Global Ecology and Biogeography, (Global Ecol. Biogeogr.) (2016) **25,** 1085–1096

Global Ecology and Biogeography A Journal of Macroecology

RESEARCH PAPER

Mapping ignorance: 300 years of collecting flowering plants in Africa

Juliana Stropp[1]*, Richard J. Ladle[2,3], Ana C. M. Malhado[2], Joaquín Hortal[4], Julien Gaffuri[5], William H. Temperley[1], Jon Olav Skøien[6] and Philippe Mayaux[7]

[1]Resource and Management Unit,, Joint Research Centre, European Commission, Via Enrico Fermi 2749, Ispra I-21027, Italy, [2]Institute of Biological and Health Sciences, Federal University of Alagoas, Av. Lourival Melo Mota, s/n, Tabuleiro do Martins, Maceió, AL 57072-900, Brazil, [3]School of Geography and the Environment, Oxford University, South Parks Road, Oxford OX1 3QY, 11, UK, [4]Departamento de Biogeografia y Cambio Global, Museo Nacional de Ciencias Naturales (MNCN CSIC), C/Jose Gutierrez Abascal 2, Madrid 28006, Spain, [5]Eurostat, European Commission, Rue Alphonse Weicker, 2721, Luxembourg, [6]Climate Risk Management Unit, Joint Research Centre, European Commission, Via Enrico Fermi 2749, Ispra I-21027, Italy, [7]Climate Change, Environment, Natural Resources Unit, International Cooperation and Development, European Commission, L-41/02-28 B-1049, Brussels, Belgium

*Correspondence: Juliana Stropp, Institute of Biological and Health Sciences (ICBS), Federal University of Alagoas, Maceió, AL, Brazil.
E-mail: justropp@gmail.com

ABSTRACT

Aim Spatial and temporal biases in species-occurrence data can compromise broad-scale biogeographical research and conservation planning. Although spatial biases have been frequently scrutinized, temporal biases and the overall quality of species-occurrence data have received far less attention. This study aims to answer three questions: (1) How reliable are species-occurrence data for flowering plants in Africa? (2) Where and when did botanical sampling occur in the past 300 years? (3) How complete are plant inventories for Africa?

Location Africa.

Methods By filtering a publicly available dataset containing 3.5 million records of flowering plants, we obtained 934,676 herbarium specimens with complete information regarding species name, date and location of collection. Based on these specimens, we estimated inventory completeness for sampling units (SUs) of 25 km × 25 km. We then tested whether the spatial distribution of well-sampled SUs was correlated with temporal parameters of botanical sampling. Finally, we determined whether inventory completeness in individual countries was related to old or recently collected specimens.

Results Thirty-one per cent of SUs contained at least one specimen, whereas only 2.4% of SUs contained a sufficient number of specimens to reliably estimate inventory completeness. We found that the location of poorly sampled areas remained almost unchanged for half a century. Moreover, there was pronounced temporal bias towards old specimens in South Africa, the country that holds half of the available data for the continent. There, high inventory completeness stems from specimens collected several decades ago.

Main conclusions Despite the increasing availability of species occurrence data for Africa, broad-scale biogeographical research is still compromised by the uncertain quality and spatial and temporal biases of such data. To avoid erroneous inferences, the quality and biases in species-occurrence data should be critically evaluated and quantified prior to use. To this end, we propose a quantification method based on inventory completeness using easily accessible species-occurrence data.

Keywords

Africa, data quality, flowering plants, GBIF, inventory completeness, spatial and temporal biases, species-occurrence data.

DOI: 10.1111/geb.12468
http://wileyonlinelibrary.com/journal/geb

INTRODUCTION

Botanical explorations over the past centuries have enormously increased our knowledge of biodiversity. Much of this knowledge is now accessible through Biodiversity Information Systems (BIS), providing new opportunities for conservation planning (Saarenmaa & Nielsen, 2002; Jetz *et al.*, 2012; Dubois *et al.*, 2015). However, biodiversity data typically show a patchy distribution due to the various purposes of biodiversity surveys, including long-term monitoring of particular sites or targeted interest in a few selected taxa only (ter Steege *et al.*, 2011). Furthermore, biodiversity data suffer from incompleteness and partially erroneous reporting because of the complexities involved in their collection and documentation (Hortal *et al.*, 2007, 2015; Rocchini *et al.*, 2011). In particular, species-occurrence data derived from specimens stored in natural history collections are known to suffer from uncertain quality and spatial and temporal biases (Nelson *et al.*, 1990; Boakes *et al.*, 2010; Anderson, 2012). The resulting knowledge shortfalls, if not properly addressed, limit the usefulness of BIS for biogeographical research and conservation planning (Soberón & Peterson, 2004; Hortal *et al.*, 2015). First, inaccuracies in the species-occurrence data themselves, for example incorrect taxonomic identification or incomplete labelling, curb the accurate prediction of species distribution (Anderson, 2012). Second, spatial bias in species-occurrence data restricts high inventory completeness to a few well-sampled regions (Hortal *et al.*, 2008), usually characterized by political stability, accessibility and proximity to research centres (Amano & Sutherland, 2013). Third, knowledge about species occurrence is expected to be particularly limited in areas progressively distant from well-sampled regions (Ladle & Hortal, 2013). This expectation is based on the principle of distance-decay of similarity in community composition (Nekola & White, 1999): environmental gradients and dispersal limitations cause more geographically distant communities to share a lower number of species than communities in close proximity. Finally, temporal biases towards old specimens render an inaccurate representation of the actual species distribution because changes in the landscape driven by habitat degradation, land cover and climate change tend to modify species assemblages (Ladle & Hortal, 2013).

Although spatial biases in species-occurrence data are frequently assessed (see Schulman *et al.*, 2007 and references therein; Sousa-Baena *et al.*, 2014; Yang *et al.*, 2014; Engemann *et al.*, 2015), temporal biases and the overall quality of species-occurrence data have received far less attention (Boakes *et al.*, 2010). Here we address this issue by assessing the quality of and spatial and temporal biases in species-occurrence data of flowering plants in Africa. Specifically, we address three questions: (1) How reliable are species-occurrence data of flowering plants in Africa that are easily available through BIS? (2) Where and when did botanical sampling occur in Africa over the past 300 years? (3) How complete is the plant inventory for Africa? We focus on a single continent (Africa) because it is more straightforward to identify some of the main factors that affect the historical acquisition, quality and distribution of biodiversity data. Moreover, Africa harbours a diverse and rich flora and has been subject to a long history of botanical sampling.

METHODS

Quality and coverage of available species-occurrence data

We retrieved 3,546,206 records from the global biodiversity information facility (GBIF, on 10 October 2012), including preserved specimens, living specimens, observations, fossils and germplasm of flowering plants collected in Africa. All records were screened by applying data filtering, as follows. First, we selected records labelled as 'basis of record = preserved specimen' in GBIF, with complete information regarding the date, latitude and longitude of collection as well as the species name. We selected only specimens that had at least two strings in the field 'scientific name interpreted'; in this way we included only specimens representing species or subspecies, but we eliminated specimens representing genera or families.

Second, we selected specimens flagged by GBIF as being free of georeferencing errors. We then visualized points of occurrence for specimens from each data provider individually. This procedure excluded specimens that were georeferenced to the country centroid and records with coordinates showing a dubious spatial pattern. Moreover, we identified and eliminated specimens for which the 'country' field was filled with countries located outside Africa. We also identified specimens for which the country field was attributed to 'unknown'. For these specimens, we searched their species name on the platform of the Missouri Botanical Garden (MOBOT, 2013) and eliminated those specimens for which the species name was registered as only occurring outside Africa.

