The collector practices that shape spatial, temporal, and taxonomic bias in herbaria

Ryan J. Schmidt[1]*, Kristen E. Saban[1], Lena Struwe[2,3], Charles C. Davis[1*]

[1] Department of Organismic and Evolutionary Biology, Harvard University Herbaria, Harvard University, 22 Divinity Avenue, Cambridge, MA 02138, USA

[2] Department of Ecology, Evolution & Natural Resources, Rutgers, The State University of New Jersey, 14 College Farm Road, New Brunswick, NJ 08901-8551, USA

[3] Department of Plant Biology, Rutgers, The State University of New Jersey, 59 Dudley Road, New Brunswick, NJ 08901-8551, USA

* Corresponding authors: ryanschmidt@g.harvard.edu, cdavis@oeb.harvard.edu

RJS: 0000-0002-4907-2270

KES: 0009-0009-5728-4001

LS: 0000-0001-6074-5758

CCD: 0000-0001-8747-1101

Word Count: 6334

Introduction: 730
Materials & Methods: 1365
Results: 2115
Discussion: 2124

Figures:

Manuscript: 8 figures. We would prefer all figures to be in color, however, figures 1 and 6 may be printed in black and white if necessary.

Supporting Information: 4 tables and 1 figure.

Summary

- Natural history collections (NHCs) are essential for studying biodiversity. Although spatial, temporal, and taxonomic biases in NHCs affect analyses, the influence of collector practices on biases remains largely unexplored.
- We utilized one million digitized specimens collected in the northeastern United States by ~10,000 collectors to investigate (a) how collector practices shape spatial, temporal, and taxonomic biases in NHCs and (b) similarities and differences between practices of more- and less-prolific collectors
- We identified six common collector practices, or collection norms: collectors generally collected (a) different species, (b) from multiple locations, (c) from sites sampled by others, (d) during the principal growing season, (e) species identifiable outside peak collecting months, and (f) species from species-poor families and genera. Some norms changed over decades, with different taxa favored during different periods. Collection norms have increased taxonomic coverage in NHCs, however, collectors typically avoided large, taxonomically-complex groups, causing their underrepresentation in NHCs. Less-prolific collectors greatly enhanced coverage by collecting during more months and from less-sampled locations.
- We assert that overall collection biases are shaped by shared predictable collection norms rather than random practices of individual collectors. Predictable biases offer an opportunity to more effectively address biases in future biodiversity models.

Keywords

herbaria; natural history collections; history of science; collection norms; biodiversity; digitization; biodiversity modelling

Introduction

Discovering and describing global patterns of species diversity and distribution remains a fundamental priority for biodiversity scientists (CBD, 2022). Although recent advances in biodiversity modeling have greatly improved our understanding of these factors, the vouchered specimens and observational data underlying these models are know to exhibit significant spatial, temporal, and taxonomic biases that remain largely unaccounted for (Meyer *et al.*, 2016; Daru *et al.*, 2018).

Herbaria and other natural history collections (NHCs) are invaluable resources for understanding global biodiversity (Funk, 2003; Johnson *et al.*, 2023; Davis, 2023, 2024; Marín-Rodulfo *et al.* 2024). The extensive sampling of NHCs over time, space, and taxa complement long-term monitoring programs such as the Atlas of the British Flora (Perring & Walters, 1962; Preston, 2013) and the USDA's Forest Inventory and Analysis (Rudis, 2003; FIA, 2023), which have provided important insights into species distributions but are limited across these key axes in important ways. Although biodiversity is not randomly distributed, to best represent biodiversity NHCs would ideally provide a representative sample of global biodiversity across time, space, and taxa. Any deviations between a spatially, temporally, and taxonomically representative sample and the representation of biodiversity in NHCs are examples of collection bias. Understanding how NHCs diverge from this ideal coverage allows us to better account for biases in our biodiversity models and discern what questions we can address using these collections. Ultimately, understanding collection biases will help guide the application and development of statistical tools to correct for biases, develop better priorities for future collecting efforts, and help us achieve more comprehensive and accurate models of global biodiversity.

