one ecoregion in the Australian dessert may have 10 species and a *C* of 0.8, meaning that there are only two species unrecorded.

Conclusion

Biodiversity data are increasingly being added to online databases and their potential applications are manifold provided data quality is accounted for. Our result showed that tree inventory completeness is not uniform across the world and sampling effort is highly biased towards the Global North, while much of the species-rich tropics remain under-inventoried. We also found taxonomic bias whereby over 85% of digitised tree records within our dataset came from just 4% of the species. Due to biases and uncertainties existing in biodiversity data, simply sampling more does not necessarily lead to knowledge increase. Hence, we illustrated how gaps in existing biodiversity data can help guide botanical surveys and highlighted areas representing future botanical exploration opportunities as well as areas of diminishing opportunities. We also provided some recommendations for conservation and botanical survey efforts. With ongoing anthropogenic impacts and escalating rates of biodiversity loss, limited resources should be allocated to systematically survey locations likely to yield new knowledge. Additionally, data digitisation and mobilisation should be supported and sufficiently funded to improve the quality and coverage of digitised biodiversity information.

References

Bakker, F. T., Antonelli, A., Clarke, J. A., Cook, J. A., Edwards, S. V., Ericson, P. G. P., Faurby, S., Ferrand, N., Gelang, M., Gillespie, R. G., Irestedt, M., Lundin, K., Larsson, E., Matos-Maraví, P., Müller, J., von Proschwitz, T., Roderick, G. K., Schliep, A., Wahlberg, N., . . . Källersjö, M. (2020). The global museum: Natural history collections and the future of evolutionary science and public education. *PeerJ*, *8*, e8225. https://doi.org/10.7717/peerj.8225

Ballesteros-Mejia, L., Kitching, I. J., Jetz, W., Nagel, P., & Beck, J. (2013). Mapping the biodiversity of tropical insects: species richness and inventory completeness of African sphingid moths. *Global Ecology and Biogeography*, *22*(5), 586-595. https://doi.org/10.1111/geb.12039

Beaumont, L. J., Gallagher, R. V., Leishman, M. R., Hughes, L., & Downey, P. O. (2014). How can knowledge of the climate niche inform the weed risk assessment process? A case study of Chrysanthemoides monilifera in Australia. *Diversity and Distributions*, *20*(6), 613-625. https://doi.org/https://doi.org/10.1111/ddi.12190

Beech, E., Rivers, M., Oldfield, S., & Smith, P. P. (2017). GlobalTreeSearch: The first complete global database of tree species and country distributions. *Journal of Sustainable Forestry*, *36*(5), 454-489. https://doi.org/10.1080/10549811.2017.1310049

BGCI. (2021a). *GlobalTreeSearch online database (version 1.5). Botanic Gardens Conservation International*. Retrieved 29 October 2021, from https://tools.bgci.org/global_tree_search.php

BGCI. (2021b). *State of the world's trees*. BGCI. https://www.bgci.org/wp/wp-content/uploads/2021/08/FINAL-GTAReportMedRes-1.pdf

Bisby, F. A. (2000). The quiet revolution: Biodiversity informatics and the internet. *Science (American Association for the Advancement of Science)*, *289*(5488), 2309-2312. https://doi.org/10.1126/science.289.5488.2309

Boyle, B., Hopkins, N., Lu, Z., Raygoza Garay, J. A., Mozzherin, D., Rees, T., Matasci, N., Narro, M. L., Piel, W. H., McKay, S. J., Lowry, S., Freeland, C., Peet, R. K., & Enquist, B. J. (2013). The taxonomic name resolution service: An online tool for automated standardization of plant names. *BMC Bioinformatics*, *14*(6), 1-14. https://doi.org/10.1186/1471-2105-14-16

Brummitt, N., Araújo, A. C., & Harris, T. (2020). Areas of plant diversity—What do we know? *Plants, People, Planet*, *3*(1), 33-44. https://doi.org/10.1002/ppp3.10110

Cao, J., Xu, J., Pan, X., Monaco, T. A., Zhao, K., Wang, D., & Rong, Y. (2021). Potential impact of climate change on the global geographical distribution of the invasive species, Cenchrus spinifex (Field sandbur, Gramineae). *Ecological Indicators*, *131*, 108204. https://doi.org/10.1016/j.ecolind.2021.108204

Chamberlain, S., Barve, V., Mcglinn, D., Oldoni, D., Desmet, P., Geffert, L., & Ram, K. (2022). *rgbif: Interface to the Global Biodiversity Information Facility API. R package version 3.6.0*. In https://CRAN.R-project.org/package=rgbif

Chamberlain, S., Szoecs, E., Foster, Z., Arendsee, Z., Boettiger, C., Ram, K., Bartomeus, T., Baumgartner, J., O'Donnell, J., Oksanen, J., Tzovaras, B. G., Marchand, P., Tran, V., Salmon, M., Li, G., & Grenié, M. (2020). *taxize: Taxonomic information from around the web. R package version 0.9.98.* In https://github.com/ropensci/taxize

Chao, A. (1984). Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of Statistics*, *11*(4), 265-270. http://www.jstor.org/stable/4615964

