Big data-big problems? How to circumvent problems in biodiversity mapping and ensure meaningful results

Alice Hughes[1], James Dorey[2], Silas Bossert[3], Huijie Qiao[4], and Michael Orr[4]

[1]Xishuangbanna Tropical Botanical Garden
[2]Flinders University of South Australia
[3]Washington State University
[4]Institute of Zoology Chinese Academy of Sciences

September 19, 2023

Abstract

Our knowledge of biodiversity hinges on sufficient data, reliable methods, and realistic models. Without an accurate assessment of species distributions, we cannot effectively target and stem biodiversity loss. Species range maps are the foundation of such efforts, but countless studies have failed to account for the most basic assumptions of reliable species mapping practices, undermining the credibility of their results and potentially misleading and hindering conservation and management efforts. Here, we use examples from the recent literature and broader conservation community to highlight the substantial shortfalls in current practices and their consequences for both analyses and conservation management. We detail how different decisions on data filtering impact the outcomes of analysis and provide practical recommendations and steps for more reliable analysis, whilst understanding the limits of what available data will reliably allow and what methods are most appropriate. Whilst "perfect" analyses are not possible for many taxa given limited data, and biases, ensuring we use data within reasonable limits and understanding inherent assumptions is crucial to ensure appropriate use. By embracing and enacting such best practices, we can ensure both the accuracy and improved comparability of biodiversity analyses going forward, ultimately enhancing our ability to use data to facilitate our protection of the natural world.

Big data-big problems? How to circumvent problems in biodiversity mapping and ensure meaningful results

Abstract: Our knowledge of biodiversity hinges on sufficient data, reliable methods, and realistic models. Without an accurate assessment of species distributions, we cannot effectively target and stem biodiversity loss. Species range maps are the foundation of such efforts, but countless studies have failed to account for the most basic assumptions of reliable species mapping practices, undermining the credibility of their results and potentially misleading and hindering conservation and management efforts. Here, we use examples from the recent literature and broader conservation community to highlight the substantial shortfalls in current practices and their consequences for both analyses and conservation management. We detail how different decisions on data filtering impact the outcomes of analysis and provide practical recommendations and steps for more reliable analysis, whilst understanding the limits of what available data will reliably allow and what methods are most appropriate. Whilst "perfect" analyses are not possible for many taxa given limited data, and biases, ensuring we use data within reasonable limits and understanding inherent assumptions is crucial to ensure appropriate use. By embracing and enacting such best practices, we can ensure both the accuracy and improved comparability of biodiversity analyses going forward, ultimately enhancing our ability to use data to facilitate our protection of the natural world.

Introduction: Changing data challenges and accelerating needs

Understanding species ranges and developing approaches to reliably monitor distributions are essential needs of any conservation target (Hughes et al., 2021a). This is especially important in the context of the Post-2020 Global Biodiversity Framework (GBF) and the accompanying Monitoring Framework, which aims to provide the metrics to measure progress towards new targets. In recent years, biological data availability has transformed, and we have moved from being data-poor to data-rich. There are now 2.3 billion records on GBIF, over 1.3 billion in eBird, and over 130 million on iNaturalist, with citizen science data increasingly outweighing prior museum specimen data for many taxa. Point-based data enable sophisticated modelling, and analysis of traits and changes over time in ways unimaginable previously (GBIF Secretariat 2021). Yet greater access to data can mean that basic ecological principles are forgotten, and analysis becomes merely a statistical exercise. Combined with the incentivisation to publish high impact (Eyre-Walker 2013) and global papers (Wyborn & Evans 2022), there is a temptation to focus on headline titles, regardless of data extent, pervasive biases, or the specific methods required to account for these issues (Hughes et al., 2021b). Thereafter, once such resources are published, the path of least resistance is to reuse them rather than reinventing the wheel (unfortunately, this often happens even if the wheel is lopsided). More problematic, the existence of studies claiming to do something reduces the novelty of any subsequent studies improving prospects, greatly reducing the potential for high-impact publication and thereby also reducing incentives to make such improvements in the first place. Furthermore, big "headlines" will be requoted and taken as accurate, and commentaries and responses are hard to publish and are unlikely to receive such attention; meaning that like anything on the internet today once an inaccurate analysis is published, the headline may be taken as dogma, and if they are inaccurate the consequences of those inaccuracies are inevitably propagated.

Complicating matters, the discrepancy between data-rich and data-poor regions generally mirrors that of GDP, and consequently the areas with the richest biodiversity may also have the poorest coverage in terms of biodiversity data (Giam et al., 2012; Stork 2018; Hughes et al., 2021b). This means that "global studies" disproportionately represent patterns in higher income economies. Reconciling these biases requires both understanding and then working to overcome sampling related issues. Such analyses are crucial across taxa, regions, and scales, forming the foundation of effective National Strategic Biodiversity Action Plans (NBSAPs) (Whitehorn et al., 2019, Schmidt-Traub, 2021).

Understanding data challenges and limits, and how and which methods can be applied, is critical. This is because every form of analysis makes its own assumptions, and every dataset has its own biases and inconsistencies. Depending on what data are available, modelling species-level data may only be locally possible for many taxa, and using data beyond sensible bounds can misdirect priorities and misinform management plans. Without an understanding of how data were generated or their biases and shortcomings, the likelihood of misuse increases. Thus, with this ever-growing wealth of data, it is critical to understand how to use it effectively, and what processes to follow to ensure that outcomes can meaningfully guide future conservation targets.

Here, we explore different methods commonly applied to biodiversity data, the assumptions they make, and the impact of applying them to data that do not meet those assumptions. We then discuss approaches and frameworks representing best practices in biodiversity data analysis, and provide a stepwise framework which can be used to ensure biodiversity data are used within their limits and that the assumptions of each step are clearly understood.

Context: The current data landscape

Biodiversity mapping is a central tenet of conservation, and it is achieved either through the use of point data or expert range maps. First came basic repositories of museum point data such as Arctos in 1996 (Jarrell et al., 2010), followed by GBIF in 2001 and OBIS for Ocean systems in 2002. More recently, a wealth of citizen science platforms have grown to provide a greater spatial, temporal, and taxonomic volume of data than has ever previously been available. Such data promised to revolutionise our understanding of global biodiversity patterns, enabling us to finally move away from the hand-drawn range maps in Birdlife, IUCN, GARD and Fishbase (all of which aggregate data and map species distributions), which themselves represented a major

step forwards in providing any spatial data to map species ranges for thousands of species (Hughes et al., 2021c). Most global papers continue to use these often hand-drawn polygon maps to visualise priorities or assess gaps because they are easy to download and simple to use, despite studies highlighting substantial scale-dependence and errors of both omission and commission (Herkt et al. 2017; Li et al., 2020; Hughes et al 2021c). Growing volumes of point data provide an option of better, and more accurate species ranges analysis than polygon approaches, yet these empirical point-based data present a different challenge; how can we use these data without over-extending their limits?

As point data continue to increase in volume, so too will the number and type of biases within those data, making their reconciliation more complex. At the same time, the drive to perform large-scale mapping is increasing (Wyborn & Evans 2021). There has never before been such a dire need for these data, with the failure to complete the Aichi Biodiversity Targets, the delays in most countries completing their National Strategic Biodiversity and Action Plans (and the need for revisiting these plans), and the Monitoring Framework. Understanding where we have data, what biases exist in those data, and how they can be overcome is critical to even begin moving towards the goals of the GBF (Mace et al. 2018). Modelling and mapping has previously even been the subject of major assessments (i.e. IPBES 2018), and thus whilst there is guidance on model approaches, understanding how different forms of data, such as species distributions, can be effectively used is essential.