Comprehensive digitization of natural history specimens from large geographic/floristic regions has revealed key spatial, temporal and taxonomic biases in NHCs (Meyer *et al.*, 2016; Daru *et al.*, 2018; Kozlov *et al.*, 2021; Eckert *et al.*, 2024). These overall biases in NHCs are a consequence of the spatial, temporal, and taxonomic collection practices of each collector—what we call collector practices. Previous studies have highlighted the connection between collector practices and overall bias in collections, documenting that a small number of mega-collectors have made disproportionately large contributions to species discovery (Bebber *et al.* 2012) and to specimen collections in NHCs (Daru *et al.* 2018). The disproportionately large impact of these mega-collectors raises an important but unanswered question: have highly prolific collectors also contributed disproportionately to the biases documented in these collections? To date, there have been no efforts to investigate how the collector practices of all collectors in a region have contributed to overall bias in NHCs. Moreover, there have been no large-scale efforts to understand the impact that less-prolific collectors have had on the spatial, temporal, and taxonomic coverage in collections.

Here, we expand the current framework for investigating biases in NHCs (*sensu* Daru *et al.*, 2018) by explicitly examining how collection biases are shaped by the practices of individual collectors which, to our knowledge, has not been broadly examined. . As a test case for our investigation, we leverage the nearly completely digitized metaherbarium that extensively documents the flora of the northeastern United States (i.e., all digitally available specimens collected in the northeastern US and housed throughout the world; Schorn *et al.*, 2016; Sweeney *et al.*, 2018; Hedrick *et al.*, 2020). Specifically, we use all digitized herbarium specimens of land plants (i.e., bryophytes and vascular plants) collected in the northeastern United States from the earliest digitized record to the present (i.e., 1781–2024). We reconstruct the contributions of collectors to investigate how overall bias in NHCs are shaped by the similarities and differences in collection practices of different collectors. We assess the relationship between these collection practices and the number of collections by each collector on a continuous scale with more- and less-prolific collectors representing opposite ends of this continuum. Mega-collectors—who have contributed a disproportionately large amount of specimens (*sensu* Daru *et al.*, 2018)—represent the uppermost extreme of this spectrum. We also investigate how what we term *collection norms*—the collector practices shared by all collectors—have influenced overall biases in NHCs. Such synthetic investigations further demonstrate the growing utility of digitized specimens within the framework of the extended specimen (Webster, 2017; Lendemer *et al.*, 2020), facilitating proper attribution for the thousands of hidden heroes that have made meaningful but previously unrecognized contributions to NHCs (Groom *et al.*, 2022) and enabling ongoing efforts to better model biodiversity in an era of rapid ecological change.

Materials and Methods

Data collection & data cleaning

We downloaded 2,365,287 records representing all digitized herbarium specimens of land plants from the northeastern United States (i.e., Connecticut, Maine, Massachusetts, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island, and Vermont; hereafter the Northeast) from GBIF (GBIF.org, 2024). These specimens are housed in 237 herbaria around the world (Table S1). We then filtered this dataset to remove the 548,895 records without a transcribed date, collector, locality, or species-level identification. This filtering left us with 1,816,392 analyzable records.

Georeferencing

About half of the cleaned records (920,633 records) contained transcribed coordinates. We batch-georeferenced an additional 401,450 specimens to municipal centroid points (i.e., the centroid points for local incorporated communities such as cities, towns, and townships; CT

DEEP, 2023; PennDOT, 2024) and removed all records that could not be georeferenced to a specific municipality (503,563 records removed). Although this method of georeferencing does not capture fine-scale differences in collection localities (Park & Davis, 2017), it is consistent with the precision for many herbarium georeferencing initiatives in the northeastern US (e.g., Mancini *et al.*, 2019) and suitable for analyses on these large spatial scales. We removed records with coordinates outside of the northeastern US (United States Census Bureau, 2024) using the `st_intersection()` function from the *sf* package in R 4.4.1 (Pebesma 2018; Pebesma & Bivand 2023; 10,254 records removed). This resulted in a total of 1,311,829 georeferenced records (see Fig. S1 for more information about the specimens removed at each step of data cleaning).

Collector disambiguation

Due to institutional differences in transcription practices, incorrect transcriptions, and orthographic variations in collector names, assigning different text strings (i.e., recordedBy strings in DarwinCore; hereafter "collector strings") to a single collector can be difficult and time consuming for large datasets (Groom *et al.*, 2022). Thanks to the large-scale availability of digitized historical and genealogical records (e.g., Ancestry.com, MyHeritage.com, and Newspapers.com) and recent initiatives by historians of science to identify and disambiguate the names of people who collected natural history specimens (e.g., Bionomia; Shorthouse, 2024; Weeks *et al.*, 2024), we are for the first time able to identify and reconstruct what we call oeuvres—all of the specimens a person has collected—of all contributors to a regional flora.