Cunze, S., Kochmann, J., & Klimpel, S. (2020). Global occurrence data improve potential distribution models for Aedes japonicus japonicus in non-native regions. *Pest management science*, *76*(5), 1814-1822. https://doi.org/10.1002/ps.5710

Devictor, V., & Bensaude-Vincent, B. (2016). From ecological records to big data: the invention of global biodiversity. *History and Philosophy of the Life Sciences*, *38*(4), 1-23. https://doi.org/10.1007/s40656-016-0113-2

Dinerstein, E., Olson, D., Joshi, A., Vynne, C., Burgess, N. D., Wikramanayake, E., Hahn, N., Palminteri, S., Hedao, P., Noss, R., Hansen, M., Locke, H., Ellis, E. C., Jones, B., Barber, C. V., Hayes, R., Kormos, C., Martin, V., Crist, E., . . . Saleem, M. (2017). An ecoregion-based approach to protecting half the terrestrial realm. *Bioscience*, *67*(6), 534-545. https://doi.org/10.1093/biosci/bix014

Edwards, J. L. (2004). Research and societal benefits of the global biodiversity information facility. *Bioscience*, *54*(6), 486-487. https://doi.org/10.1641/0006-3568(2004)054[0486:RASBOT]2.0.CO;2

Edwards, J. L., Lane, M. A., & Nielsen, E. S. (2000). Interoperability of biodiversity databases: Biodiversity information on every desktop. *Science (American Association for the Advancement of Science)*, *289*(5488), 2312-2314. https://doi.org/10.1126/science.289.5488.2312

Elith, J., & Leathwick, J. R. (2009). Species distribution models: Ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics*, *40*(1), 677-697. https://doi.org/10.1146/annurev.ecolsys.110308.120159

FAO, & UNEP. (2020). *The state of the world's forests 2020*. FAO and UNEP. https://doi.org/10.4060/ca8642en

Franklin, J., & Miller, J. A. (2009). *Mapping species distributions: Spatial inference and prediction*. Cambridge University Press.

Gatti, R. C., Reich, P. B., Gamarra, J. G. P., Crowther, T., Hui, C., Morera, A., Bastin, J.-F., de-Miguel, S., Nabuurs, G.-J., Svenning, J.-C., Serra-Diaz, J. M., Merow, C., Enquist, B., Kamenetsky, M., Lee, J., Zhu, J., Fang, J., Jacobs, D. F., Pijanowski, B., . . . Liang, J. (2022). The number of tree species on Earth. *Proceedings of the National Academy of Sciences, 119*(6). https://doi.org/doi:10.1073/pnas.2115329119

GBIF.org. (02 March 2022). *GBIF occurrence download*. https://doi.org/10.15468/dl.aptyh5

Girardello, M., Chapman, A., Dennis, R., Kaila, L., Borges, P. A. V., & Santangeli, A. (2019). Gaps in butterfly inventory data: A global analysis. *Biological conservation, 236*, 289-295. https://doi.org/10.1016/j.biocon.2019.05.053

Gotelli, N. J., & Colwell, R. K. (2011). Estimating species richness. In *Biological diversity: Frontier in measurement and assessment* (Vol. 12, pp. 39-54). Oxford University Press.

Grantham, H. S., Duncan, A., Evans, T. D., Jones, K. R., Beyer, H. L., Schuster, R., Walston, J., Ray, J. C., Robinson, J. G., Callow, M., Clements, T., Costa, H. M., DeGemmis, A., Elsen, P. R., Ervin, J., Franco, P., Goldman, E., Goetz, S., Hansen, A., . . . Watson, J. E. M. (2020). Anthropogenic modification of forests means only 40% of remaining forests have high ecosystem integrity. *Nature Communications, 11*(1), 5978. https://doi.org/10.1038/s41467-020-19493-3

Guisan, A., Tingley, R., Baumgartner, J. B., Naujokaitis-Lewis, I., Sutcliffe, P. R., Tulloch, A. I. T., Regan, T. J., Brotons, L., McDonald-Madden, E., Mantyka-Pringle, C., Martin, T. G., Rhodes, J. R., Maggini, R., Setterfield, S. A., Elith, J., Schwartz, M. W., Wintle, B. A., Broennimann, O., Austin, M., . . . Buckley, Y. M. (2013). Predicting species distributions for conservation decisions. *Ecology Letters*, *16*(12), 1424-1435. https://doi.org/https://doi.org/10.1111/ele.12189

Hansen, M. C., Potapov, P. V., Moore, R., Hancher, M., Turubanova, S. A., Tyukavina, A., Thau, D., Stehman, S. V., Goetz, S. J., Loveland, T. R., Kommareddy, A., Egorov, A., Chini, L., Justice, C. O., & Townshend, J. R. G. (2013). High-resolution global maps of 21st-century forest cover change. *Science*, *342*(6160), 850-853. https://doi.org/doi:10.1126/science.1244693

Haque, M. M., Nipperess, D. A., Baumgartner, J. B., & Beaumont, L. J. (2018). A journey through time: Exploring temporal patterns amongst digitized plant specimens from Australia.