Third, we excluded duplicate specimens. Collecting duplicate specimens of the same individual plant is a common practice in botany. These duplicate specimens are often distributed to several herbaria to help expand the coverage of collections and to provide backup security for the scientific information. There are four attributes of a voucher specimen that together can be used to identify duplicate specimens: (1) species identity, (2) date of collection, (3) geographical coordinates and (4) the name of the collector. Here, we defined duplicates by screening for unique combinations of species name, date of collection and location within proximity of 0.25° latitude and longitude. We did not consider 'name of collector' because this attribute is not yet standardized on the GBIF database: a single collector may be represented in the GBIF database with different spellings. Standardizing the name of collectors in our dataset was not feasible given the large number of specimens. Our choice of identifying duplicates located within 0.25° latitude and longitude follows the resolution with which data providers georeferenced their specimens. For instance, the data provider PRECIS (South Africa) georeferenced their specimens using a grid of 0.25° resolution. We considered it unlikely that the same species would be collected twice on the same day at locations less than 0.25° away from each other.

Fourth, we assessed the validity of 52,537 taxa names by submitting all names to the Taxonomic Name Resolution Service (TNRS) version 3.2 (Boyle *et al.*, 2013) in May 2013. We selected only specimens for which the names of species and subspecies matched those provided by the TNRS with an overall match score of >0.9. Match scores provided by the TNRS output range from zero to one, where one indicates a complete match between the string to be checked and a valid taxa name in the core database and a score of zero indicates no match. For brevity, we use the term species to refer to both species and subspecies in our dataset. Last, we selected only specimens sampled within the African coastline obtained from a digital elevation model using data from the Shuttle Radar Topography Mission (SRTM).

Inventory completeness

We estimated the inventory completeness of flowering plants on the entire African continent by defining sampling units (SUs) of 25 km × 25 km. For each individual SU, we considered the cumulative number of specimens and species collected from 1700 until 2012. We estimated inventory completeness based on Sousa-Baena *et al.* (2014). We obtained the number of sampling events for each individual SU. Each sampling event is a unique combination of the location (i.e. latitude and longitude) where a specimen was collected and its date of collection. We then obtained the number of species observed in each sampling event. Subsequently, Sousa-Baena's estimate of inventory completeness was calculated with the following equation:

$$C_i = \frac{S_{\mathrm{obs},i}}{S_{\mathrm{obs},i} + (a_i^2/2b_i)}$$

where C_i is the estimated inventory completeness for SU *i*; $S_{\mathrm{obs},i}$ is the number of species observed in SU *i*; a_i and b_i represent the number of species observed in one sampling event and the number of species observed in two sampling events in SU *i*, respectively. C_i ranges from zero to one, with one indicating a complete inventory. SUs with a small number of records may present artefactual values of *C*. This is because random effects may change *a* and *b*, causing estimates of *C* to be unstable in SUs with a small number of records (Sousa-Baena *et al.*, 2014). To define the range at which values of *C* are stable, and thus more reliable, we assessed the relationship between *C* and number of unique records (i.e. a unique combination of date and location of collection and species name). We found that a monotonic relationship exists above 200 unique records; therefore, in the main text we present estimates of inventory completeness for SUs that have more than 200 unique records. We present estimates of inventory completeness for SUs with ≥50 specimens in Fig. S1 in the Supporting Information.

We verified the results obtained from this method through two complementary approaches. First, we estimated inventory completeness based on Chao & Jost (2012), who proposed obtaining sample coverage from field-based biodiversity inventories. Here, we adapt their approach as follows:

$$K_i = \frac{f_1}{n_i}\left[\frac{(n_i-1)f_{1i}}{(n_i-1)f_{1i}+2f_{2i}}\right]$$

where K_i is the estimated inventory completeness; n_i, f_{1i} and f_{2i} are, respectively, the numbers of specimens, singletons and doubletons found in SU *i*. K_i ranges from zero to one, with one indicating a complete inventory.

Second, we used the curvilinearity of smoothed species accumulation curves (SACs) as a proxy for inventory completeness (Hortal *et al.*, 2004, 2008, Yang *et al.*, 2013). We calculated smoothed SACs with the method 'exact' of the function 'specaccum' in the R package vegan (Oksanen *et al.*, 2013). The mean slope of the last 10% of SACs obtained for each SU (hereafter referred to r_i) was used to estimate inventory completeness (Yang *et al.*, 2013). A flat slope (i.e. r_i values close to zero) indicates saturation in the sampling and thus a high inventory completeness. To convert the estimated r_i values into a normalized scale from zero to one, with values approaching one indicating a complete inventory, we subtracted the number one from the value of the slope parameter r_i obtained for each SU (i.e. $R_i = 1 - r_i$). We quantified the congruence of the results obtained from the three methods described above by applying a modified *t*-test that is suitable for quantifying the correlation of spatial variables (Clifford *et al.*, 1989).

In a next step, we calculated the geographical distance (in km) between all SUs and the closest well-sampled SU, characterized by having at least 200 unique records and $C_i \geq 0.5$. We excluded North Africa from this analysis because this region is environmentally more homogeneous than sub-Saharan Africa. We expect that with increasing distance to well-sampled SUs, floristic and environmental similarity decrease. The geographical distance to well-sampled SUs could help to identify areas where deficits in species-occurrence data persist.

Spatio-temporal distribution of inventory completeness

To determine whether inventory completeness in individual countries emerges from recently collected or historically old specimens we applied a modified *t*-test (Clifford *et al.*, 1989) correlating C_i with the median of the year in which specimens were collected in each SU *i*.

We performed Moran's I test to determine whether the spatial clustering of well-sampled SUs (i.e. number of unique records ≥ 200; $C_i \geq 0.5$) was related to the decade in which inventory completeness had been reached. For this, we calculated C_i using the cumulative number of specimens and species observed in each SU. We started by considering the period from 1700 to 1900, which we successively expanded by 10 years each, covering 11 distinct time periods (i.e. 1700–1910, 1700–1920, . . ., 1700–2000, 1700–2012). Additionally, Moran's *I* statistic was applied to assess whether

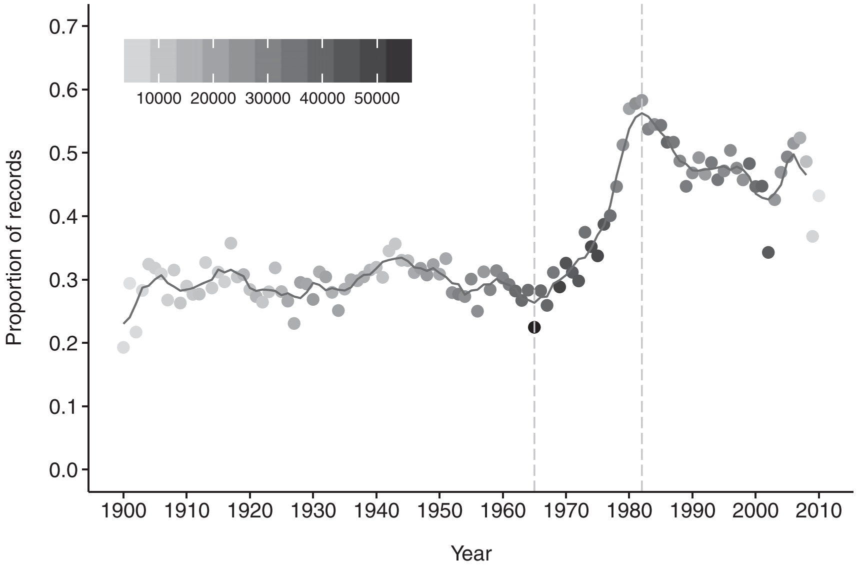

Figure 1 Proportion of records retained after filtering over. Vertical dashed lines depict the period of intense data collection and improvement in the quality of species-occurrence data (i.e. a high number of specimens retained after data filtering). Shades of grey indicate the number of records in each year prior to data filtering. For detailed description of criteria used to select records free of errors and with complete information see Methods.

spatial clustering of well-sampled SUs was related to the time span for which a SU has been subject to botanical sampling. Here we obtained for each well-sampled SU the interquartile range (IQR) of the year in which specimens were collected.