To disambiguate collector strings, we extracted the first collector in each collector string, separating what we consider the principal collector (henceforth referred to as the collector) from any associated collectors. Although associated collectors are crucial parts of any collection team and deserve proper credit for their efforts, we focused our analysis on principal collectors in this initial study. Our rationale is that the principal collector is usually responsible for recording field notes and is likely to take on the major role of depositing the specimens in an herbarium collection. We then separated the collector strings into words using the `unnest_tokens()` function from `tidytext` (Silge & Robinson, 2016) and concatenated these words in alphabetical order to standardize different transcriptions of the same text (e.g., "C. F. Parker", "C F Parker", and "Parker, C. F." would all become "c,f,parker"). We then merged all records with identical concatenated strings and manually validated each cluster—merging records with different concatenated strings that represent the same collector—to ensure that each cluster represented a single collector. We used biographical information from historical and genealogical databases (e.g., Ancestry.com and Newspapers.com) and databases of natural history collectors (i.e., Bionomia and Harvard Index of Botanists; Shorthouse, 2024; Harvard University Herbaria, 2024) to reconstruct the oeuvres of collectors that collected under multiple names, including their spouses' names. For instance, we identified "Mrs. C. S. Phelps" as Ora Almira Phelps (née Parker) who collected under the names Mrs. Charles Sheppard Phelps, Orra A. Phelps, Mrs. O. P. Phelps, and Orra Parker Phelps.

We excluded any collector strings that were ambiguous either because of obvious transcription errors that could not be verified with a digital image of the specimen or had limited information. To ensure that we were not conflating multiple collectors, we excluded records with only initials (e.g., C.A.B.), only a surname (e.g., Boice), or only the initial of the first name and the surname (e.g., C. Boice; 233,321 records removed,1,078,508 records remaining). We then removed duplicate specimens (i.e., specimens collected by the same collector with the same specimen number in DarwinCore's `recordNumber` field) so that each collection event is represented by a single specimen (89,251 records removed). We did not remove any specimens without a transcribed specimen number (i.e., those with "s.n.", "sn", or a blank `recordNumber` field) since we could not confirm that that these specimens were duplicate collections. This resulted in our final dataset of 989,257 specimens (Table S2).

Temporal Bias

To investigate temporal trends in botanical collections, we calculated the number of specimens, distinct species, sampling localities, and active collectors for each year during 1781–2024. We then evaluated the relationship between these metrics and the oeuvre size of each collector on a continuous scale from less-prolific (small oeuvres) to more-prolific collectors (large oeuvres). We investigated seasonal variations in collection intensity by comparing the number of specimens collected in each month and analyzed how this distribution changed with respect to the oeuvre size of the collector who gathered the specimen.

Spatial Bias

We quantified spatial bias by gridding the georeferenced specimens into 10-km grid squares (hereafter localities) to help mitigate the effects of batch georeferencing and create equal-area polygons for comparison (Franklin & Miller, 2009; Schmidt *et al.*, 2023). We calculated the revisitation proportion for each collector as the number of specimens per unique collecting locality. We also calculated the average oeuvre size of collectors active in each locality, weighted by the number of collections of each collector (higher values indicate more activity by highly prolific collectors) to investigate the geographical bias of more- versus less-prolific collectors.

To understand how collectors of different sizes contributed to overall spatial sampling, we found the number of unique grids sampled for different subsets of the data. To determine if more- or less-prolific collectors expand overall spatial coverage, we arranged specimens by decreasing and increasing oeuvre size, respectively and found the number of unique grids sampled for increasingly larger subsets of the data in 10,000 specimen increments (i.e., after arranging by oeuvre size, we extracted the first 10,000 specimens, first 20,000 specimens, 30,000 specimens, etc.). We assessed how spatial bias from collectors with different oeuvre sizes differs from two different null models: we randomly ordered specimens from our dataset to determine if collections by more- or less-prolific collectors are more spatially clustered than the overall specimens (randomized specimens); and simulated a new dataset by randomly sampling from all localities in the northeastern US to determine how collections differ from spatially random collections (simulated random sampling).