Understanding the dimensions of biases and their consequences

Accurate mapping of species ranges has some basic tenets that must be followed to produce reliable results (Malavasi 2020). Depending on the amount and quality of available data, increasingly sophisticated approaches can be used, but even basic methods have great potential when used carefully and appropriately. For the most basic analyses, data must still be representative, with minimal biases, but data in popular databases were largely collected from areas within 2km of roads (Hughes et al., 2021b), meaning generalist and disturbance-tolerant species will be over-represented and relatively few areas will be protected in most cases (given the higher level of development near mapped roads; Hughes 2018). These biases can be even stronger in citizen science data or social media data, and must be carefully accounted for (Bird et al., 2014; Johnston et al. 2020; Hughes et al. 2021b; Barber et al., 2022; Chowdhury et al., 2023a).

Common uses of biodiversity data include species-level (i.e. IUCN Red List) to regional and global mapping. Thus, spatial and taxonomic biases will be a particular problem with biodiversity mapping approaches which do not include point projection or extrapolation to circumvent data gaps and biases. On a species-level, approaches such as extent of occurrence and area of occurrence (EOO and AOO; which bases ranges on minimum spanning polygons and occupied cells, respectively) will disproportionately represent the most disturbed areas where most data exist, and whilst they are only sometimes calculated with available point data (unless further additional data can be collated for verification), they are a core component of IUCN range assessments. Further, without careful data cleaning, errors can obscure any patterns or trends, as common errors (switched coordinates, incorrect georeferences, misidentifications, etc.) may mislay species distribution points across different continents, even different hemispheres, and a projective approach consequently risks overestimating distributions across much of the global land-surface. For all these reasons, to assess if a dataset can be used for meaningful analysis, we first need to determine if the data are taxonomically or spatially representative, if it represents the current ranges of species, and how most data were generated, before assessing if methods used for assessing ranges can give a useful understanding of species ranges.

Globally, biases are pervasive, and representative and accessible data simply do not exist for many taxa, especially hyper-diverse and challenging taxa such as insects (Garcia-Rosello et al., 2023). Most data come from high-income countries, despite only a minority of species coming from these regions (Hughes et al., 2021; Orr et al., 2020). For example, Asian bees represent 15% of globally described species, but only 1% of data (Orr et al., 2020). Notably, problematic knowledge gaps still remain even in some of the best-known areas like Europe (Leandro et al. 2017). Ultimately, any research attempting global-scale analysis based on point data would strongly over-represent patterns in high-income economies and better-known regions within them (Orr et al., 2021), though different data-sharing policies from different countries or institutes may alter

completeness of available data. Countries and regions may limit both data availability, and permitting for research, both of which can alter the availability and therefore representativeness of data, particularly in developing economies. These policies may also limit collaborative efforts, particularly between neighbouring countries, and make transnational analysis particularly challenging due to a lack of standardisation between regional data-collection efforts. Thus, equitable international partnerships are a paramount consideration for research going forward, and understanding how to reconcile differences is crucial to best use data that has already been collected.

Taking a recent example, Chowdhury et al., 2023b) aimed to assess the adequacy protection coverage of protected areas for all insects. Using their occurrence data DOI, we can see that 69% of their data came from Europe, and 21% from North America, whereas conversely each of Africa, Asia, Oceania, and South America only contributed about 2% of records each, despite having much greater diversity (Orr et al., 2020). This is consistent with other analyses across taxa, highlighting why interpolation requires extreme caution when interpreting global trends (i.e. Orr et al., 2021; Hughes et al., 2021b); a recent attempt at global bee decline analysis faced similar geographic biases and did not account for them (Zattara & Aizen 2021), and even many regional analyses (e.g Kerr et al., 2015) do not adequately account for potential changes in collector aims and behaviours over time. There are similar biases in taxonomic representation (Chowdhury et al., 2023b); of the eight genera with over one million points in GBIF, seven are butterflies, and in fact 51% of all GBIF invertebrate data are for Lepidoptera. In another example, Bolam et al. (2023) analyzed most IUCN-assessed threatened species and claimed that over half of threatened species require recovery actions, but which species have been assessed is itself biased by spatial effort for most groups, as well as only representing a subset of species (Hughes et al., 2021a), so these results may not be generalizable. This inability to generalise given the lack of spatial or taxonomic representation may be the case for many studies using subsets of total diversity for which there are sufficient data (Visconti et al. 2016; Pacifici et al. 2020). Furthermore, taxonomic expertise may be limited, and in certain taxa, reliance even on single experts may impact the reliability of results across space and time. Given acknowledged spatial biases (Hughes et al. 2021b; Rocha-Ortega et al. 2021), many of these species likely already occupy human-modified areas and would benefit little from protected areas, or might even do worse if they show anthropophilic tendencies. Furthermore, overall 80% of insect families have under 10% of species covered, and the remaining 20% have 11-13%, meaning that no generalizable conclusions can be drawn when the majority of species are not covered (because those with sufficient data do not largely represent rarer species). In terms of how records were generated, 62% were human observation (majority citizen science data) and 31% were specimens (from possibly centuries ago, as a time filter was not applied). Some specimens might even be fossils of species which are still extant (as fossil records were downloaded and only extinct fossil species were removed).

Citizen science data can also complicate matters when used uncritically, yet such data makes up most contemporary data for many groups. Taxonomic biases are amplified in these data, with birds clearly dominating (Dobson et al., 2020; Di Cecco et al., 2021), but other groups may also exhibit biases. For example, 11% of all insect data from Chowdhury et al. (2023b) was just from UK butterfly and moth citizen science monitoring programs, and a further 10% were from global citizen science programs. This would exacerbate the aforementioned biases of collections in developed areas (Hughes et al. 2021b), as seen for bees but was unaccounted for in Zattara & Aizen (2021). Citizen science data are useful, especially for phenological monitoring, but occur disproportionately in developed areas and for common and easy to recognise species, especially of "charismatic" or "beautiful" species, meaning that, without careful steps to correct for or counteract biases, these data can compound biases and not necessarily improve our knowledge (Dickinson et al. 2012; Bird et al. 2014; Ward 2014). This can also impact the ability of such data to detect trends (Kamp et al., 2016). Generally, regional differences in organisations and their taxonomic foci means that these biases may vary by region; the UK in particular has a huge aggregation of Lepidoptera data, despite not having particularly high diversity relative to other regions. This is further complicated by differences in data sharing policies. For example, iNaturalist is free to use, but the Bees, Wasps & Ants Recording Society of the UK shares data publicly only on 10 km grids and, as such, cannot be included in GBIF downloads (https://www.bwars.com/content/bwars-data-download).

These issues mean that the public data for insects are not spatially or taxonomically representative, and cannot be regarded as such, so approaches to circumvent these issues require either resampling or reprojection methods if we wish to reconstruct meaningful global, or even regional patterns (i.e. Orr et al., 2021). Given the spatial and taxonomic gaps, clear assessments of representativeness of data are needed, especially in tropical regions (Giam et al., 2012; de Araujo et al., 2022), and supplemental data may be needed for the most basic view of many taxa. Many regions have disproportionate volumes of data with different regional biases, for example Chowdhury et al., (2023b) have largely provided a metric with a disproportionate emphasis on European Lepidoptera. However, even within Europe, further assessment of the percentage of the area would be needed to judge the true degree of protection, unless additional modelling and calibration were used to reconcile the biases within these regions. Stating any result beyond these regions and taxa risks being misleading, while giving the impression that we already have sufficient data for such types of global analyses, which we simply do not for many taxa (Wyborn & Evans, 2021).

Measuring diversity in the face of biases

Following the case study outlined above, it is clear that there are right and wrong ways to carry out and interpret analyses. Whilst some shortcuts may have relatively little impact, others may invalidate interpretations when assumptions are not met. It is also worth considering the cumulative contribution of smaller biases in undermining accurate interpretation. Now we come to the methods: how can we use spatially and taxonomically biased data to recover biodiversity patterns? Avoiding biases entirely is virtually impossible in large datasets, so finding methods to minimise their impacts is critical.