All analyses were carried out with R (R Core Team, 2015), with the exception of that on the geographical distance between all SUs and the closest well-sampled SU, which was generated in QGIS (QGIS Development Team, 2009). Geographical data were manipulated with the package rgdal (Bivand *et al*., 2015). The modified *t*-test was conducted using the function 'modified.ttest' of the package SpatialPack (Osorio *et al*., 2012) and Moran's I test was performed with the function 'Moran.I' of the package *ape* (Paradis *et al*., 2014). Maps were produced using QGIS. The R script containing the methods to calculate inventory completeness is given in Appendix S10.

RESULTS

Quality and coverage of available herbaria data

Data retrieved from GBIF as of 10 October 2012 contained 3,546,206 records, of which 90% (3,258,622) were specimens representing 97,335 taxon names, including family, genera, species and subspecies. Our data filtering led to an exclusion of 74% of the initial records and 52% of the initial taxon names. The first step of the data filtering process (i.e. the exclusion of specimens with missing information on collection date, latitude, longitude and taxon with one string in the field 'scientific name interpreted' excluded 59% (2,097,370) of the initial records, two-thirds of which lacked information on collection date. The exclusion of duplicates eliminated 24% of the remaining specimens (322,550). Of the 52,537 taxon names that were checked for validity, 90% (47,238) had an overall match score >0.9. Therefore, the exclusion of specimens due to incomplete or incorrect labelling, faulty georeferencing and lack of updated taxonomy precluded us from using 70% (2,200,980) of the data initially available. Our final dataset therefore, contained 934,676 specimens belonging to 47,238 species (or subspecies), collected in 57 countries in the period from 1700 to 2012 (Table S1).

Improvement in the quality of species-occurrence data was achieved predominantly between the mid-1960s and late 1970s, i.e. a large number of specimens collected during this period were retained after data filtering. This period coincides with intense data collection (Fig. 1). As a consequence, specimens collected after the 1960s contributed a greater proportion to the total number of specimens present in our final dataset (57%, 533,996) than the specimens collected in the previous 270 years. However, half of the total number of specimens collected in recent decades are still excluded due to erroneous or incomplete labelling, mainly as a result of lack of information on geographical coordinates.

We found a clear country and temporal bias in our dataset. The number of specimens collected in South Africa alone (512,680) surpassed the total number of specimens collected in all other African countries. Madagascar and Tanzania, both participating in GBIF, rank second (78,752) and third (50,694), respectively. A few countries, such as Cameroon and Gabon contribute a relatively large number of specimens (33,282 and 35,938; respectively) despite their small area and lack of participation in GBIF. In contrast, Congo and Uganda, both participating in GBIF, hardly provide any data on flowering plants (see Table S2). The 15 French-speaking countries in Africa have a greater number of collected specimens (197,548; 0.02 specimen per km^2) than the 17 English-speaking ones (174,578; 0.01 specimen km^{-2}), excluding South Africa.

Over the entire continent, the number of specimens collected per decade rose between the 1970s and 1980s and slightly decreased after the 1990s (Fig. 2, Table S3). Within each language group, a few countries contribute disproportionally to the total number of specimens. Madagascar holds 40% of the specimens collected in French-speaking countries, Tanzania 30% of the specimens collected in the English-speaking countries (excluding South Africa), Cameroon 96% of the specimens collected in the English–French-speaking

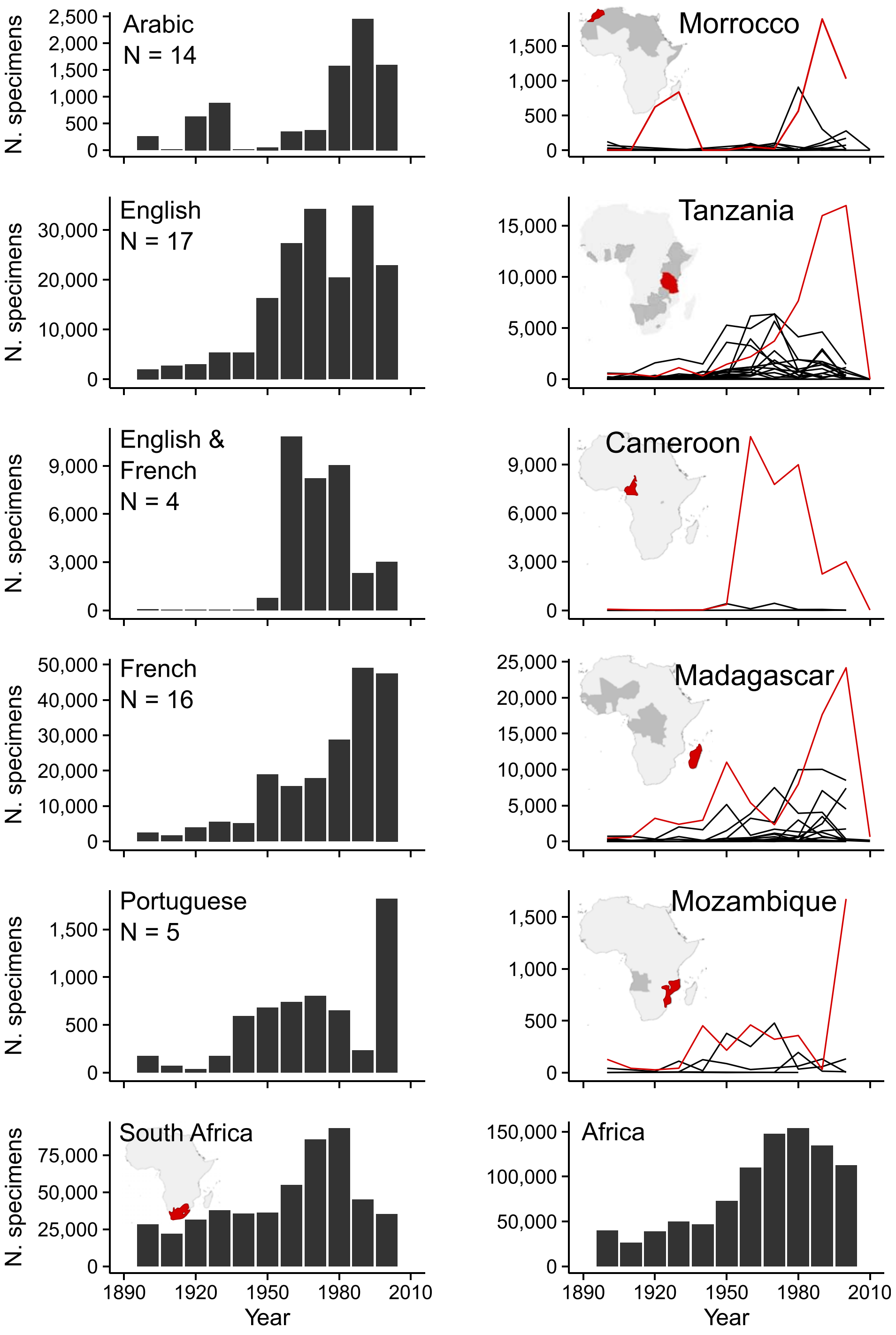

Figure 2 The number of specimens collected in Africa per 10-year period until 2012. Bars represent the total number of specimens collected in groups of countries sharing the same official language; lines represent the total number of specimens collected in individual countries grouped by official language. 'N' stands for the number of countries in each language group. Groups are presented in alphabetical order. Light grey areas in the maps indicate countries sharing the same official language, whereas red areas (and red lines) depict the country with the largest number of specimens in their language group.

countries, Mozambique 63% of specimens collected in Portuguese-speaking countries and Morocco 61% of specimens in Arabic-speaking countries.