Taxonomic Bias

To determine the relative representation of different taxa in herbarium collections, we calculated collection depth as the average number of specimens per species in a given taxon in the northeastern US (i.e., total specimens/unique species for each genus and family based on the `acceptedScientificName` field from GBIF). We evaluate taxon size on a continuous scale, whereby taxa with fewer species in the northeastern US are considered smaller and those with more species are considered larger. Taxa with higher collection depths were considered better represented in herbaria.

To assess how frequently collectors collect a species that they have already collected, we calculated the proportion of species re-collected by each collector (i.e., total specimens/unique species for each collector). Collectors who collected many specimens of the same species would have a high re-collection proportion while those that collected only one specimen of each species would have a re-collection proportion of one.

To investigate whether some taxa (i.e., species, genera, and families) were favored by collectors over other taxa, we plotted the number of collections per taxon against the number of collectors who collected each taxon. We fit a generalized additive model (GAM) to these points to estimate how many collectors we expected to have collected each taxon based on the total number of specimens of that taxon. Taxa that fell above this GAM curve were collected by more people than expected (hereafter, favored taxa) and taxa that fell below the curve were collected by fewer people than expected (hereafter, commonly avoided taxa).

Results

Collectors

We identified 9247 collectors who collected plant specimens in the northeastern US (Fig. 1a; Table S3). This is no doubt an underestimate of the total number of people who have contributed to collections in the region since many collectors were excluded from our analysis due to incomplete or ambiguous collector names and insufficient locality information (45% of analyzable specimens removed) and those whose specimens have yet to be digitized. There was a large variation in the number of specimens that each collector collected. We do not define a threshold between more- and less-prolific collectors for any of our temporal, spatial, or taxonomic analyses and instead evaluate variation in collector practices along a continuum of oeuvre sizes with more- and less-prolific representing opposite ends of this range. However, we briefly present results for some subsets of collectors below to demonstrate the overall variation in the contributions of collectors with different oeuvre sizes. The vast majority (more than 90%; 8385 people) collected fewer than 100 specimens. Only 1.8% of collectors (171 people) collected more than 1000 specimens (contributing 71% of the total number of collections). The most prolific collector in our dataset was Robert L. Schaeffer, Jr., who collected 50,287 specimens (Fig. 1b). Half of all specimens from the northeastern US were collected by only 57 collectors (0.6% of collectors). Most collectors (70%; 6,549 people) collected fewer than ten specimens (contributing 1.5% of collections).

People who collected less than 1000 specimens tended to collect only one specimen of each species (Fig. 1b) and about ten specimens per locality (Fig. 1c). For collectors who collected more than 1000 specimens, they tended to collect only one specimen for each species for the first 1000 specimens they collected. After collecting about 1000 specimens, they collected multiple specimens of the same species, and the number of species they collect plateaus near 2000 species. E. H. Eames collected the most plant species of any collector in our dataset

(2574 species, both vascular and nonvascular; Whelan, 1948). Most people collected either vascular plants (85%) or non-vascular plants (7%), with only 8% collecting both types.