Different researchers and organizations employ different methods, but consistency is important to enable comparison. One of the major methods used by the IUCN in their generation of species ranges (as part of IUCN assessments) is that of the extent of occurrence and area of occupancy (AOO, which is the occupied subset of the EOO-which is an MCP minimum convex polygon) to quantify species distributions. Such methods have also been employed in some research articles (Bradshaw et al., 2014; Chowdhury et al., 2023), sometimes with useful improvements that help alleviate biases (Kass et al., 2021). However, mapping these ranges requires a certain level of data confidence and completeness in existing distributional data, and the uses of either technique are limited if systematically collated data were not used for mapping ranges. In such cases, the AOO approach cannot be usefully applied alone because most points are from developed and disturbed areas (for many species, we would not expect the areas inhabited to be protected as they are less likely to represent high-quality habitat targeted for such protections). Thus, if an AOO based on point data is used, a null-dataset (bias weighting and assessment of representativeness) may be needed to ensure that surveyed cells are adequate, as biased sampling again means that more developed and less protected areas will have more species data. Areas of occurrence would therefore require further stratified sampling, such as grid- or transect-based approaches, or would need to rely on percentage presence approaches (the percentage of surveyed cells that a species were present in) to try to assess patterns; and may be entirely unsuitable if the degree of data coverage is too low. Furthermore, the AOO system of analysis has basic requirements to be performed well; based on inventory of species presence within the estimated range; and such a requirement is unlikely to be met in global analysis where degree of sample coverage will underestimate occurrence wherever data gaps occur.

Methods aiming to understand entire species ranges (and with inadequate data for occupancy-like approaches) may attempt to use point-based data to assess entire species ranges, but this requires extreme caution. For interpolation-based approaches, points must be an accurate reflection of species ranges, and multiple measures are needed to curate the data and ensure that they are accurate. Because of the source of data and possible encoding errors, spatial filters are needed to ensure that data are accurate. The development and cleaning of databases is an important, but challenging endeavour, but is needed to enable higher-resolution analysis. Filters could include admin-area checklists (following correction of synonyms using a curated list) or hemisphere filters (Orr et al., 2021) to remove points where coordinate errors may exist or points may be in private collections or even zoos. Chowdhury et al. (2023) used CoordinateCleaner (Zizka et al. 2019) to filter for occurrence quality (Table S1). Without clearer filters for realms or continents

this still leaves considerable room for errors. Recently, two more complimentary packages, bdc (Ribeiro et al. 2022) and BeeBDC (Dorey et al. *In review* ; Table S1), have been released that address additional and complementary quality checks. A failure to adequately clean and filter data can substantially inflate species ranges. Additionally, even after filtering, data must be examined and analysed with a critical eye and with hypotheses kept in mind to ensure that results fall within the limits of what the data permit.

Conversely, there can be issues if a range is calculated based on a minimum number of points without spatial thinning or removal of duplicates when calculating the number. For example, Chowdhury et al. (2023) aimed to assess the percentage of ranges protected for all insects but analysed 217 species with a polygon area of "3" and 1105 with under "10" (presumedly kilometres); these low values suggest that many species ranges are underestimated, possibly reported only from small-scale inventories. If they are not from a protected area, such species would automatically be 100% unprotected. Conversely, a species with three geographically disparate records might have its range greatly overestimated, potentially also exaggerating the percentage of its range that is unprotected. Points included in studies such as Chowdhury et al. (2023) lack such filters, and whilst attempts were made (using the rangeBuilder package) such errors will prevail when data quantity is low and filters have not been implemented to clean distributional errors. However in a better scenario, where filters had been applied would a convex hull be appropriate to represent current species ranges?

Given the paucity of data for most insects (and many other taxa), different approaches have been applied to circumvent issues in existing studies aiming to map richness across continents. These include interpolating richness by modelling it directly as a function of ecological drivers (termites, bees, Collembola: Liu et al., 2022; Orr et al., 2021; Potapov et al., 2023, respectively), species-level modelling within MCPs to delimit ranges, or modelling following the building of a comprehensive point based database (ants: Kass et al., 2022). Even for smaller regions, where there is high diversity, filtering suitable habitat within a polygon provides a much more targeted and realistic understanding of species ranges (Chesshire et al., 2023), and in all of these studies the impact of data biases were limited either through interpolation of richness for data poor regions, or via aggregation of further data followed by modelling, and in all of them habitat filters were applied to ensure that analyses were accurate. If these steps are not taken, results could be misleading and, at times, counterproductive, given the limited funds and support available for conservation and area protection.

Using data to enable conservation planning

Most forms of biodiversity data can be used for environmental management and conservation planning, and tailoring analysis to the available data is critical. Target development can use two major approaches, either focusing on species distribution mapping, or, if data of sufficient resolution, quality and quantity are not available, attempting to map diversity patterns and reconcile richness patterns in the face of bias. If sufficient data are available either models or filtered convex polygons can be used to map species ranges. Whereas when insufficient data are available and richness is mapped (Liu et al., 2022; Orr et al., 2021; Potapov et al., 2023), inventories of species richness are used and richness itself is reprojected using models. These approaches can reproject richness patterns to a reasonable degree, if sufficient inventories have been carried out across all major environmental conditions, and assuming that biogeographic differences will not influence overall richness patterns. These approaches are useful in groups where insufficient data are available for higher resolution analysis, and can also be used to identify areas for further research if there is a potential for high hidden diversity (Orr et al. 2021; Kass et al. 2022).

For large spatial or taxonomic scopes, if species-level analysis is impossible (the majority of global analyses) then interpolation-based methods are likely to be the most appropriate. In these types of studies, one should employ either a subsampling approach (Qiao et al., 2023) or interpolation based on community level inventories (to model richness overall rather than individual species ranges: Orr et al., 2021 Liu et al., 2022; Potapov et al., 2023). For subsampling, there is still a minimum data requirement, as most areas lack data. Thus, for well sampled taxa such as birds, it can be fairly widely applied (e.g. almost all urban areas), but there is so little sampling for most taxa that an index-based approach may not be possible without then interpolating. Approaches based on biodiversity indices (Hill numbers, Shannon, Simpson, etc.) all require both a minimum number of samples and a minimum coverage (Qiao et al., 2023). Species area curves

are a common way to estimate completeness for any given region, yet these assume representative coverage throughout that region and a localised inventory of a small proportion may asymptote even when it is not representative of the whole area to which the assessment is applied. Thus, for such curves to be useful first assessments of the percentage of the area with data is needed.

For many taxa, including most invertebrates, interpolation based on modelling is needed. Such methods rely on interpolating richness based on community-level samples and using species modelling techniques to relate richness data to conditions present. Inevitably, this method also involves assumptions about the representativeness of the data. For a community projection approach, a minimum sample-size and species number should be used (to remove the possibility of selective sampling or overrepresentation of generalists to the neglect of specialists), and all biome types should be represented so that the richness (or richness index) of these varying biomes and conditions can be assayed. However, it should be noted that such an approach will assume that there are no biogeographic variations in drivers between regions, and consequently cannot be applied to oceanic islands, as such models cannot inherently incorporate biogeographic processes or dispersal. Thus, for interpolation approaches to be applied, the number of records per species, and even to a degree the accuracy of identification within sites is less important (provided it is consistent within a site), and provided there is coverage across environmental conditions these approaches provide a powerful mechanism for global analysis, enabling analysis even in poorly-known regions.

For species-specific approaches, both the volume and accuracy of the data must be substantially higher, as they are much more vulnerable to spatial bias and sensitive to data errors, with even greater consequences for poorly known species. Firstly, data must be clean and accurate for any species-level assessment, so cleaning checks and filtering of bad records is a critical first step (see Box 1). The first question is whether the data are sufficiently representative for species level analysis both in terms of taxa, and the region under analysis, furthermore any form of species level assessment requires sufficient data to assess range. When examined critically, public data sources alone are insufficient for modelling most species (Garcia-Rosello et al., 2023), even across vertebrates, so some of the most diverse regions might be underestimated.