Spatial and temporal distribution of inventory completeness

The estimates of inventory completeness in the three methods were strongly correlated with each other (Figs S2 & S3). Sousa-Baena's method may be less sensitive to variation in sampling evenness across SUs than the other two methods applied here because it accounts for the number of species observed in sampling events. For this reason, we choose to present in the following paragraphs the results based on Sousa-Baena *et al.* (2014) and refer reader to Tables S4 & S5 for an overview of the results obtained using the other two methods.

Out of a total of 41,985 SUs, 31% have been subject to floristic sampling (containing at least one specimen). Only 1002 (2.4%) SUs contained at least 200 unique records. The

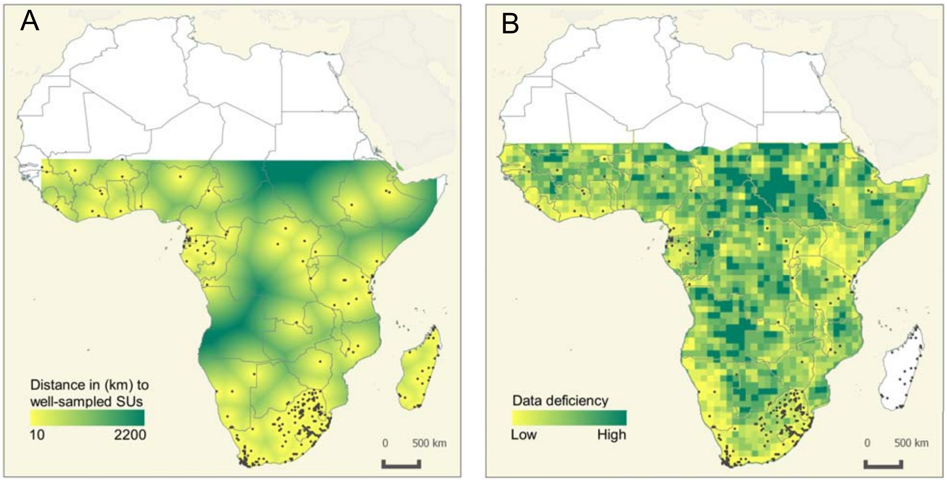

Figure 3 Maps of (a) distance to well-sampled sampling units (SUs), i.e. SUs with inventory completeness (number of unique records ≥ 200 and $C_i \geq 0.5$) obtained using 934,676 specimens and 47,238 species, and (b) data-deficient areas estimated as the ratio between documented and modelled number of species per 1° grid cell obtained using 185,427 specimens and 5873 species (adapted from Fig. 2B in Küper *et al.*, 2006). In both maps, black dots indicate well-sampled SUs determined in this study, deeper shades of green indicate regions with high deficit of species-occurrence data and white areas depict regions for which data deficiency was not estimated. Note that acute deficits of species-occurrence data remain, for example, in central Angola and the Democratic Republic of Congo and northern and southern Mozambique, despite the difference in the number of specimens used to generate the maps.

estimates of inventory completeness ranged from 0.07 to 0.75 with a median of 0.45 and an interquartile range (IQR) of 0.18. Although the overall number of specimens increased over the past 300 years, limited sampling effort and unevenness in spatial coverage restrict well-sampled SUs ($\geq$200 unique records and $C_i \geq 0.5$) to just 0.6% of Africa's territory (261 SUs). Furthermore, 58% of the well-sampled SUs (152 SUs) occur in South Africa. The share of well-sampled SUs is low even in countries with a large number of specimens: only 8% of SUs in South Africa are well sampled, dropping to 3% and 1% in Madagascar and Tanzania, respectively.

We identified acute data deficits in central and southeastern Africa (Fig. 3). In Angola and the Democratic Republic of Congo, the average distance to well-sampled areas is 620 and 340 km, respectively, with distances of about 800 km in areas bordering the two countries. In Mozambique, the average distance to well-sampled SUs is 250 km. The spatial pattern of data deficits depicted here closely resembles the map of data-deficient areas presented by Küper *et al.* (2006), which is based on 185,427 specimens and 5873 species.

We found that in South Africa inventory completeness was negatively correlated with the median year in which specimens were collected. This finding indicates that apparently high levels of inventory completeness emerge from data collected decades ago (Fig. 4). Furthermore, we found that spatial clustering of well-sampled SUs is related to temporal components of botanic inventory making: SUs geographically close to each other tend to reach inventory completeness around similar decades and experience a similar time span of botanic sampling (Moran's I = 0.10, $P < 0.001$ for a decade; Moran's I = 0.14, $P < 0.001$ for IQR). Clusters of well-sampled SUs in southern Africa seem to reach completeness earlier and share a longer history of botanic sampling than those in central and western Africa (Fig. 5). Maps and data presented here can be downloaded or viewed online at: http://rris.biopama.org/plant_completeness.

DISCUSSION

The increasing availability of species-occurrence data for Africa did not remove the substantial spatial biases in botanical sampling. As early as 1968, Léonard (1968) reported acute gaps in botanical sampling in Africa. Nearly, half a century later, and with tremendous increase in the number of specimens, the location of poorly sampled areas remains almost unchanged. Moreover, the pronounced temporal bias towards old specimens in South Africa, the country that holds half of the available data for the entire continent, implies that high inventory completeness is largely based on specimens collected several decades ago. In the following sections, we discuss our findings and highlight the importance of explicitly communicating limitations in the quality, coverage and longevity of species-occurrence data.

How reliable are the species-occurrence data available through BIS for Africa?

Incomplete or incorrect labelling, lack of updated taxonomy and faulty georeferencing precluded the use of 70% of the initially available data. The records excluded by our filtering are unsuitable for the analysis presented here; they may, however, still be appropriate for other purposes. Furthermore, it should be stressed that duplicate specimens (322,550) may not contain incomplete or incorrect labelling. We excluded them from our dataset due to methodological concerns,