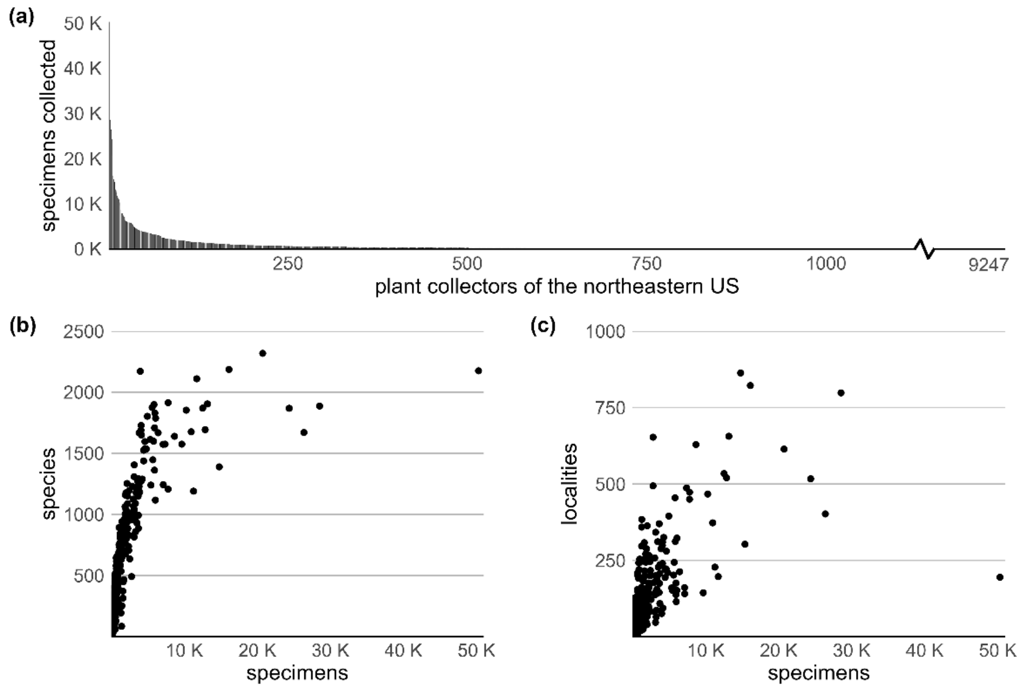

Figure 1. We identified 9247 people who collected herbarium specimens in the northeastern US. The bar plot shows (a) the total number of unique specimens for each plant collector in the northeastern US. The scatter plots show the relationship between the number of specimens each person collected and (b) the number of species they collected and (c) the number of localities in which they collected.

Temporal Bias

The number of collectors active in a given year has varied substantially through time with peaks during 1880–1916 and again during 1932–1941 (Fig. 2a–d). The number of active collectors is strongly correlated with the number of specimens collected in a given year (cross-correlation value of 0.90, p<0.001), species (0.94, p<0.001), and localities (0.90, p<0.001). The number of specimens collected (Fig. 2a), species collected (Fig. 2b), and collectors active in a given year (Fig. 2c) also peaked during 1880–1916 and 1935–1941 whereas the number of sampling localities peaked only from 1935–1941 (Fig. 2d). All metrics have declined since 1950.

About 90% of specimens from the northeastern US were collected during spring and summer (i.e., May to September)—the main growing season in northern temperate zones—with relatively few specimens collected during off-peak months (i.e., from October through April; Fig. 3). The highest proportion of collections by less-prolific collectors were also during May–September. However, collections by more-prolific collectors had a much narrower temporal distribution with collections almost exclusively from June, July, and August.

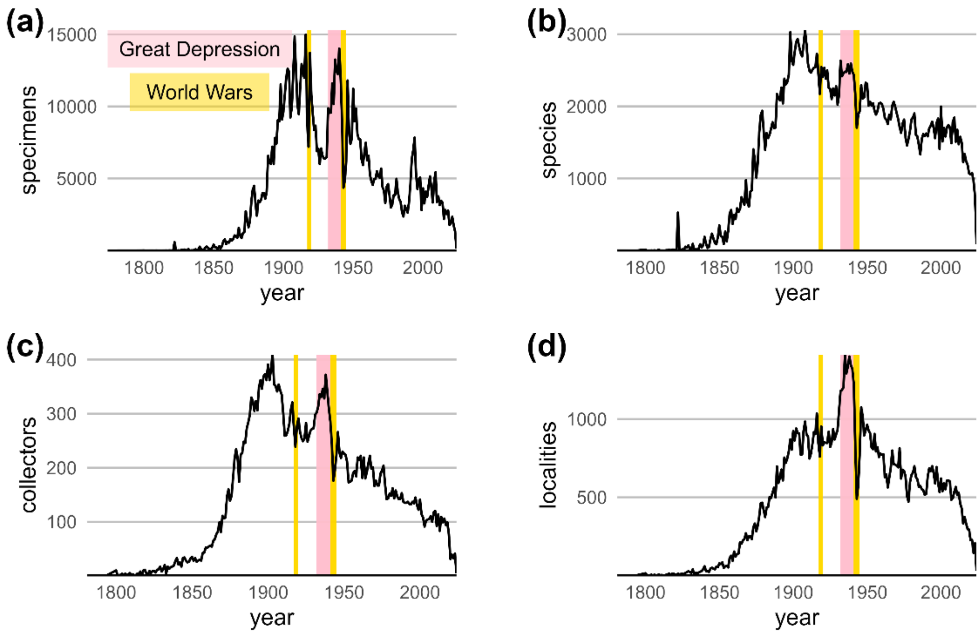

Figure 2. The line plots show the annual variation in (a) the number of specimens collected, (b) the number of species collected, (c) the number of active collectors, and (d) the number of localities in which specimens were collected from 1781–2024. The yellow bars indicate the years when the US was involved in World Wars I and II (1917–1920 and 1941–1946, respectively) and the pink bars represent the duration of environmental projects sponsored by the US federal government during the Great Depression (1929–1939).