Systematics and biodiversity, *16*(6), 604-613. https://doi.org/10.1080/14772000.2018.1472674

Haque, M. M., Nipperess, D. A., Gallagher, R. V., & Beaumont, L. J. (2017). How well documented is Australia's flora? Understanding spatial bias in vouchered plant specimens. *Austral Ecology*, *42*(6), 690-699. https://doi.org/10.1111/aec.12487

Heberling, J. M., Miller, J. T., Noesgaard, D., Weingart, S. B., & Schigel, D. (2021). Data integration enables global biodiversity synthesis. *Proceedings of the National Academy of Sciences - PNAS*, *118*(6). https://doi.org/10.1073/pnas.2018093118

Hortal, J., Bello, F. d., 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*(1), 523-549. https://doi.org/10.1146/annurev-ecolsys-112414-054400

Hsieh, T. C., Ma, K. H., & Chao, A. (2016). iNEXT: an R package for rarefaction and extrapolation of species diversity (Hill numbers). *Methods in Ecology and Evolution*, *7*(12), 1451-1456. https://doi.org/https://doi.org/10.1111/2041-210X.12613

Hughes, A. C., Orr, M. C., Ma, K., Costello, M. J., Waller, J., Provoost, P., Yang, Q., Zhu, C., & Qiao, H. (2021). Sampling biases shape our view of the natural world. *Ecography*, *44*(9), 1259-1269. https://doi.org/10.1111/ecog.05926

Hunt, D. R. (1996). The genera of temperate broadleaved trees. *Broadleaves*, *2*, 4 - 5.

Kier, G., Mutke, J., Dinerstein, E., Ricketts, T. H., Küper, W., Kreft, H., & Barthlott, W. (2005). Global patterns of plant diversity and floristic knowledge. *Journal of Biogeography*, *32*(7), 1107-1116. https://doi.org/10.1111/j.1365-2699.2005.01272.x

Kreft, H., & Jetz, W. (2007). Global patterns and determinants of vascular plant diversity. *Proceedings of the National Academy of Sciences - PNAS*, *104*(14), 5925-5930. https://doi.org/10.1073/pnas.0608361104

Lomolino, M. V. (2004). Conservation biogeography. In M. V. Lomolino & L. R. Heaney (Eds.), *Frontiers of biogeography: new directions in the geography of nature* (pp. 293-296). Sinauer Associate. https://doi.org/10.2980/1195-6860(2006)13[424:FOBNDI]2.0.CO;2

Lovett, J. C., Rudd, S., Taplin, J., & Frimodt-Møller, C. (2000). Patterns of plant diversity in Africa south of the Sahara and their implications for conservation management. *Biodiversity & Conservation*, *9*(1), 37-46. https://doi.org/10.1023/A:1008956529695

Maldonado, C., Molina, C. I., Zizka, A., Persson, C., Taylor, C. M., Albán, J., Chilquillo, E., Rønsted, N., & Antonelli, A. (2015). Estimating species diversity and distribution in the era of Big Data: to what extent can we trust public databases? *Global Ecology and Biogeography*, *24*(8), 973-984. https://doi.org/10.1111/geb.12326

Meyer, C., Kreft, H., Guralnick, R., & Jetz, W. (2015). Global priorities for an effective information basis of biodiversity distributions. *Nature Communications*, *6*, 8221. https://doi.org/10.1038/ncomms9221

Meyer, C., Weigelt, P., & Kreft, H. (2016). Multidimensional biases, gaps and uncertainties in global plant occurrence information. *Ecology Letters*, *19*(8), 992-1006. https://doi.org/10.1111/ele.12624

Mora, C., Tittensor, D. P., & Myers, R. A. (2008). The completeness of taxonomic inventories for describing the global diversity and distribution of marine fishes. *Proceedings of the Royal Society B: Biological Sciences*, *275*(1631), 149-155. https://doi.org/doi:10.1098/rspb.2007.1315

O'Donnell, J., Gallagher, R. V., Wilson, P. D., Downey, P. O., Hughes, L., & Leishman, M. R. (2012). Invasion hotspots for non-native plants in Australia under current and future climates. *Global Change Biology*, *18*(2), 617-629. https://doi.org/10.1111/j.1365-2486.2011.02537.x

Olson, D. M., Dinerstein, E., Wikramanayake, E. D., Burgess, N. D., Powell, G. V. N., Underwood, E. C., D'amico, J. A., Itoua, I., Strand, H. E., Morrison, J. C., Loucks, C. J., Allnutt, T. F., Ricketts, T. H., Kura, Y., Lamoreux, J. F., Wettengel, W. W., Hedao, P., & Kassem, K. R. (2001). Terrestrial ecoregions of the world: A new map of life on earth: A new global map of terrestrial ecoregions provides an innovative tool for conserving biodiversity. *Bioscience*, *51*(11), 933-938. https://doi.org/10.1641/0006-3568(2001)051[0933:Teotwa]2.0.Co;2

Pelayo-Villamil, P., Guisande, C., Vari, R. P., Manjarrés-Hernández, A., García-Roselló, E., González-Dacosta, J., Heine, J., González Vilas, L., Patti, B., Quinci, E. M., Jiménez, L. F., Granado-Lorencio, C., Tedesco, P. A., & Lobo, J. M. (2018). Global diversity patterns of freshwater fishes – potential victims of their own success. *Diversity and Distributions*, *21*(3), 345-356. https://doi.org/https://doi.org/10.1111/ddi.12271