Sophisticated models can be developed for well-sampled individual species using approaches such as Maxent, or other species niche modelling methods can be applied (though many of these will map all relevant habitats, lacking any geospatial reference point to differentiate functional and realised niches). However, such models have very high data requirements, as sufficient and even data must exist from across a species' range to effectively model its distribution and pair it with environmental characteristics. This means that, unless considerable effort is devoted to collating representative global data with many partners, or taxa are already well sampled, sophisticated models may not be representative or appropriate. Assessing these models, not only using statistical approaches (AUC, Boyce index, AIC) alone is also not sufficient, and work with experts to assess if ranges capture species ranges is also likely needed to assess whether they are reliable, and also recognise biogeographic boundaries (which may be missed in models, especially in complex areas or where there are major differences between fundamental and realised niche). MCPs may also be used when data are scarce or analysis is regional, but understanding how to curate data is a first essential step before mapping species ranges (Zizka et al. 2019; Ribeiro et al. 2022; Dorey et al. *In review*).

Filtering for success

For basic analysis of large numbers of species, automated and repeatable pipelines are critical. Creating an MCP is one method to delimit the majority of a species' known range. However, for vertebrates it has been known for centuries that species have finer-grain habitat requirements, and even in IUCN maps the need to refine habitat within the range polygon is becoming a basic standard (Lack, 1953; Brooks et al., 2019); points may completely surround cities or other unsuitable regions, yet the species may no longer be present there. Failure to remove clearly unsuitable habitat would both dramatically increase range size and could reduce the proportion of range protected (as much of a city is developed). Coastal filters are also needed, as a failure to realistically trim MCPs may render oceans suitable for land animals.

Sensible filters can transform species ranges and entirely rearrange diversity patterns. To demonstrate how

decisions on data-refinement and cleaning impact on range sizes and degree of protection, we selected a range of species and imposed different levels of filtering on the data, all of which can be conducted with small datasets, or when some species may have small volumes of data available. This includes adding spatial filters, adding a habitat filter, trimming by coastline, and comparing it to known IUCN ranges for species. It should be noted that most IUCN ranges are also inaccurate and overinflate species ranges (Li et al., 2019; Hughes et al, 2021c), yet uncritical MCPs are exponentially larger (whilst still missing parts of the range as they will not capture species range limits, where abundance is typically lower). For example, an IUCN range is only 7-8% the size of those recovered using basic MCPs for the species shown here (as in Chowdhary et al., 2023a). If these ranges are being mapped to assess hotspots for protection, or the degree of protection, then the area covered and the location will entirely determine the outcomes of assessment, and if care to filter data appropriately is not applied, then analysis on such data may have little relationship with the real patterns of distribution or degree of protection of species.

Even when more carefully delineated ranges (IUCN, birdlife, GARD: http://www.gardinitiative.org/) are likely to overestimate the degree of protection, their area is still smaller than an MCP, especially if a habitat filter is not applied (Table 1). We used a general habitat filter, so more specialist filters and other steps outlined throughout could greatly improve range estimates and make them more similar to those in expert range maps (de Barros et al., 2021; Huang et al., 2020; Xu et al., 2022). In all cases, the lack of filtering means ranges are projected as many times larger than they are likely to be. Thus, as we show here, the cleaning of data can transform where species are mapped, richness patterns, and the efficacy of protection. We selected a range of species for which sufficient data exist to map ranges, and where the IUCN has mapped ranges for comparison (thus most of our examples are mammals, though one bee, *B. dahlbomii* , is also present), our previous work has also examined the prevalence of biases in these types of data, and how they persist across taxa (Hughes et al2021b, 2021c; Li et al., 2019).

Table 1. Percentage of species range protected with different filters applied for species minimum convex polygons (MCP), as well as for International Union for the Conservation of Nature (IUCN) ranges. The filters that were applied are noted in column headers: Hem-hemisphere filter, Coast-removal of ocean areas within the polygon, habitat-a simple habitat filter based on basic classifications of land-use types. We used the species *Ailuropoda melanoleuca* (Carnivora: Ursidae),*Bombus dahlbomii* (Hymenoptera: Apidae), *Panthera onca* and*Panthera tigris* (Carnivora: Felidae), *Priodontes maximus*(Cingulata: Chlamyphoridae), *Tapirus pinchaque* and *Tapirus terrestris* (Perissodactyla: Tapiridae).

	MCP Area	MCP Area_basic	MCP Area_hem	MCP Area_hem_coast	MCP Habitat	IUCN IUCN
Ailuropoda melanoleuca	2.06	7.49	7.49	7.49	7.82	23.05
Bombus dahlbomii	1.91	1.91	9.37	11.92	7.34	25.03
Panthera onca	12.21	12.21	16.35	24.54	27.21	38.87
Panthera tigris	1.08	1.18	4.92	6.84	7.96	10.62
Priodontes maximus	26.01	26.01	26.01	33.44	35.53	33.92
Tapirus pinchaque	23.65	23.65	23.65	23.65	23.64	47.08
Tapirus terrestris	29.38	29.38	29.38	29.60	32.28	30.61

If we then compare the IUCN range maps to the MCPs (using the habitat-filtered maps, as these can be as small as 8% (*B. dahlbomii, P. onca)* of the largest MCPs), we find large areas of commission error (mapped areas that fall outside species actual ranges due to either wrong points, or unsuitable habitat within the MCP), highlighting that much of these areas are still unsuitable (Table 2). Whilst modelling could be used, and would further refine distributions (i.e. Hughes et al., 2021c), these steps highlight that even if data were adequately cleaned, the final outcomes would still not be indicative of species ranges; at a minimum, a habitat filter must be imposed, and coarser filters based on intactness can be easy to apply (e.g. Lu et al., 2021). Thus, whilst global prioritisation and ambitious analyses are needed, species-specific models or

maps for many groups, such as insects, are simply not possible globally without additional efforts to collate data (Garcia-Rosello et al., 2023); such attempts would simply not be representative of data-poor regions, or indeed, any biodiversity hotspot (especially if no habitat filter was implemented). Until data-gaps are filled (as in the case of concerted work for ants: Kass et al., 2022), meaningful analysis will remain challenging or even impossible at global scales, and focusing on taxa or regions where such data are available is necessary.

Table 2. Overlap between IUCN ranges and habitat-filtered MCPs.

	Ailuropoda melanoleuca	*Bombus dahlbomii*	*Panthera onca*	*Panthera tigris*	*Priodontes maximu*
MCP- Commission	80.68	72.05	43.80	91.52	17.96
IUCN-Omission	8.27	6.28	4.23	2.13	14.71
Overlap	11.05	21.66	51.97	6.35	67.33

Defining your research question, selecting and processing the data

Mapping species ranges is important, but any study seeking to understand species ranges must follow sensible steps to ensure that analyses are appropriate, and that the caveats and assumptions at each step are clearly understood. Any researcher should keep several questions explicitly in mind from the outset of their studies, and need to assess if the data exists, or can be collated to answer those questions (Figure 1).

1. Defining the questions: What is being attempted, what data are needed, and do those data exist?
2. Selecting appropriate data to use: What assumptions are you making, what limits exist in potential data use? For instance, can citizen science data be used for your specific aim, and if so how could such data be sufficiently cleaned to be reliable?
3. Understanding the scale and representativeness: Are the data representative of the species and region under study?