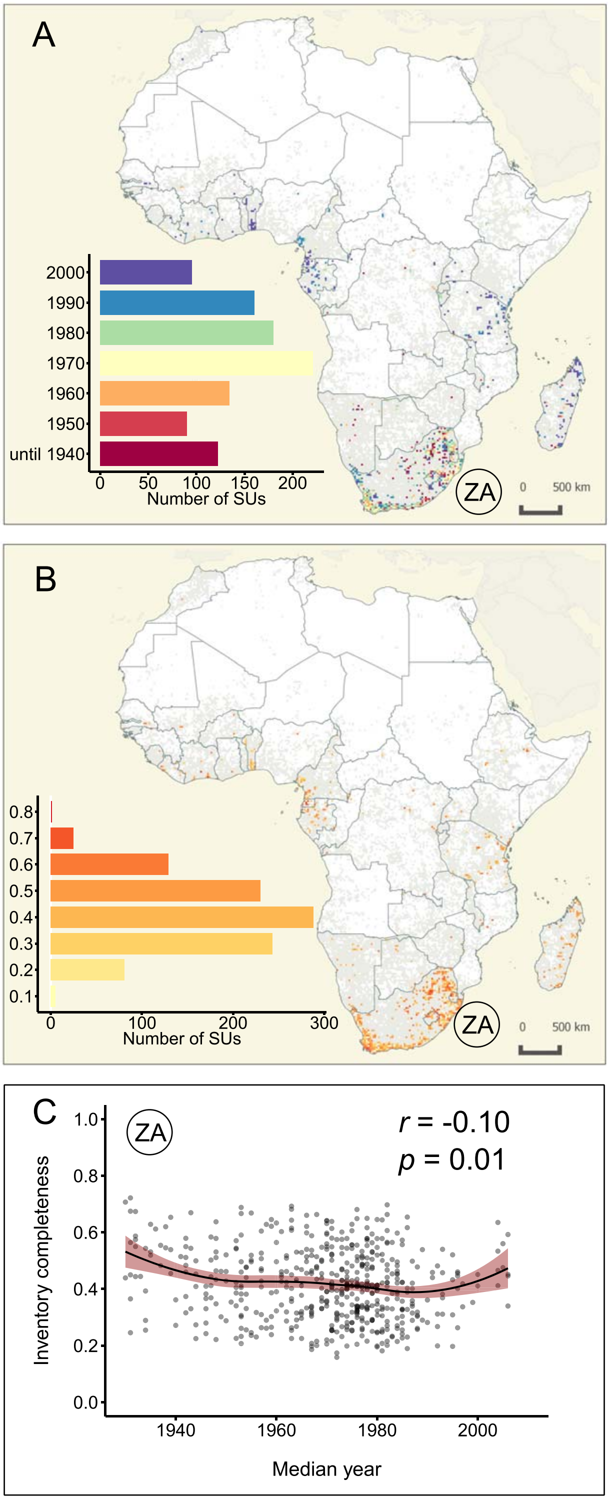

Figure 4 Maps of temporal and spatial biases in the African flowering plant inventory. (a) Median of the year in which specimens were collected; shades of blue indicate areas subject to recent botanical collections, whereas shades of red indicate old botanical collections. (b) Inventory completeness (according to the method of Sousa-Baena *et al.*, 2014); deeper shades of red indicate relatively well-sampled areas. In both maps, squares indicate sampling units of 25 km × 25 km; sampling units with less than 200 specimens are indicated in light grey, whereas areas with no data are indicated in white. Bar graphs represent frequency distribution of the median of the year in which specimens were collected (panel a) and of inventory completeness (panel b). (c) Scatterplot relate inventory completeness and median of year in which specimens were collected; r and p represent the correlation coefficient and significance level, respectively. Negative correlation observed for South Africa (ZA) suggest that inventory completeness emerges from botanical collections made decades ago.

including duplicate specimens in our estimates of inventory completeness could introduce noise into the results. The number of duplicate specimens of the same species is influenced by collector behaviour and species identity (ter Steege *et al.*, 2011); therefore, it may not approximate species abundance in the field.

Although errors in specimen labels were particularly common in older data, about half of the specimens collected in the past three decades were still subject to error or incomplete labelling. The occurrence of errors in specimen labels was reported for biodiversity-rich areas in South America, for example the Andes and the Amazon. However, the magnitude of the errors is rarely quantified (Hopkins, 2007; Boakes *et al.*, 2010). We argue here that quantifying and communicating errors in specimen labels constitutes the first step in ensuring quality control of species-occurrence data (see Maldonado *et al.*, 2015). Removing persisting errors may lead to the discovery of new species (Bebber *et al.*, 2010) and can substantially improve the reliability of subsequent species distribution modelling (Boakes *et al.*, 2010). If not resolved, errors in specimen labels limit the effectiveness of best practice guides and quality standards for digitizing herbaria data.

Where did botanical sampling occur in Africa over the past 300 years?

The spatial and temporal pattern of species-occurrence data reported here reflects the data that are easily available through GBIF, and consequently disregards complementary data sources such as inventory plots, field observations and local herbaria. Clearly, GBIF only represents a proportion of the available species-occurrence data and there may be systematic deviations in the willingness of countries to contribute to the global GBIF database. As of July 2015, only 13 out of 57 African countries were participating in GBIF. The top three countries (South Africa, Tanzania and Madagascar) of our analysis in terms of the total number of collected specimens are, unsurprisingly, all GBIF participants (see Table S2). It follows that our results for countries that do not participate in GBIF (e.g. Angola, Mozambique and Democratic Republic of the Congo) should be interpreted with caution; there, species-occurrence data may exist but are not easily available. For a few countries that are not participants in GBIF, external institutions may play an important role in

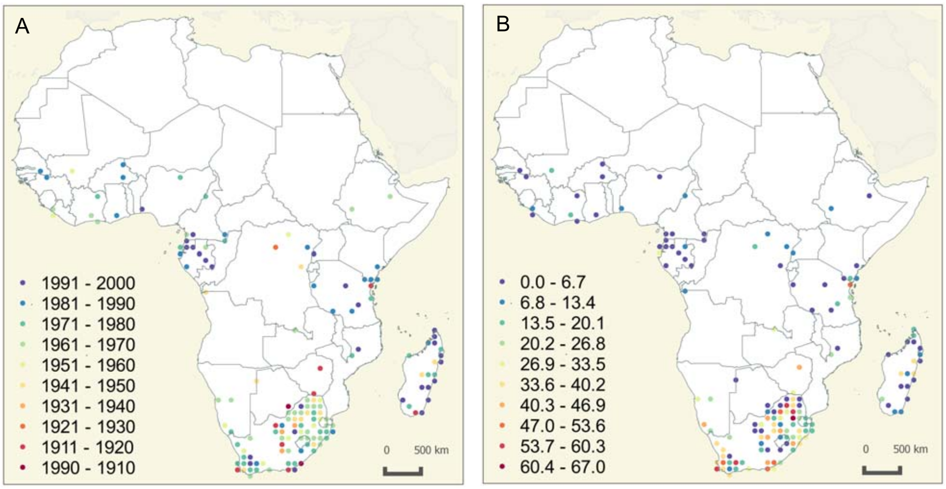

Figure 5 Maps of temporal attributes of well-inventoried sampling units (number of unique records ≥ 200 and $C_i \geq 0.5$; according to the method of Sousa-Baena *et al.*, 2014). Deeper shades of blue indicate (a) areas that have reached inventory completeness in recent decades and (b) areas that have experienced botanical inventory over a longer time span, i.e. those with higher interquartile range (IQR) of years in which specimens were collected. Note that sampling units in central and western Africa have more recently reached inventory completeness and experienced a shorter history of botanical sampling than southern Africa. To improve visualization, the 25 km × 25 km grid used in the analysis was aggregated to 1° resolution by calculating the median of (a) the decade in which inventory completeness was reached and (b) IQR.

providing species-occurrence data to GBIF as suggested by the cases of Cameroon and Gabon—neither of these participates in GBIF, yet a relatively high number of specimens are provided to GBIF.

Our results indicated distinct country-level biases in the species-occurrence data, with specimens identified in South Africa making up 54% of records for the entire continent. There are many factors that have contributed to this observation and it is difficult to disentangle their effects without an in-depth, qualitative analysis of socio-economic and scientific factors. It is likely that the strong historical, cultural and academic ties between South Africa and the UK, a well-developed university system and the presence of the highly distinctive Cape phytogeographical region (Cox, 2001) have all played a role in making South Africa pre-eminent. South Africa also hosts 55 herbaria, more than any other African country (Thiers, 2014), although many of these are small.

Our finding that French-speaking countries in Africa possess a greater number of collected specimens than English-speaking countries (excluding South Africa) is generally concordant with previous studies. Specifically, in their study of inventory completeness of sphingid moths in Africa, Ballesteros-Mejia *et al.* (2013) observed that former French and Belgian regions are better sampled than former British and Portuguese regions. However, the observed association of sampling effort with language group is largely driven by a few countries that contribute disproportionately to the total number of specimens. Examples include Madagascar that holds 40% of the specimens collected in French-speaking countries and Tanzania that holds 30% of the specimens collected in English-speaking countries. In the case of Madagascar, a large number of specimens is derived from externally funded projects with no obvious link to the colonial history of the country (see below).

When did botanical sampling occur?