Poorter, L., van der Sande, M. T., Thompson, J., Arets, E. J. M. M., Alarcón, A., Álvarez-Sánchez, J., Ascarrunz, N., Balvanera, P., Barajas-Guzmán, G., Boit, A., Bongers, F., Carvalho, F. A., Casanoves, F., Cornejo-Tenorio, G., Costa, F. R. C., de Castilho, C. V., Duivenvoorden, J. F., Dutrieux, L. P., Enquist, B. J., . . . Peña-Claros, M. (2015). Diversity enhances carbon storage in tropical forests. *Global Ecology and Biogeography*, *24*(11), 1314-1328. https://doi.org/https://doi.org/10.1111/geb.12364

POWO. (2022). *Plants of the World Online. Facilitated by the Royal Botanic Gardens, Kew.* http://www.plantsoftheworldonline.org/

Pyke, G. H., & Ehrlich, P. R. (2010). Biological collections and ecological/environmental research: A review, some observations and a look to the future. *Biological reviews of the Cambridge Philosophical Society*, *85*(2), 247-266. https://doi.org/10.1111/j.1469-185X.2009.00098.x

Qian, H., Deng, T., & Sun, H. (2019). Global and regional tree species diversity. *Journal of Plant Ecology*, *12*(2), 210-215. https://doi.org/10.1093/jpe/rty013

Qian, H., Zhang, Y., Zhang, J., & Wang, X. (2013). Latitudinal gradients in phylogenetic relatedness of angiosperm trees in North America. *Global Ecology and Biogeography*, *22*(11), 1183-1191. https://doi.org/https://doi.org/10.1111/geb.12069

R Core Team. (2021). *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.* In https://www.R-project.org/

Rivers, M. (2017). The global tree assessment - Red listing the world's trees. *BGjournal*, *14*(1), 16-19. https://www.jstor.org/stable/26369172

Rocchini, D., Hortal, J., Lengyel, S., Lobo, J. M., Jiménez-Valverde, A., Ricotta, C., Bacaro, G., & Chiarucci, A. (2011). Accounting for uncertainty when mapping species distributions: The need for maps of ignorance. *Progress in Physical Geography: Earth and Environment*, *35*(2), 211-226. https://doi.org/10.1177/0309133311399491

Salinas-Rodríguez, M. M., Sajama, M. J., Gutiérrez-Ortega, J. S., Ortega-Baes, P., & Estrada-Castillón, A. E. (2018). Identification of endemic vascular plant species hotspots and the effectiveness of the protected areas for their conservation in Sierra Madre Oriental, Mexico. *Journal for Nature Conservation*, *46*, 6-27. https://doi.org/https://doi.org/10.1016/j.jnc.2018.08.012

Serra-Diaz, J. M., Enquist, B. J., Maitner, B., Merow, C., & Svenning, J.-C. (2018). Big data of tree species distributions: how big and how good? *Forest Ecosystems*, *4*(1), 30. https://doi.org/10.1186/s40663-017-0120-0

Shirey, V., Belitz, M. W., Barve, V., & Guralnick, R. (2021). A complete inventory of North American butterfly occurrence data: Narrowing data gaps, but increasing bias. *Ecography (Copenhagen)*, *44*(4), 537-547. https://doi.org/10.1111/ecog.05396

Shreeve, T. G., & Dennis, R. L. H. (2011). Landscape scale conservation: Resources, behaviour, the matrix and opportunities. *Journal of Insect Conservation*, *15*(1), 179-188. https://doi.org/10.1007/s10841-010-9336-9

Slik, J. W. F., Arroyo-Rodríguez, V., Aiba, S.-I., Alvarez-Loayza, P., Alves, L. F., Ashton, P., Balvanera, P., Bastian, M. L., Bellingham, P. J., Berg, E. v. d., Bernacci, L., Bispo, P. d. C., Blanc, L., Böhning-Gaese, K., Boeckx, P., Bongers, F., Boyle, B., Bradford, M., Brearley, F. Q., . . . Venticinque, E. M. (2015). An estimate of the number of tropical tree species.

Proceedings of the National Academy of Sciences, *112*(24), 7472-7477. https://doi.org/doi:10.1073/pnas.1423147112

Soares-Filho, B., Moutinho, P., Nepstad, D., Anderson, A., Rodrigues, H., Garcia, R., Dietzsch, L., Merry, F., Bowman, M., Hissa, L., Silvestrini, R., & Maretti, C. (2010). Role of Brazilian Amazon protected areas in climate change mitigation. *Proceedings of the National Academy of Sciences*, *107*(24), 10821-10826. https://doi.org/doi:10.1073/pnas.0913048107

Soberón, J., Jiménez, R., Golubov, J., & Koleff, P. (2007). Assessing completeness of biodiversity databases at different spatial scales. *Ecography (Copenhagen)*, *30*(1), 152-160. https://doi.org/10.1111/j.2006.0906-7590.04627.x

Soberón, J., & Peterson, T. (2004). Biodiversity informatics: Managing and applying primary biodiversity data. *Philosophical transactions. Biological sciences*, *359*(1444), 689-698. https://doi.org/10.1098/rstb.2003.1439