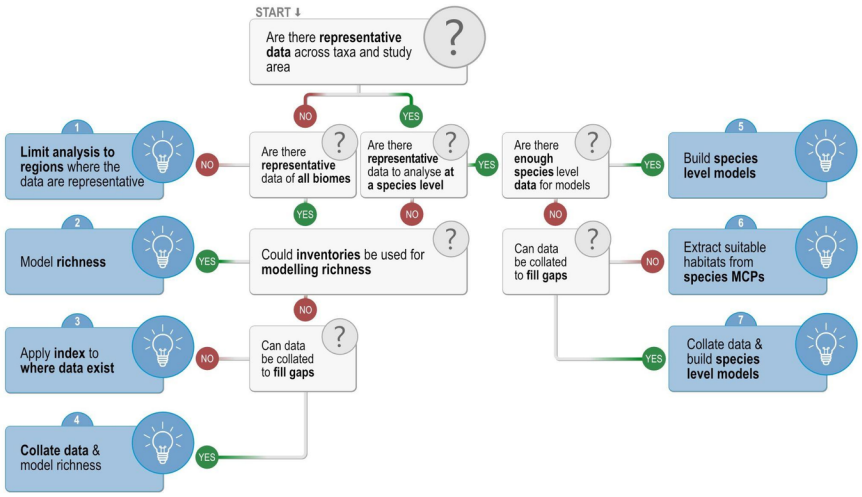

Figure 1. Decision tree of how to accurately use data to analyse patterns of richness. Questions are given in the light grey boxes and solutions are provided in blue

Understanding species ranges and how well-protected they are is crucial for filling gaps in protection coverage and effectively protecting species into the future, but poor analysis can mislead and misdirect attention and waste limited conservation resources. Here, we demonstrate the potential pitfalls of uncritical species mapping, highlighting the importance of thoroughly vetting data sources and engaging them with the same ecological principles that are expected for the best-known species. Ultimately, best-available-data arguments

seeking to overreach beyond what is possible (i.e. globally, Wyborn & Evans 2022) are bad-data arguments that can also have bad outcomes for conservation. Furthermore, "effective protection" needs may also differ across species (Butchart et al., 2015), such that generalists and specialists would respond differently, and this must also be considered, rather than naïve aggregations that treat all species as generalists or only include generalist species and then claim representation for entire groups.

Various analyses have been conducted which use these approaches (Figure 1) to sensibly and sensitively map species ranges and biodiversity patterns. For example, in terms of limiting analysis to areas where data are representative, there are many regional analyses which use such approaches (Figure 1; solution 1), as well as studies which increase the resolution of studies by incorporating other types of data not available at wider scales (Fukaya et a., 2020; Cosentino et al., 2023). For modelling richness patterns in the absence of sufficient data for species-level analysis, examples exist for data-poor invertebrates, ferns, and lycophytes (Orr et al., 2020; Weigand et al., 2020; Liu et al., 2022; Potapov et al., 2023) (Figure 1; solution option 2), and such work may also include collating more data to ensure that it is spatially and taxonomically representative (Figure 1; solution 4) (Qiao et al. 2023). A good recent example is for ants, where a large database was collated to ensure that it was representative before conducting individual species models to map richness (Figure 1; solutions 5 and 7) (Janicki et al., 2016; Guénard et al., 2017; AntWeb; https://antweb.org; Kass et al., 2022). Alternatively, indices can also be applied to fill gaps, either for limited regions such as urban areas (Hughes et al., 2022) (Figure 1; solution 3), or indices combined and interpolated more widely (Qiao et al., 2023). Only where representative data exist across taxa can species-specific analyses be conducted. If limited species-specific data are available, this may include simple approaches, such as selecting suitable habitat from within MCPs (Figure 1; solution 6) (Chesshire et al., 2022), or following approaches similar to those advocated for refining IUCN maps to better represent species true ranges (landcover and elevation filters within similar ecoregions or biomes, Brooks et al., 2019). In all cases, care and calibration at each step is needed to ensure that data are used within reasonable limits, to not over-extend their utility, or overreach on interpretation.

Figure 2. Projecting richness with different approaches. A. Maxent species models (Figure 1 solution5). B. Interpolating richness from inventories (solution 2). C. Filtering habitat within convex polygons (solution 6). D. IUCN richness. E. IUCN richness filtered by suitable habitat, F. MCP unfiltered richness.

Each approach has its own inherent assumptions and level of detail (see Figure 2), so some caveats are necessary. Patterns from each should be comparable (provided that there are sufficient data for analysis), but they will not be identical, and richness values will vary between approaches (thus, relative patterns rather than absolute values should be explored). Selection of input variables, particularly land-cover variables, can strongly influence the projected distribution of species, and if some regions are under-sampled or under-represented then they may not be accurately reflected in analysis; this problem increases with scale. Categorical variables can artificially delimit ranges, and thus whilst we used the IUCN ecosystem typology to increase the sensitivity of analysis (as more accurate National maps are not available for continental regions such as Africa), this may still cause more extreme transitions in apparent diversity (especially in small, inaccessible and under-sampled regions) but this is still necessary to refine overgeneralised maps (Figure 2E vs D for IUCN and F vs C for MCPs); yet more sensitive models using interpolation or modelling based on continuous variables (such as vegetation height and density) can be better where they can be applied (Figure 2 a, b). Furthermore, selection of models requires care, as joint-species distribution models are becoming increasingly popular (Zurell et al., 2018; 2019), however whilst popular these should be applied only where a functional relationship between species (competition, mutualism etc) exists, as correlations based on bias sampling could emerge from applying such models without verified ecological relationships (Poggiato et al., 2021). Similarly tied to scale, interpolations over very large and heterogeneous regions cannot incorporate biogeographic differences if relying solely on interpolation of richness rather than species-specific approaches, and may better represent drivers of patterns in better sampled regions. Furthermore, interpolations between mainland areas and islands cannot encompass island biogeography, as dispersal would be assumed equal through the study area. If using MCPs, the edges of species distributions will be excluded (as the approach can only assess range within the recorded maximum distribution of the species based on known locations;

see Figure 2C, F). Analyses using unfiltered MCPs (Figure 2F) are vulnerable to "mid-domain" type effects, where central areas will appear richer regardless of land-cover type due to the probability of overlap between MCPs in central areas of maps (Figure 2).

The importance of reflecting habitat in maps is clear by comparing patterns when habitat variables are incorporated versus richness maps based on simple clipped maps of maximum range extent (Figure 2C vs 2E). Inventory-based approaches also rely on data being representative throughout the study area. Assessing the performance of inventory-based approaches can be challenging without having complete inventory data for cross-referencing (Figure 2B). Hence, expert knowledge may be the only way to assess if patterns "make sense". Furthermore, as interpolation-based approaches may work better over large regions where the full environmental gradient is adequately sampled they might be more appropriate over very large scales (i.e. global) given that some regions (such as Europe and the US) have very good sampling. However, at a continental scale this can be challenging without additional fieldwork or data-collation to add additional inventory data (which was not possible in the case of this study, 2B). In addition, less species-rich areas may be more poorly sampled, so inventory-based interpolations need at least some site-based inventories in addition to calculations of richness by larger unit areas. Assessment and ground-truthing at each step is a fundamental necessity of analysis, and whilst all models are wrong, some models are useful. That usefulness ultimately depends on whether a model provides sufficient accuracy to guide interventions and further work effectively. Here we show that many methods can create broadly similar patterns (Figure 2A-C), species specific models will provide the best approach if sufficient species level data are available, but on extended (i.e. global) scales this is rarely done. In most of these examples (Figure 2) the relative patterns are similar, but unfiltered approaches overgeneralise, inflating richness of regions predicted to host lower diversity in Central Africa, whilst missing areas to the Northeast and Southeast, whilst other approaches do highlight the same areas as potentially hosting higher diversity.