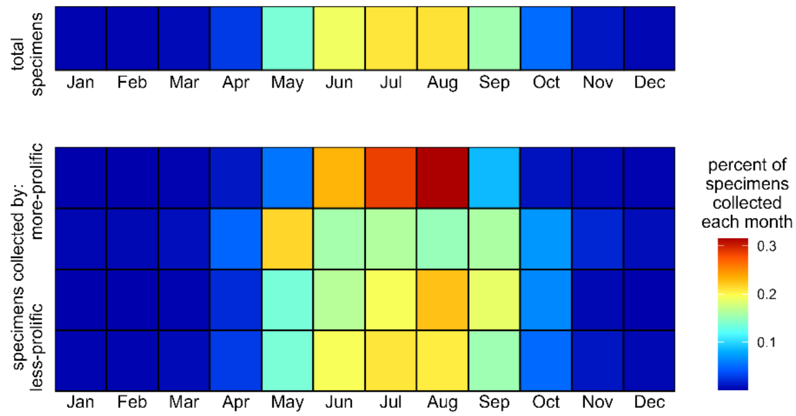

Figure 3. This graph shows the percentage of specimens collected in each month for all specimens (total specimens) and divided into quartiles based on oeuvre size (i.e., going from the first quartile of specimens collected by the least prolific collectors at the bottom to the fourth quartile of specimens collected by the most prolific collectors at the top).

Spatial Bias

The specimens collected by more-prolific collectors were more spatially clustered and had lower geographic coverage than those collected by less-prolific collectors (Fig. 4). Additionally, collections by less-prolific collectors included areas not represented by more-prolific collectors, but more-prolific collectors did not capture areas not represented by less-prolific collectors.

Certain spatial clusters that dominate overall specimen clustering in the northeastern US are driven almost exclusively by collections from more-prolific collectors (Fig. 5). Some of the areas with the highest collection density are driven by a few, prolific collectors (e.g., the hotspot in near Allentown, PA is driven primarily by R. L. Shaeffer, Jr.), whereas other areas with high collection density are driven by many less-prolific collectors (e.g., many of the hotspots in upstate NY). The overall density of collections and the different drivers of collection intensity change quickly over some state borders. For example, there are dense collections in PA and very sparse collections in adjacent NY.

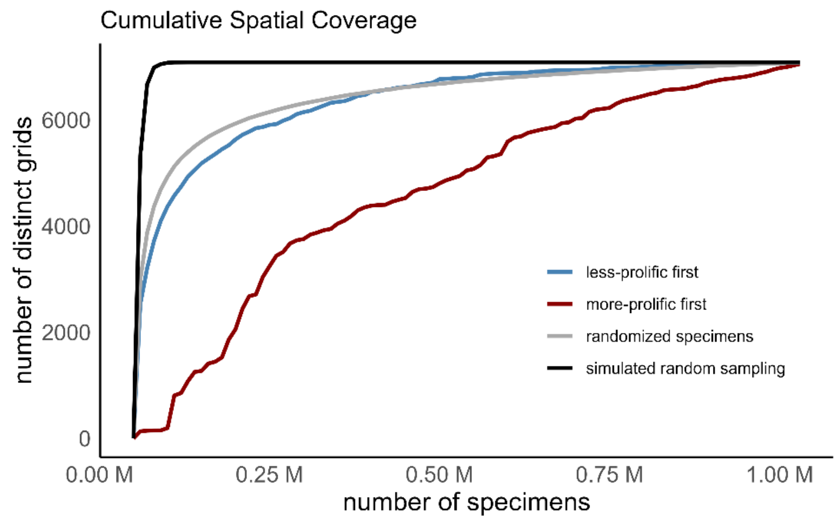

Figure 4. Accumulation curves for the cumulative spatial coverage of gridded herbarium specimens based on the oeuvre size for collectors. Each curve contains all 989,257 specimens in our dataset with specimens added in different orders to demonstrate differences in the spatial coverage of specimens collected by more- and less-prolific collectors. Specimens were added by decreasing oeuvre size for the red curve (more-prolific collectors added first); increasing oeuvre size for the blue curve (less-prolific collectors added first); and in a random order independent of oeuvre size for the gray curve (randomized specimens, median of 99 permutations). The black curve shows randomly simulated specimens to represent our null model of random spatial sampling in the region (simulated random sampling).