Sousa-Baena, M. S., Garcia, L. C., Peterson, A. T., & Brotons, L. (2014). Completeness of digital accessible knowledge of the plants of Brazil and priorities for survey and inventory. *Diversity and Distributions*, *20*(4), 369-381. https://doi.org/10.1111/ddi.12136

Stropp, J., Ladle, R. J., M. Malhado, A. C., Hortal, J., Gaffuri, J., H. Temperley, W., Olav Skøien, J., & Mayaux, P. (2016). Mapping ignorance: 300 years of collecting flowering plants in Africa. *Global Ecology and Biogeography*, *25*(9), 1085-1096. https://doi.org/10.1111/geb.12468

Telenius, A. (2011). Biodiversity information goes public: GBIF at your service. *Nordic journal of botany*, *29*(3), 378-381. https://doi.org/10.1111/j.1756-1051.2011.01167.x

The Nature Conservancy. (2005). *Alaska-Yukon arctic ecoregional assessment*. The Nature Conservancy. https://arcticatlas.geobotany.org/catalog/dataset/05f55a78-b874-4ba9-850c-e3241b1bf890/resource/c7811276-3266-475d-b3c4-ca2380c0da3f/download/tnc-arctic-update-2-northern-alaska-ecosystems-map-2003.pdf

The Nature Conservancy. (2006). *The North Atlantic coast ecoregional assessment, conservation status report & resource CD 2006*. The Nature Conservancy. https://www.conservationgateway.org/ConservationByGeography/NorthAmerica/UnitedStates/edc/Documents/NAC-execsum.pdf

Tobler, M., Honorio, E., Janovec, J., & Reynel, C. (2007). Implications of collection patterns of botanical specimens on their usefulness for conservation planning: An example of two neotropical plant families (Moraceae and Myristicaceae) in Peru. *Biodiversity and Conservation*, *16*(3), 659-677. https://doi.org/10.1007/s10531-005-3373-9

UNCC. (2021). *COP26: The Glasgow climate pact* https://ukcop26.org/wp-content/uploads/2021/11/COP26-Presidency-Outcomes-The-Climate-Pact.pdf

Wikramanayake, E., Manandhar, A., Bajimaya, S., Nepal, S., Thapa, G., & Thapa, K. (2010). Chapter 10 - The Terai Arc landscape: A tiger conservation success story in a human-dominated landscape. In R. Tilson & P. J. Nyhus (Eds.), *Tigers of the world (Second Edition)* (pp. 163-173). William Andrew Publishing. https://doi.org/https://doi.org/10.1016/B978-0-8155-1570-8.00010-4

Wilson, E. O. (2016). *Half-earth: our planet's fight for life*. WW Norton & Company.

Yang, W., Ma, K., & Kreft, H. (2014). Environmental and socio-economic factors shaping the geography of floristic collections in China. *Global Ecology and Biogeography*, *23*(11), 1284-1292. https://doi.org/https://doi.org/10.1111/geb.12225

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*(11), e1124. https://doi.org/10.1371/journal.pone.0001124

Zizka, A., Antunes Carvalho, F., Calvente, A., Rocio Baez-Lizarazo, M., Cabral, A., Coelho, J. F. R., Colli-Silva, M., Fantinati, M. R., Fernandes, M. F., Ferreira-Araújo, T., Gondim Lambert Moreira, F., Santos, N. M. C., Santos, T. A. B., Dos Santos-Costa, R. C., Serrano, F. C., Alves da Silva, A. P., de Souza Soares, A., Cavalcante de Souza, P. G., Calisto Tomaz, E., . . . Antonelli, A. (2020). No one-size-fits-all solution to clean GBIF. *PeerJ*, *8*, e9916. 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*(5), 744-751. https://doi.org/https://doi.org/10.1111/2041-210X.13152