Moving forward, a set of best practices are needed to ensure that, as accessible data grow, they are used in a way that builds on understanding rather than bias (e.g., see Box 1). For example, understanding if data are representative across regions and taxa for species level analysis (as noted in Figure 1), as global insect data do not meet these criteria, global analysis is not yet possible.This is because most data are concentrated in North America and Europe and taxonomic representation is too incomplete. For insects, such analysis may be possible within extremely well-studied taxa, such as ants, because work to collate representative data for species level analysis has been conducted (Kass et al., 2022). However, for most insects (and many other organisms) species-specific analysis of protected area coverage or stacked diversity patterns cannot be conducted meaningfully on a global basis, and only coarser metrics of diversity are possible. We also need to prepare for new types of data (such as social-media-generated data), and constantly update guidelines to ensure their effective use. These different types of data can amplify different forms of bias or introduce new biases, and thus knowing how they can be used will likely need continued revisiting, not to mention safeguards to ensure their use is also ethical. As we have seen, the greatest growth in data availability has been in developed areas of high-income economies; if we uncritically apply methods fit only for data-rich areas to data-lacking areas, we risk misleading rather than facilitating effective management.

Whilst it is true that we will continue to generate data, and the issues highlighted here are not new, developing best practice guidelines and facilitating sensible and sensitive data-use remains crucial if we are to lead rather than mislead management and conservation prioritisation. Inappropriate data-use can misrepresent hotspots, and incorrectly gauge levels of protection species and hotspots have. Therefore, ensuring how to use data effectively, however limited or biased those data may be, is key to enabling effective use and solution generation as we continue to grow biodiversity databases.

Box 1; Steps for cleaning data.

Table S1. The possible occurrence data filtering steps and the functions, packages, descriptions, and citations required to undertake each sub-step. Sub-steps include data flagging (adding a column with the test results), data carpentry (changing the data itself), and data filtering (removing occurrences based on data flags). The packages bdc and BeeBDC also have a selection of functions that are useful for data visualisation and

critical for checking the "common-sense" of results.

Step	Function	Package	Brief descrip
Pre-filtering	*bdc_scientificName_empty*	*bdc*	Flag occurrenc
	bdc_coordinates_empty	*bdc*	Flag occurrenc
	bdc_coordinates_outOfRange	*bdc*	Flag occurrenc
	bdc_basisOfRecords_notStandard	*bdc*	Flag occurrenc
	bdc_country_from_coordinates	*bdc*	Get country na
	jbd_CfC_chunker	*BeeBDC*	A chunked and
	bdc_country_standardized	*bdc*	Standardises c
	bdc_coordinates_transposed	*bdc*	Flags occurren
	jbd_Ctrans_chunker	*BeeBDC*	A chunked and
	bdc_coordinates_country_inconsistent & jbd_coordCountryInconsistent	*bdc and BeeBDC*	Flag occurrenc
	bdc_coordinates_from_locality	*bdc*	Extracts occur
	flagAbsent	*BeeBDC*	Flags occurren
	flagLicense	*BeeBDC*	Flags occurren
	GBIFissues	*BeeBDC*	Flags occurren
	clean_fossils	*CoordinateCleaner*	Flags species t
Taxonomy	*bdc_clean_names*	*bdc*	Cleans scientif
	bdc_query_names_taxadb	*bdc*	Harmonizes ta
	harmoniseR	*BeeBDC*	Harmonizes ta
Space	*bdc_coordinates_precision & jbd_coordinates_precision*	*bdc and BeeBDC*	Flags occurren
	coordUncerFlagR	*BeeBDC*	Flags occurren
	clean_coordinates	*CoordinateCleaner*	Flags occurren
	cd_ddmm	*CoordinateCleaner*	Flags occurren
	diagonAlley	*BeeBDC*	Flags records f
	cd_round	*CoordinateCleaner*	Flags occurren
	countryOutlieRs	*BeeBDC*	Flags occurren
	manualOutlierFindeR	*BeeBDC*	Flags occurren
Time	*dateFindR*	*BeeBDC*	Attempts to fi
	bdc_eventDate_empty	*bdc*	Flags occurren
	bdc_year_outOfRange	*bdc*	Flags occurren
Duplicates	*dupeSummary*	*BeeBDC*	Iteratively sea
	cc_dupl	*CoordinateCleaner*	Flags duplicat
Filtering	*bdc_filter_out_flags*	*bdc*	Filters occurre
	bdc_filter_out_names	*bdc*	Filters occurre
	summaryFun	*BeeBDC*	Filters occurre

References

Predictors of contraction and expansion of area of occupancy for British birds. Proceedings of the Royal Society B: Biological Sciences, 281(1786), 20140744.

Brooks, T. M., Pimm, S. L., Akçakaya, H. R., Buchanan, G. M., Butchart, S. H., Foden, W., ... & Rondinini, C. (2019). Measuring terrestrial area of habitat (AOH) and its utility for the IUCN Red List. Trends in ecology & evolution, 34(11), 977-986.

Butchart, S. H., Clarke, M., Smith, R. J., Sykes, R. E., Scharlemann, J. P., Harfoot, M., ... & Burgess, N. D. (2015). Shortfalls and solutions for meeting national and global conservation area targets. Conservation Letters, 8(5), 329-337.

Chesshire, P.R., Fischer, E.E., Dowdy, N.J., Griswold, T.L., Hughes, A.C., Orr, M.C., Ascher, J.S., Guzman, L.M., Hung, K.-L.J., Cobb, N.S. and McCabe, L.M. (2023), Completeness analysis for over 3000 United

States bee species identifies persistent data gap. Ecography e06584. *https://*doi.org/10.1111/ecog.06584

De Araujo, M. L., Quaresma, A. C., & Ramos, F. N. (2022). GBIF information is not enough: national database improves the inventory completeness of Amazonian epiphytes. Biodiversity and Conservation, 31(11), 2797-2815.

A., Falcon-Brindis, A., et al. *(* 2023*)* . BeeBDC: An R package and globally synthesised and flagged bee occurrence dataset. BioRxiv.

Garcia-Rosello, E., Gonzalez-Dacosta, J., Guisande, C., & Lobo, J. M. (2023). GBIF falls short of providing a representative picture of the global distribution of insects. Systematic Entomology.

GBIF Secretariat. (2021). GBIF Science Review 2020.

Herkt, K. M. B., Skidmore, A. K., & Fahr, J. (2017). Macroecological conclusions based on IUCN expert maps: A call for caution. Global Ecology and Biogeography, 26(8), 930-941.

Huang, Q., Lothspeich, A., Hernandez-Yanez, H., Mertes, K., Liu, X., & Songer, M. (2020). What drove giant panda *Ailuropoda melanoleuca*expansion in the Qinling Mountains? An analysis comparing the influence of climate, bamboo, and various landscape variables in the past decade. Environmental Research Letters, 15(8), 084036.

Hughes, A. C. (2018). Have Indo-Malaysian forests reached the end of the road?. Biological Conservation, 223, 129-137.

Hughes, A. C., Qiao, H., & Orr, M. C. (2021a). Extinction targets are not SMART (Specific, measurable, ambitious, realistic, and time Bound). BioScience, 71(2), 115-118.

Hughes, A. C., Orr, M. C., Ma, K., Costello, M. J., Waller, J., Provoost, P., ... & Qiao, H. (2021b). Sampling biases shape our view of the natural world. Ecography, 44(9), 1259-1269.

Jarrell, G. H., Ramotnik, C. A., & McDonald, D. L. (2010). ARCTOS: a relational database relating specimens, specimen-based science, and archival documentation.

apatric species using distribution models and support vector machines. Ecological Applications, 31(1), e02228.

3). Darwin's finches. Scientific American, 188(4), 66-73.

Leandro, C., Jay-Robert, P., & Vergnes, A. (2017). Bias and perspectives in insect conservation: a European scale analysis. Biological Conservation, 215, 213-224.

Li, J., Hughes, A. C., & Dudgeon, D. (2019). Mapping wader biodiversity along the East Asian—Australasian flyway. PloS one, 14(1), e0210552.