The peak of botanical sampling in Africa occurred in the 1970s and 1980s (Fig. 2). We observed an overall decrease in the number of specimens collected in the past three decades. Part of this observation may be explained by the inevitable time lag between data collection, digitization and publishing through GBIF (Gaiji *et al.*, 2013), which affects the availability of recently collected specimens. The overall decline of botanical sampling in Africa is strongly influenced by the decreased efforts in South Africa, as the country holds more than half of the specimens collected across the entire continent. Nevertheless, the recent decline in botanical sampling is a phenomenon shared among several countries (cf. Fig. 2). The general trend may reflect the global decline in the number of plant surveys following the decreasing number of active plant taxonomists in recent decades (Bebber *et al.*, 2010). In contrast to the general trend, a few countries such as Benin and Madagascar show a steep increase in plant collection over the last two decades. This trend might be partly related to ambitious externally funded research projects [see, for example, the Vahinala project in Madagascar (Missouri Botanical Garden, 2014) and the Biota West Africa for Benin (e.g., Assede *et al.*, 2012)].

How complete is the plant inventory for Africa?

The methods we used to estimate inventory completeness are sensitive to differences in the degree of evenness at which species are sampled within each SU. This issue can be illustrated using two contrasting scenarios of sampling evenness. In the first scenario, sampling effort is directed towards maximizing taxonomic representation without capturing the relative abundances of species (i.e. as typically done in collecting trips; see ter Steege *et al.*, 2011). In this case, inventory completeness may be underestimated due to overrepresentation of rare species in SUs. In the second scenario, sampling effort is directed towards sampling abundant species (as is typically done in monitoring programmes). In such a situation, inventory completeness may be overestimated because of underrepresentation of rare species in the SU. In our opinion, such a scenario is unlikely to happen as our estimates of inventory completeness are based on data provided by several herbaria. In fact, the number of specimens of the same species in individual SUs is generally low. For instance, the most collected species in the SU with the highest number of specimens, *Eragrostis curvula* (Poaceae), amounted to 0.5% of all specimens collected. A similar pattern is observed in SUs with a relatively low number of species per specimen: in a SU with 2777 specimens and 637 species, the most collected species (*Santiria trimera*; Burseraceae) accounts for 1.6% of all specimens (see Fig. S4). The overrepresentation of rare species in our dataset shows systematic underestimation of inventory completeness in both well-sampled and poorly sampled SUs, making the results for individual SUs comparable with each other. We regard our results as robust, but recommend that future case studies benchmark their estimates of inventory completeness based on herbarium data against those obtained from actual inventory plots.

Our assessment revealed that 31% of SUs contain at least one specimen, whereas only 0.6% of the SUs analysed here can be considered well-sampled, holding $\geq$200 unique records and reaching $C_i \geq 0.5$. Moreover, we find clear spatial aggregation of well-sampled SUs. These findings are not unique in a global context: it is well known that botanical sampling in the Amazon is geographically associated with larger cities, major rivers, roads and proximity to research centres (Nelson *et al.*, 1990; Hopkins 2007; Schulman *et al.*, 2007; Engemann *et al.*, 2015). Similarly, botanical sampling (and that of other biological groups) in Africa has been associated with accessibility and the availability of research infrastructure. Küper *et al.* (2006) observed that botanical sampling overlaps with the location of climate stations and sampling of passerine birds, which tend to be clustered near urban areas and transport infrastructure (see maps in New *et al.*, 2002; Reddy & Davalos, 2003). The spatial aggregation of well-sampled SUs produces a long gradient in deficit of species-occurrence data across the African continent, as indicated by the map of distance to well-sampled SUs (Fig. 3). The spatial pattern of data deficiencies (depicted in Fig. 3) closely resembles the map of data-deficient areas presented by Küper *et al.* (2006). Both maps depict acute deficits in species-occurrence data, for example in central and western Angola, Botswana, the central Democratic Republic of Congo and northern and southern Mozambique. Küper *et al.* (2006) noted that areas highlighted as data deficient had already been repeatedly identified back in 1968 and 1979 (see Fig. 3 and references in Küper *et al.*, 2006). They proposed that deficits in species-occurrence data could be more effectively reduced by targeting botanical sampling in poorly sampled areas rather than by digitizing already collected data that are likely to come from relatively well-sampled regions. Our results support this proposal: a five- and eight-fold increase in the number of specimens and species used here compared with the study of Küper *et al.* (2006), have done little to reduce the spatial bias in species-occurrence data. The persisting spatial bias is particularly problematic because it casts doubts on our current understanding of patterns of species diversity and distribution at a broader spatial scale (Küper *et al.*, 2006; Engemann *et al.*, 2015).

The spatio-temporal clustering of well-sampled SUs, as revealed by the Moran's I, reflects the history of botanical sampling in Africa. A causal explanation for this pattern deserves an in-depth analysis and is beyond the scope of this study. We therefore interpret our findings by focusing on specific examples, matching botanical sampling with scientific activities in individual countries. Gabon, for example, seems to have experienced a relatively short history of intensive botanical sampling that started in the 1960s, peaked in the 1990s and continued until the 2000s (see Table S3). Consequently, inventory completeness was predominantly reached around the 1980s and 1990s (Fig. 5). The intensification of botanical sampling in Gabon may have been driven by the project *Flore du Gabon*, which had its first volume published by the National Museum of Natural History (MNHN) of Paris in 1961 (Aubreville, 1961). As of February 2016, 45 volumes of *Flore du Gabon* have been published. The project *Flore du Gabon* is expected to be finished by 2019 and is now led by two external institutions: the Naturalis Biodiversity Center (the Netherlands) and the Botanic Garden Meise in Belgium (Botanic Garden Meise, 2014).

In South Africa, particularly in the vicinities of Cape Town and Johannesburg, SUs may have experienced botanical sampling over a period of up to 60 years; many reaching inventory completeness around the 1940s and 1970s (see Fig. 5). This observation reflects the long history of botanical exploration in South Africa. In particular, the Cape Region was one of the first areas outside Europe to be botanically explored, with its first botanical specimens dating from the 1700s (Goldblati, 1978). Botanical exploration in South Africa continued and intensified in the 1970s and 1980s, declining after the 1990s (see review in Crouch *et al.*, 2013). The *Flora of Southern Africa* was first published in 1963 (Codd *et al.*, 1963). Twenty years later, intense botanical sampling in southern Africa enabled the compilation of a sequence of regional plant checklists (Germishuizen & Meyer, 2003). As these became available, the National Herbarium in Pretoria (PRECIS) started producing a series of plant checklists for Southern Africa, including Botswana, Lesotho, Namibia, South Africa, Swaziland (Germishuizen & Meyer, 2003, and references therein). Although botanical sampling peaked in South Africa in the 1980s, perhaps due to the implementation of the International Biological Programme in the 1970s (Huntley, 1987), there were warnings that progress being made on the *Flora of Southern Africa* was already slowing down in the 1980s (Leistner, 1983). Despite these warnings, botanical sampling declined from 1990 onwards (see Fig. 2), probably driven by socio-economic changes.

Many SUs in southern Africa that are characterized by moderate or high completeness have a high frequency of specimens collected between the 1950s and 1970s. In view of accelerating land-use change and the subsequent loss of habitats, it is unclear whether the species collected several decades ago can still represent the set of species found on the ground today. This observation points to the risk of spurious or obsolete knowledge of species occurrence, i.e. the assumption that SUs are well sampled, while in fact the set of collected species has long been lost from that area (see Ladle & Hortal, 2013). The resulting temporal biases in inventory completeness can increase the probability of errors of commission in predictive species distribution models, highlighting the importance of data users being responsible for checking the date of collection of species-occurrence data.