Supporting Information

Appendix S1

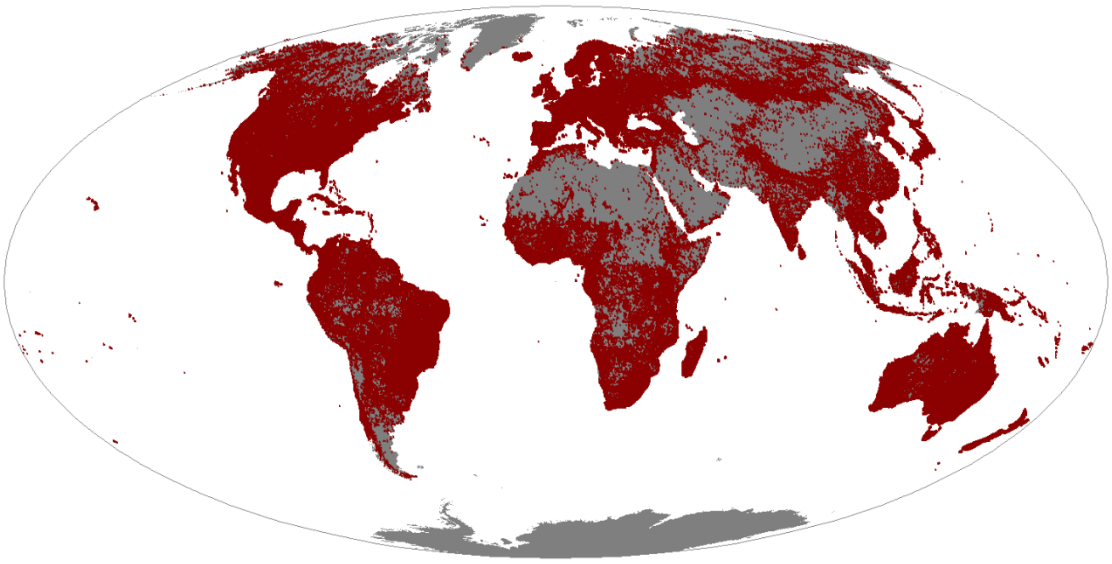

Figure S1.1 Spatial distribution of tree occurrence records.

Appendix S2

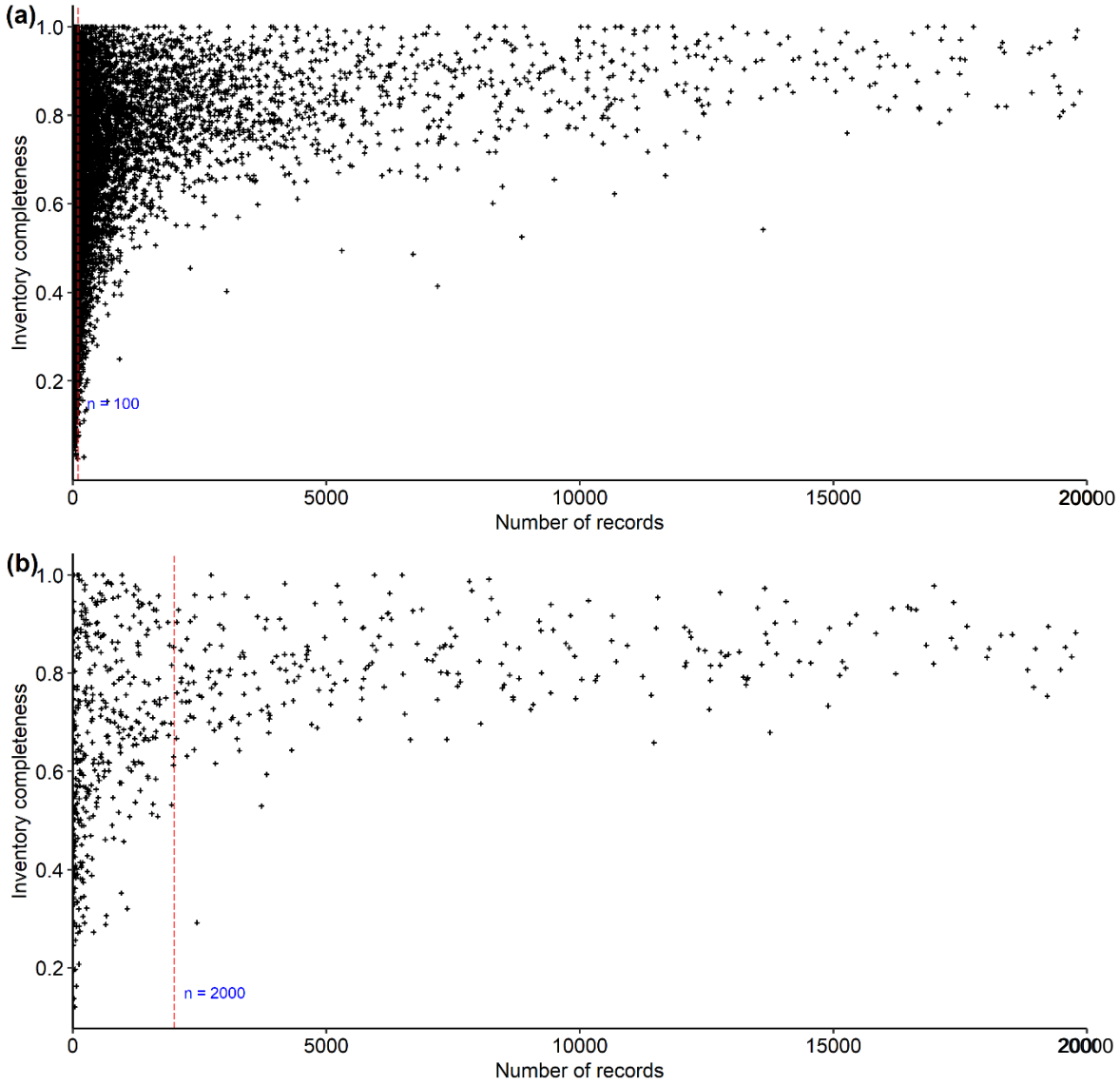

Figure S2.2 Scatter plots of the number of records versus inventory completeness for (a) ecoregions and (b) SUs. Red dashed lines represent the median sample size for ecoregions (rounded to the nearest thousand) and SUs (rounded to the nearest hundred).

Appendix S3

Table S2.1 Summary of sampling effort and inventory completeness at a biome level. SD = Standard deviation, *C* = inventory completeness.