Liu, S., Xia, S., Wu, D., Behm, J. E., Meng, Y., Yuan, H., ... & Yang, X. (2022). Understanding global and regional patterns of termite diversity and regional functional traits. Iscience, 25(12), 105538.

Mace, G. M., Barrett, M., Burgess, N. D., Cornell, S. E., Freeman, R., Grooten, M., & Purvis, A. (2018). Aiming higher to bend the curve of biodiversity loss. Nature Sustainability, 1(9), 448-451.

Poggiato, G., Munkemuller, T., Bystrova, D., Arbel, J., Clark, J. S., & Thuiller, W. (2021). On the interpretations of joint modeling in community ecology. Trends in Ecology & Evolution, 36(5), 391-401.

Potapov, A. M., Guerra, C. A., van den Hoogen, J., Babenko, A., Bellini, B. C., Berg, M. P., ... & Scheu, S. (2023). Globally invariant metabolism but density-diversity mismatch in springtails. Nature communications, 14(1), 674.

Qiao, H., Orr, M., Hughes, A. C. (2023). Measuring metrics: what biodiversity indicators are most appropriate for different forms of data bias. OSF-preprints. https://osf.io/jgqst/download

Ribeiro, B.R., Velazco, S.J.E., Guidoni-Martins, K., Tessarolo, G., Jardim, L., Bachman, S.P., and Loyola, R. (2022). bdc: A toolkit for standardizing, integrating and cleaning biodiversity data. Methods Ecol. Evol. 13, 1421-1428. https://doi.org/10.1111/2041-210X.13868.

cology & Evolution, 5(10), 1325-1327.

Wyborn, C., & Evans, M. C. (2021). Conservation needs to break free from global priority mapping. Nature Ecology & Evolution, 5(10), 1322-1324.

Zizka, A., Silvestro, D., Andermann, T., Azevedo, J., Duarte Ritter, C., Edler, D., Farooq, H., Herdean, A., Ariza, M., Scharn, R., et al. (2019). CoordinateCleaner: Standardized cleaning of occurrence records from biological collection databases. Methods Ecol. Evol. 10, 744-751. https://doi.org/10.1111/2041-210X.13152.

Zurell, D., Pollock, L. J., & Thuiller, W. (2018). Do joint species distribution models reliably detect interspecific interactions from co-occurrence data in homogenous environments?. Ecography, 41(11), 1812-1819.

Zurell, D., Zimmermann, N. E., Gross, H., Baltensweiler, A., Sattler, T., & Wuest, R. O. (2020). Testing species assemblage predictions from stacked and joint species distribution models. Journal of Biogeography, 47(1), 101-113.

Supplemental Methods

IUCN richness maps

We developed a range of maps of richness using different techniques outlined here. Firstly, as a comparator we mapped richness using IUCN polygons as well as trimming the IUCN polygons as advocated by Brooks et al 2019. To map richness of African bats based on the IUCN (though this does not include all known species) we downloaded as all Chiroptera from the African continent from the IUCN website (on June 3rd). Firstly we clipped the polygons for a rectangular area over the African continent, then dissolved each species polygon to give a single polygon for each species. We used the "count overlapping polygons" toolbox, split the polygon. To do this we split the polygon into groups of 30 species (the tool works better on lower numbers), numbered outputs sequentially, then converted to rasters (with a resolution of 0.008 degrees, approximately 1km at the equator) then used the raster calculator to add all results to map richness according to the IUCN.

Simple polygon richness maps

To map richness based on point-based data we downloaded data from a number of sources (see supplemental data), this included the ACR bat database (using the 2023 dataset), as well as additional datasets (Hughes et al., 2017; Tanalgo et al., 2022). In addition, we downloaded updates to the bat interactions database (May 2023) and data collected since 2000 stored in the Arctos dataset, plus the updated bat records in GBIF for 2021-2023 (shown in supplemental data). This data was then cleaned to remove synonyms and correct spellings throughout. This yielded a dataset with 38208 points for 343 species.

Minimum convex polygons were created for each species using the "minimum bounding geometry" tool in ArcMap 10.8. We then trimmed this by the outline of the African continent, then split these into a geodatabase using the split by attributes tool using "species" as the attribute. MCPs were then treated in two ways, the overall map of unfiltered MCP richness was calculated in the same way as the IUCN simple richness using the count overlapping polygons toolbox in ArcMap to show how unfiltered maps of richness misrepresent range.

Habitat trimmed maps

We needed to filter out habitat. This step is key, and likely will introduce errors, as whilst widely advocated, most landuse maps misrepresent actual landcover by delineating landcover types in a fairly arbitrary way. This will over-emphasize the importance of more sampled habitats, and may miss undersampled and smaller habitats; thus the choice of this layer is key. Here we used the IUCN ecosystem typology (published in 2022) because it provides overlapping classes of ecologically relevant landcover and thus provides the most nuanced mechanism to map landcover for regions where accurate national maps of landuse may not be available.

National maps are generally better calibrated for a given region than global maps, which must necessarily simplify land-cover categories, and may not be sufficiently tested at a local level. Furthermore, the IUCN ecosystem typology is recognised by the UN System of Environmental Economic Accounting (SEEA) and is therefore already a global standard. An alternative here if the region of interest is predominantly forested is to use tree density or height as the delimiter for habitats, however, in a region with diverse habitats, more nuance may be needed given that trimming in this way cannot account for climate differences across the recorded maximum bounds of species distribution. Notably urban areas were not included as biases in sample collection to more populated areas can skew results (Hughes et al., 2021a).

After removing duplicate species points we then extracted the landcover types based on the typology under all locality points. We then used the summary statistics tool to calculate the total number of points per species in each landcover category as well as overall, then used this to calculate the percentage of locality points for each species within each category. We then filtered out habitat with at least 10% of locality points within each landcover category which fell within the polygon for the species. To do this we masked all habitat types suitable for the species with their individual MCP then mosaiced all parts of the species range together using the mosaic to new raster tool, all suitable habitats were given a value of one. Because of errors in the mosaic to new raster tool if applied to stack richness for large numbers of species we then mosaiced all species range maps onto a mask of the African region, then used the raster calculator to sum richness for groups of 40 species, numbered these sequentially, then added the numbered MCP filtered outputs to sum overall MCP filtered richness.

In addition we used the same filtered habitat requirements and repeated the process using the IUCN polygons rather than the MCPs. This has the advantage of better representing overall species range (MCPs will necessarily not include the edges of species ranges as they can only delimit the maximum bounds of recorded range), but we know from former work IUCN maps include pervasive biases and also artificially underestimate many parts of species ranges (Hughes et al., 2021, Li et al., 2020). Furthermore, IUCN maps are not available for most plants and invertebrates, as well as being notably incomplete in some small bodied vertebrate taxa. Filtered IUCN maps were then stacked and richness calculated in the same way as MCP filtered richness using the raster calculator.

Projecting richness from inventories

Where insufficient data exists for species specific modelling, projecting richness from inventories may be the only possibility. This requires a lower resolution than many other approaches (as inventories must be drawn from an area) and can also not reflect biogeographic differences, meaning that where island biogeography is important it cannot account for that. Projecting richness using this approach relies on site-based inventories of species present. To do this we first used the same dataset as above, we created a $10km^2$ fishnet as a polygon grid. The fishnet was trimmed to continental boundaries using the same clip as previously used. This was then imported along with point data into QGIS 3.26.3 and the sampling point tool used to intersect grids with point data, with the FID of each grid used in an additional column as a unique ID for the grid. We then reimported this into ArcMap 10.8 and used the summary statistics tool to calculate the number of points and species per grid cell. This was then reconnected to the original grid using joins and relates, and cells with at least 30 unique records were selected as inventories, the latitudes and longitudes of the centroids added, and this data exported to a CSV. To better reflect appropriate data from adequately inventoried but potentially less diverse sites we then added the publically available data from the Darkcide database (Tanalgo et al 2022) once we removed any listings with only a single species present (which may reflect species specific inventories, for example a number of sites only listed *Otomops* species). The final dataset had 417 inventories for the region which were then used for modelling.