CONCLUSIONS AND RECOMENDATIONS

Species-occurrence data are increasingly being applied to broad-scale biogeographical research and biodiversity conservation (Jetz *et al*., 2012; Dubois *et al*., 2015). To avoid erroneous inferences, it is crucial to scrutinize such data before they are used and explicitly communicate the related spatial and temporal biases (Hortal *et al*., 2015). Our results show that the quality and completeness of species-occurrence data of flowering plants in Africa, easily available through BIS, are low: 70% of the available species-occurrence data contained erroneous or incomplete information and only 1% of the SUs are relatively well sampled. To overcome this shortcoming, we suggest that BIS implement easy pathways for community feedback on data quality (see Maldonado *et al*., 2015). Moreover, we suggest disseminating best practices for collecting and vouchering botanical data, paying special attention to correct georeferencing and specification of the date of collection.

Our study identified pronounced temporal biases towards older specimens, particularly in South Africa, and persisting deficits in botanical sampling in Central Africa. Knowledge of species occurrence is necessarily limited and scale dependent (Hortal *et al*., 2015) and it is currently impossible to obtain very accurate species lists for relatively large areas (D'Alessandro & Fattorini, 2002). In practice, this implies that users of species-occurrence data should be informed of the quality and biases related to such data in order to be able to assess the associated uncertainty. Finally, our findings show that using only one metric (e.g. spatial bias) may be insufficient to communicate the many shortcomings inherent to species-occurrence data (Hortal *et al*., 2015). A more comprehensive view may be obtained by establishing several maps depicting a set of metrics related to quality and spatial and temporal biases in such data. Therefore, following the reasoning of Boggs (1949) and, more recently, Rocchini *et al*. (2011), Ladle & Hortal (2013) and Ruete (2015), we strongly recommend the systematic development of 'maps of ignorance' for biodiversity. Such maps can provide a complete measure of the reliability of species-occurrence data that are commonly available through BIS.

As discussed by Rocchini *et al*. (2011), only by knowing where we should trust (or doubt) our knowledge of species occurrence we will be able to make legitimate decisions using the results of species distribution models and where best to allocate limited resources for improving the quality and coverage of species-occurrence data.

ACKNOWLEDGEMENTS

This work was funded by the European Development Fund (EDF 10) project Biodiversity and Protected Areas Management (BIOPAMA, http://rris.biopama.org/). R.J.L. was funded by the Brazilian National Council for Scientific and Technological Development (CNPq; no. 310953/2014-6). J.H. was funded by a Spanish DGCyT Ramón y Cajal grant and by a CNPq Special Visiting Researcher grant (no. 401471/2014-4). We thank Wenjing Yang for sharing the R code for calculating SACs. We thank Martin Weiss and Martha B. Dunbar for their comments on earlier drafts of this article. Finally, we thank the exceedingly constructive comments we received from the editor Richard Field, referee Duccio Rocchini and two anonymous referees.

REFERENCES

Aubreville, A. (ed.) (1961) *Flore du Gabon*. Museum National d'Histoire Naturelle, Paris. Cited in Prance, G.T. (1977) Floristic inventory of the tropics: where do we stand? *Annals of the Missouri Botanical Garden*, **64**, 659–684.

Amano, T. & Sutherland, W.J. (2013) Four barriers to the global understanding of biodiversity conservation: wealth, language, geographical location and security. *Proceedings of the Royal Society B: Biological Sciences*, **280**, 20122649.

Anderson, R.P. (2012) Harnessing the world's biodiversity data: promise and peril in ecological niche modeling of species distributions. *Annals of the New York Academy of Sciences*, **1260**, 66–80.

Assede, E.P., Adomou, A.C. & Sinsin, B. (2012) Magnoliophyta, biosphere reserve of Pendjari, Atacora Province, Benin. *Check List*, **8**, 642–661.

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**, 586–595.

Bebber, D.P., Carine, M.A., Wood, J.R., Wortley, A.H., Harris, D.J., Prance, G.T., Davidse, G., Paige, J., Pennington, T.D & Robson, N.K. (2010) Herbaria are a major frontier for species discovery. *Proceedings of the National Academy of Sciences USA*, **107**, 22169–22171.

Bivand, R., Keitt, T. & Rowlingson, B. (2015) rgdal: Bindings for the Geospatial Data Abstraction Library. R package version 1.1-3. available at: https://CRAN.R-project.org/package=rgdal

Boakes, E.H., McGowan, P.J., Fuller, R.A., Chang-Qing, D., Clark, N.E., O'connor, K. & Mace, G.M. (2010) Distorted views of biodiversity: spatial and temporal bias in species occurrence data. *PLoS Biology*, **8**, e1000385.

Boggs, S.W. (1949) An atlas of ignorance: a needed stimulus to honest thinking and hard work. *Proceedings of the American Philosophical Society*, **93**, 253–258.

Botanic Garden Meise (2014) *Botanic Garden Meise: annual report 2014*.

Boyle, B., Hopkins, N., Lu, Z., Garay, J.A.R., Mozzherin, D., Rees, T., Matasci, N., Narro, M.L., Piel, W.H. & Mckay, S.J. (2013) The taxonomic name resolution service: an online tool for automated standardization of plant names. *BMC Bioinformatics*, **14**, 16.

Chao, A. & Jost, L. (2012) Coverage-based rarefaction curve and extrapolation: standardizing samples by completeness rather than size. *Ecology*, **93**, 2533–2547.

Clifford, P., Richardson, S. & Hemon, D. (1989) Assessing the significance of the correlation between two spatial processes. *Biometrics*, **45**, 123–134.

Codd, L.E., De Winter, B., Rycroft, H.B. (1963) Flora of Southern Africa. Government Printer, Pretoria. Cited in Prance, G.T. (1977) Floristic inventory of the tropics: where do we stand? *Annals of the Missouri Botanical Garden*, **64**, 659–684.

Cox, B. (2001) The biogeographic regions reconsidered. *Journal of Biogeography*, **28**, 511–523.

Crouch, N.R., Smith, G F. & Figueiredo, E. (2013) From checklists to an E-Flora for Southern Africa: past experiences and future prospects for meeting target 1 of the 2020 global strategy for plant conservation. *Annals of the Missouri Botanical Garden*, **99**, 153–160.

D'Alessandro, L & Fattorini, L. (2002) Resampling estimators of species richness from presence absence data: why they don't work. *Metron*, **61**, 5–19.

Dubois, G., Schulz, M., Klooster, J., Verbeeck, B., Mayaux, P., Skøien, J., Cottam, A., Temperley, W., Clerici, M. & Drakou, E. (2015) *The digital observatory for protected areas (DOPA) Explorer 1.0*. EUR 27162 EN, EC. Publications Office of the European Union, Luxembourg.

Engemann, K., Enquist, B.J., Sandel, B., Boyle, B., Jørgensen, P.M., Morueta-Holme, N., Peet, R.K., Violle, C. & Svenning, J.-C. (2015) Limited sampling hampers 'big data' estimation of species richness in a tropical biodiversity hotspot. *Ecology and Evolution*, **5**, 807–820.

Gaiji, S., Chavan, V., Ariño, A.H., Otegui, J., Hobern, D., Sood, R. & Robles, R. (2013) Content assessment of the primary biodiversity data published through GBIF network: status, challenges and potentials. *Biodiversity Informatics*, **8**, 94–172.

Germishuizen, G. & Meyer, N.L. (eds.) (2003) Plants of southern Africa: an annotated checklist. *Strelitzia*, **14**. National Botanical Institute, Pretoria.

Goldblati, P. (1978) An analysis of the flora of Southern Africa: its characteristics, relationships, and origins. *Annals of the Missouri Botanical Garden*, **65**, 369–436.

Huntley, B.J. (1987) Ten years of cooperative ecological research in South Africa. *South African Journal of Science*, **83**, 72–79.