Biome name and number	No. of ecoregions	No. of SUs	Total records	Mean sampling effort	Sampling effort SD	Mean richness	Richness SD	Mean estimated richness	Estimated richness SD	Mean *C* value	*C* value SD	Mean unobserved species	Unobserved species SD
Forested biomes													
1. Tropical & subtropical moist broadleaf forests	230	1945	2,603,713	0.36	0.74	939	881	1175	1003	0.74	0.18	236	175
2. Tropical & subtropical dry broadleaf forests	56	410	449,305	0.33	0.53	566	502	749	595	0.71	0.15	183	129
3. Tropical & subtropical coniferous forests	15	73	208,563	0.38	0.33	635	609	783	674	0.72	0.19	148	88
4. Temperate broadleaf & mixed forests	83	1257	13,526,937	1.40	2.72	176	197	206	241	0.88	0.13	29	53
5. Temperate conifer forests	47	392	976,222	0.35	0.75	146	215	176	263	0.83	0.13	31	52
6. Boreal forests/taiga	26	1553	970,755	0.02	0.08	40	19	45	24	0.90	0.10	6	8
14. Mangroves	19	34	58,579	0.22	0.30	530	469	802	616	0.60	0.17	273	221
Non-forested biomes													
7. Tropical & subtropical grasslands, savannas & shrublands	58	2139	1,270,955	0.11	0.18	662	760	819	994	0.79	0.14	156	262
8. Temperate grasslands, savannas & shrublands	48	1058	484,732	0.14	0.33	101	97	128	126	0.78	0.16	27	39
9. Flooded grasslands & savannas	25	110	29,907	0.06	0.14	123	158	213	215	0.59	0.23	90	98
10. Montane grasslands & shrublands	46	475	161,310	0.34	1.29	176	262	257	375	0.67	0.16	80	120
11. Tundra	51	836	126,107	0.02	0.08	16	11	19	13	0.86	0.14	3	4
12. Mediterranean forests, woodlands & scrub	40	332	4,956,429	1.51	3.19	144	99	166	114	0.87	0.12	22	22
13. Deserts & xeric shrublands	102	2645	412,422	0.07	0.16	138	171	196	230	0.69	0.18	58	71