This data could then be modelled for richness. This has the advantage that we can use the same variables as used for species specific modelling, including not requiring the use of a landcover map as a filter and thus enabling more nuance. Variables selected are noted in supplemental data, these were chosen to reflect climate parameters, and continuous metrics of habitat structure. Variables included actual evapotranspiration, annual mean potential evapotranspiration, aridity, two metrics of distance to bedrock (bdticm, bedroom

from ISRIC world soil grids, as well as Estimated soil organic carbon stock as a measure of fertility), continental moisture index, continentality, embergers pluviothermic quotient, growing degree days 5, potential evapotranspiration of the driest quarter (resources during the most limiting time of year) potential evapotranspiration seasonality, thermicity (from Envirem) and bio 3,4,5,6,12,13,14 and 15 from Worldclim as well as vegetation canopy height.

Therefore for both forms of modelling (richness modelling based on inventories and species modelling) we used canopy height of all vegetation, a range of soil variables, and variables representative of climate, moisture and seasonality. Richness inventories were then grouped into classes (i.e. under 5 species, 5-10 species, etc) and modelled as individual classes using Maxent, all outcomes had an AUC of over 0.9. Within Maxent we modelled each richness level, ran three replicates (and used an average) and used default parameters. The average was then reclassified in ArcMap 10.8 using the 10 percentile cloglog threshold as a minimum bound of suitability, and then using an equal division between the threshold and the maximum value of 1 to reflect the maximum and minimum values of the richness level, with areas "unsuitable" given a value of 0. The mosaic to new raster tool was then used with the selection set to "maximum" to give the maximum number of species any given area was suitable for based on model outcomes.

Species specific modelling

When representative data (spatial and taxonomic) is available species-specific models are always the best approach. Using the same point data outlined above with the same environmental variables (but at resolution of 1km (0.008 degrees) rather than 10km (0.08 degrees)). Species with at least 3 points were modelled (this is too low for individual species models, but can still be useful in an approach where only patterns are being examined, though output accuracy can then be assessed by comparing to the IUCN ranges as whilst ranges within IUCN are artificial the overall patterns in terms of geographic regions should be broadly comparable). As noted variables should represent the conditions likely to delimit species ranges, reflect ecophysiological thresholds, and other habitat constraints, including habitat structure. Variable correlation is accounted for within Maxent, and species-specific variable selection can also be made using various R packages (i.e. ENMEval).

Once again, we ran models using the default parameters and used 3 replicates, and the average reclassified to a binary map (0:1) using the 10 percentile cloglog threshold. As models will note all environmentally suitable habitat regardless of biogeographic constraints, to reflect biogeography we then used the IUCN redlist to note species endemic to Madagascar, in Madagascar and continental Africa, and those limited to the African continent. It should be noted that a number of Madagascan endemic species lacked sufficient data to model, thus richness in Madagascar may be under-estimated, especially given the small ranges of some bats, but all models had an AUC exceeding 0.9, additional indicies such as Boyce index or AIC can be used to give independent measures of model performance and accuracy. These were then masked in batches using masks of each of those three regions. The raster calculator was then used to batch sum groups of 30 species within each of these three groups and numbered sequentially. We then combined all those species restricted to continental Africa using the raster calculator, before using the mosaic to new raster tool to combine the three types of biogeographic map to refine ranges and map modelled richness.

All maps are provided to show how they note richness patterns across the African continent.

Supplemental data

1-Species data

ACR data downloaded June 1st 2023 https://africanbats.org/publication/african-chiroptera-report-2022/

Arctos Data Downloaded June 1st 2023 https://arctos.database.museum/download.cfm?file=ArctosData0Iq7LVJgPN.csv

Alroy, J. (2019). Latitudinal gradients in the ecology of New World bats. Global Ecology and Biogeography, 28(6), 784-792.

Bat-Eco Interactions database Downloaded June 1st blob:https://batbase.org/05e4c768-ec29-43a2-b9d1-7fc7855484e8

Becker, D. J., Crowley, D. E., Washburne, A. D., & Plowright, R. K. (2019). Temporal and spatial limitations in global surveillance for bat filoviruses and henipaviruses. Biology Letters, 15(12), 20190423.

Di Gregorio, C., Iannella, M., & Biondi, M. (2021). Revealing the role of past and current climate in shaping the distribution of two parapatric European bats, Myotis daubentonii and M. capaccinii. The European Zoological Journal, 88(1), 669-683.

GBIF.org (06 May 2023) GBIF Occurrence Downloadhttps://doi.org/10.15468/dl.xc2jvq

GBIF.org (08 May 2022) GBIF Occurrence Download https://doi.org/10.15468/dl.82sgat

Jargalsaikhan, A., Ariunbold, B., Dalannast, M., Purevee, E., Paek, W. K., Tsagaan, K., & Ganbold, O. (2022). Molecular phylogenetic analysis of bats in the family Vespertilionidae in Mongolia. Journal of Asia-Pacific Biodiversity, 15(3), 329-335.

Kruskop, S. V., Artyushin, I. V., Yuzefovich, A. P., Undrakhbayar, E., Speranskya, A. S., Lisenkova, A. A., ... & Lebedev, V. S. (2020). Genetic diversity of Mongolian long-eared bats (Plecotus; Vespertilionidae; Chiroptera). Acta Chiropterologica, 22(2), 243-255.

Kuo, H. C., Chen, S. F., Fang, Y. P., Flanders, J., & Rossiter, S. J. (2014). Comparative rangewide phylogeography of four endemic T aiwanese bat species. Molecular Ecology, 23(14), 3566-3586.

Mokrani, Y., Mimeche, F., Nouidjem, Y., & Saheb, M. (2018). Rapid assessment of cave-dwelling bat diversity in the Chebket ES-Sellaoua Mountains (Eastern Algeria). Arxius de Miscel lania Zoologica, 16, 112-120.*

Raman, S., Shameer, T. T., Pooja, U., & Hughes, A. C. (2023). Identifying priority areas for bat conservation in the Western Ghats mountain range, peninsular India. Journal of Mammalogy, 104(1), 49-61.

Scherrer, D., Christe, P., & Guisan, A. (2019). Modelling bat distributions and diversity in a mountain landscape using focal predictors in ensemble of small models. Diversity and Distributions, 25(5), 770-782.

Tanalgo, K. C., Tabora, J. A. G., de Oliveira, H. F. M., Haelewaters, D., Beranek, C. T., Otalora-Ardila, A., ... & Hughes, A. C. (2022). DarkCideS 1.0, a global database for bats in karsts and caves. Scientific Data, 9(1), 155.

Zhou, H., Ji, J., Chen, X., Bi, Y., Li, J., Wang, Q., ... & Shi, W. (2021). Identification of novel bat coronaviruses sheds light on the evolutionary origins of SARS-CoV-2 and related viruses. Cell, 184(17), 4380-4391.

Variable	**Reference**
Climate layers	*Worldclim https://data.biogeo.ucdavis.edu/data/worldclim/v2.1/base/wc2.1_30s_bio.zip*
Tree density	*Crowther, T. W., Glick, H.B., Covey, K. R., et al. (2015). Mapping tree density at a global sc*
Soil layers	*ISRIC Soil grids downloaded May 8th 2023. https://data.isric.org/geonetwork/srv/eng/catalog.*
Vegetation height	*Lang,N., Jetz, W., Schindler, K, Wegner, J.D.(2019) A high-resolution canopy height model o*
IUCN ecosystem typology	*https://global-ecosystems.org/analyse?area=-18.984375+-38.272689%2C-18.984375+39.909736*
Soil water balance	*Trabucco, A.,Zomer, R. J. (2019): Global High-Resolution Soil-Water Balance. figshare. Datas*