Hopkins, M.J.G. (2007) Modelling the known and unknown plant biodiversity of the Amazon Basin. *Journal of Biogeography*, **34**, 1400–1411.

Hortal, J., Garcia-Pereira, P. & García-Barros, E. (2004) Butterfly species richness in mainland Portugal: Predictive models of geographic distribution patterns. *Ecography*, **27**, 68–82.

Hortal, J., Lobo, J.M. & Jimenez-Valverde, A. (2007) Limitations of biodiversity databases: case study on seed-plant diversity in Tenerife, Canary Islands. *Conservation Biology*, **21**, 853–863.

Hortal, J., Jiménez-Valverde, A., Gómez, J.F., Lobo, J.M., & Baselga, A. (2008) Historical bias in biodiversity inventories affects the observed environmental niche of the species. *Oikos*, **117**, 847–858.

Hortal, J., de Bello, F., Diniz-Filho, J.A.F., Lewinsohn, T.M L., J.M. & Ladle, R.J. (2015) Seven shortfalls that beset large-scale knowledge on biodiversity. *Annual Review of Ecology, Evolution, and Systematics*, **46**, 523–549.

Jetz, W., McPherson, J.M. & Guralnick, R.P. (2012) Integrating biodiversity distribution knowledge: toward a global map of life. *Trends in Ecology and Evolution*, **27**, 151–159.

Küper, W., Sommer, J., Lovett, J. & Barthlott, W. (2006) Deficiency in African plant distribution data – missing pieces of the puzzle. *Botanical Journal of the Linnean Society*, **150**, 355–368.

Ladle, R & Hortal, J. (2013) Mapping species distributions: living with uncertainty. *Frontiers of Biogeography*, **5**, 8–9.

Leistner, O.A. (1983) Progress report on the Flora of southern Africa (F.S.A.). Proceedings of the Xth AETFAT Congress, Pretoria, South Africa. *Bothalia*, **99**, 153–160. Cited in Crouch, N.R., Smith, G.F., Figueiredo, E. (2013) From checklists to an E-Flora for Southern Africa: past experiences and future prospects for meeting target 1 of the 2020 global strategy for plant conservation. *Annals of the Missouri Botanical Garden*.

Léonard, J. (1968) Statistiques des progrès accomplis en 13 ans dans la connaissance de la flore phanérogamique Africaine et Malgache (1953–65). Cited in Hedberg, O., Hedberg, I. (eds) Conservation of vegetation in Africa south of the Sahara. Proceedings of the 6th AETFAT Congress at Uppsala, Sweden. *Acta Phytogeographica Suecica*, **54**, 297–299.

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**, 973–984.

Missouri Botanical Garden (2014) *Catalogue of the vascular plants of Madagascar*. Available at: http://www.tropicos.org/project/mada (accessed October 2015).

MOBOT (2013) *W3 Tropicos: vascular tropicos nomenclatural database*. Available at: http://mobot.org/W3T/Search/vast.html (accessed February 2013).

Nekola, J.C. & White, P.S. (1999) The distance decay of similarity in biogeography and ecology. *Journal of Biogeography*, **26**, 867–878.

Nelson, B.W., Ferreira, C.A.C., da Silva, M.F. & Kawasaki, M.L. (1990) Endemism centres, refugia and botanical collection density in Brazilian Amazonia. *Nature*, **345**, 714–716.

New, M., Lister, D., Hulme, M & Makin, M. (2002) A high resolution data set of surface climate over global land areas. *Climate Research*, **21**, 1–25.

Oksanen, J., Blanchet, G.F., Kindt, R., Legendre, P., Minchin, P.R., O'hara, R.B., Simpson, G.L., Solymos, P., Henry, M., Stevens, H. & Wagner, H. (2013) vegan: community ecology package. R package version 2.0-10. Available at: http://CRAN.R-project.org/package=vegan.

Osorio, F., Vallejos, R. & Cuevas, F. (2012). SpatialPack: package for analysis of spatial data. R package version 0.2. Available at: http://CRAN.R-project.org/package=SpatialPack.

Paradis, E., Claude, J. & Strimmer, K. (2004) APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, **20**, 289–290.

QGIS Development Team (2009) QGIS geographic information 624 system. Open Source Geospatial Foundation. Available at: http://qgis.osgeo.org.

R Core Team (2015) *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Available at: https://www.R-project.org.

Reddy, S. & Dávalos, L.M. (2003) Geographical sampling bias and its implications for conservation priorities in Africa. *Journal of Biogeography*, **30**, 1719–1727.

Rocchini, D., Hortal, J., Lengyel, S., Lobo, J.M., Jimenez-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*, **35**, 211–226.

Saarenmaa, H. & Nielsen, E.S. (Eds.) (2002) Towards a global biological information infrastructure. Challenges, Opportunities, Synergies, and the Role of Entomology. European Environment Agency, Copenhagen..

Ruete, A. (2015) Displaying bias in sampling effort of data accessed from biodiversity databases using ignorance maps. *Biodiversity Data Journal*, **3**, 1–15.

Schulman, L., Toivonen, T. & Ruokolainen, K. (2007) Analysing botanical collecting effort in Amazonia and correcting for it in species range estimation. *Journal of Biogeography*, **34**, 1388–1399.

Soberón, J. & Peterson, T. (2004) Biodiversity informatics: managing and applying primary biodiversity data. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **359**, 689–698.

Sousa-Baena, M.S., Garcia, L.C. & Peterson, A.T. (2014) Completeness of digital accessible knowledge of the plants of Brazil and priorities for survey and inventory. *Diversity and Distributions*, **20**, 369–381.

ter Steege, H., Haripersaud, P.P., Bánki, O.S. & Schieving, F. (2011) A model of botanical collectors' behavior in the field: never the same species twice. *American Journal of Botany*, **98**, 31–37.

Thiers, B. (continuously updated) *Index Herbariorum: a global directory of public herbaria and associated staff*. New York Botanical Garden's Virtual Herbarium. Available at: http://sweetgum.nybg.org/ih/ (accessed December 2014).

Yang, W., Ma, K. & Kreft, H. (2013) Geographical sampling bias in a large distributional database and its effects on species richness-environment models. *Journal of Biogeography*, **40**, 1415–1426.

Yang, W., Ma, K. & Kreft, H. (2014) Environmental and socioeconomic factors shaping the geography of floristic collections in China. *Global Ecology and Biogeography*, **23**, 1284–1292.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article at the publisher's web-site:

Appendix S1 Number of records by data provider.
Appendix S2 List of countries and their respective initial and final number of specimens.
Appendix S3 Number of specimens and species for individual countries.
Appendix S4 Relationship between estimate of inventory completeness obtained according to Sousa-Baena *et al.* (2014) and number of unique records (i.e. unique combination of date and location of collection and (a) species name, (b) number of specimens.
Appendix S5 Correlation between the outputs of the three methods applied in this study to estimate inventory completeness.
Appendix S6 Maps of estimates of inventory completeness according to Chao & Jost (2012), Sousa-Baena *et al.* (2014) and curvilinearity of species accumulation curves.
Appendix S7 Descriptive statistics of the estimates of inventory completeness according to the three methods applied in this study and correlation between inventory completeness and year in which specimens were collected for each African country; SUs $\geq$ 200 unique records.
Appendix S8 Descriptive statistics of the estimates of inventory completeness according to the three methods applied in this study and correlation between inventory completeness and year in which specimens were collected for each African country; SUs $\geq$ 50 specimens.
Appendix S9 Relative abundance distribution of the botanical collections for a few selected individual sampling units.
Appendix S10 R-script.

BIOSKETCH

Juliana Stropp is a post-doctoral researcher at the Federal University of Alagoas. Her major research interests are conservation biogeography and tropical forest ecology.

Editor: Maria Dornelas