Appendix S4

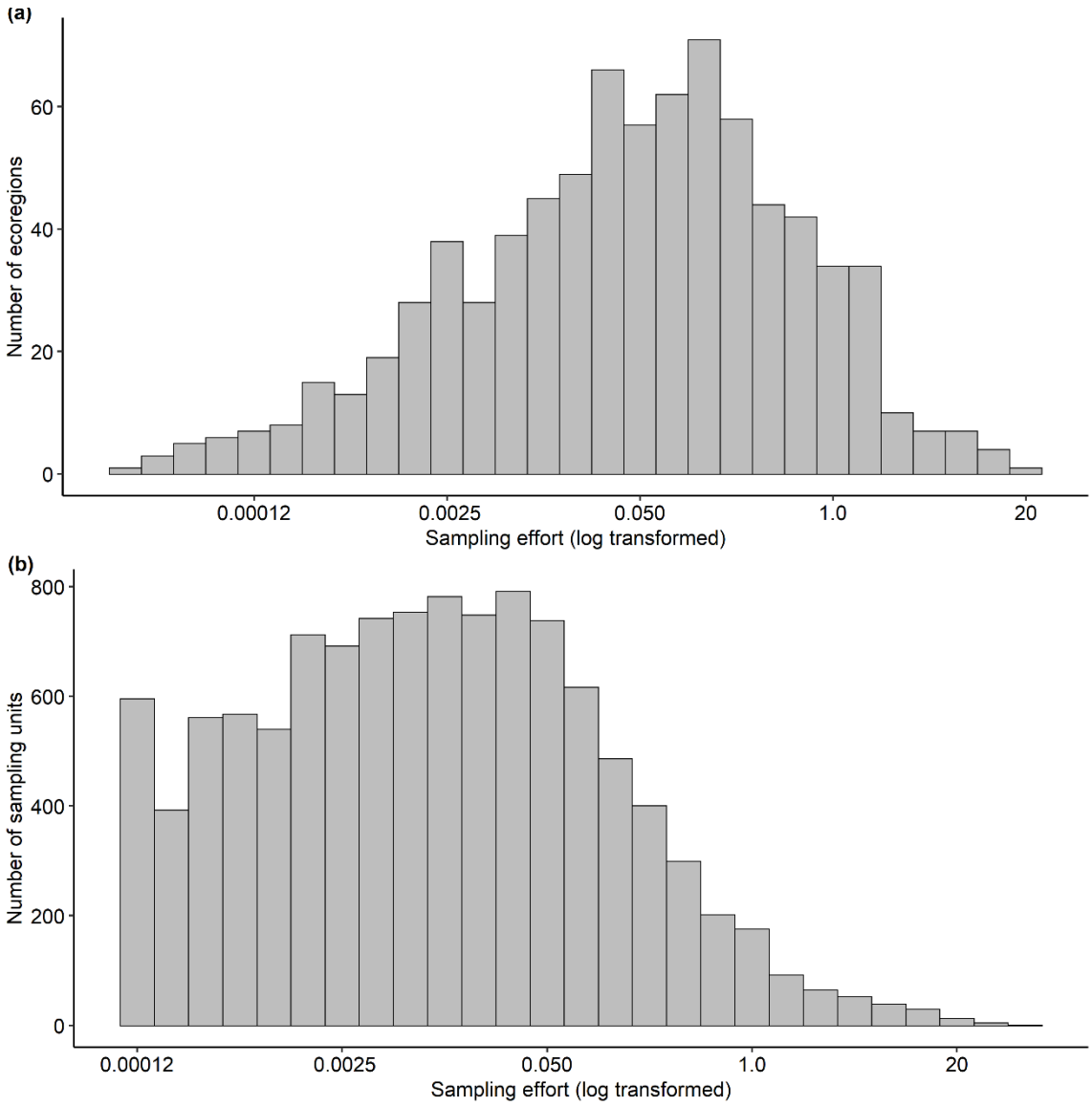

Figure S4.3 Frequency histogram of sampling effort for (a) ecoregions and (b) SUs (x-axis is stretched on a natural log scale).

Appendix S5

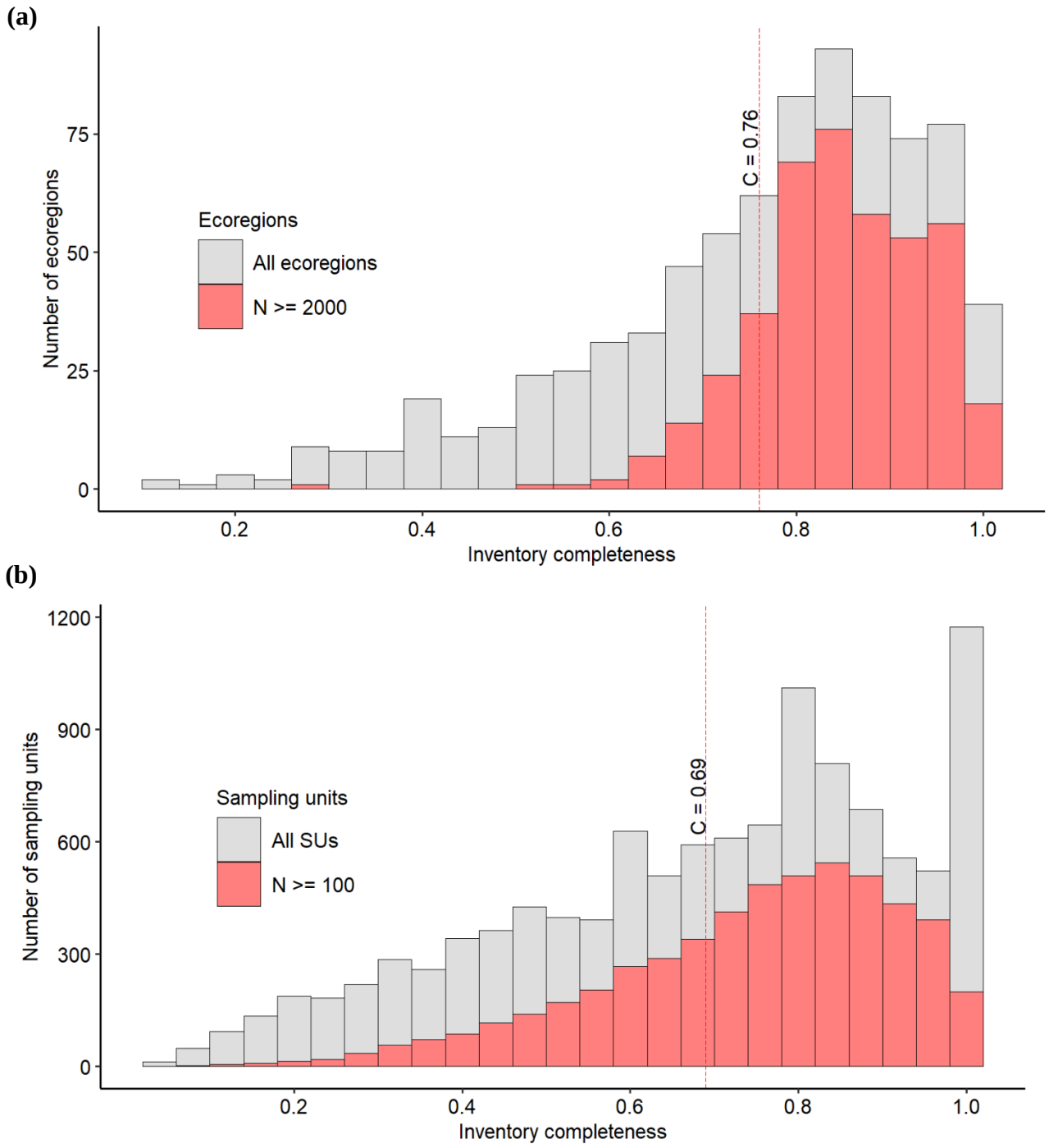

Figure S5.4 Frequency histogram of tree inventory completeness (*C*) for (a) ecoregions (mean *C* is 0.76 ± 0.18) and (b) 100 x 100 km SUs (mean *C* is 0.69 ± 0.24). N is the number of occurrence records.