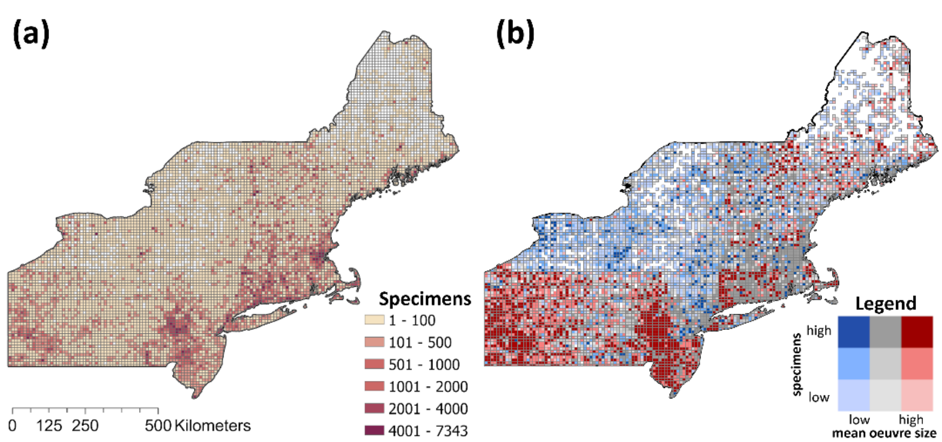

Figure 5. The maps show (a) the density of collections in the northeastern US and (b) the relationship between collection density and areas where collections have been driven primarily by less-prolific collectors (blue; localities in the lowest tercile based on

mean oeuvre size), more-prolific collectors (red; localities in the middle tercile based on mean oeuvre size), or a mix of collector types (gray; localities in the top tercile based on mean oeuvre size).

Taxonomic Bias

Smaller genera are more likely to have a greater collection depth than larger genera; the same is the case for smaller families (Fig. 6). Despite the overrepresentation of smaller genera, several of the most frequently collected species are from large genera (e.g., three species of *Carex*; for a list of the hundred most frequently collected species, see Table S4). Ferns are dramatically overrepresented among the most frequently collected species (11 of the top 20 collected species were ferns). Within each year, 90% of specimens were collected during May–September but only 46% of species were collected only during these five months. Species that have been collected outside of the peak collection window (i.e., with at least one collection during October–April) are far more likely to be overrepresented in herbaria compared with species that have not been collected outside of peak collection months (Fig. S2). These non-peak species include all but 18 of the 1000 most commonly collected species in the Northeast; 11 of these 18 are species of *Carex*. Despite also being collected in off-peak months, the top species have been preferentially collected throughout the year, including during peak months; 96% of the top 1000 most collected species remain in the top 1000 when only collections from peak months are considered.

Some species are overrepresented in collections because they were collected by many people (e.g., *Arisaema triphyllum* (L.) Schott, *Onoclea sensibilis* L., and *Polystichum acrostichoides* (Michx.) Schott; see Table S5 for information about the number of people who collected each taxa mentioned in this section), whereas others are overrepresented because they were collected intensively by a few people (e.g., *Sceptridium dissectum* (Spreng.) Lyon, *Scirpus cyperinus* (L.) Kunth, and *Viola sororia* Willd.).

Some species were collected by far more people than expected from our GAM model (e.g., *Cypripedium acaule* Aiton and *Solanum dulcamara* L.) whereas *Dichanthelium acuminatum* (Sw.) Gould & C.A.Clark was collected by far fewer people than expected. Similarly, some genera were collected by more people than expected from our GAM model (e.g., *Lobelia, Lysimachia*, and *Trifolium*), whereas others by fewer people than expected (e.g., *Crataegus, Dichanthelium, Potamogeton, Salix, Sphagnum*). Some families were also collected by more people than expected from our model (e.g., Apocynaceae, Asteraceae, Ericaceae, Fabaceae, and Orchidaceae) and others by fewer than expected (e.g., Cyperaceae, Poaceae, Juncaceae, Salicaceae, and Violaceae). Commonly favored families—collected by more people than expected—typically had peaks in annual collections in the 1910s and 1930s, mirroring overall trends in collections through time (Fig. 7). Commonly avoided families—collected by fewer people than expected—typically had only a single peak during the 1910s. Some commonly avoided families (e.g., Potamogetonaceae and Sphagnaceae), had relatively low collections through time and its peaks correspond to specialist collectors rather than overall trends in collections.

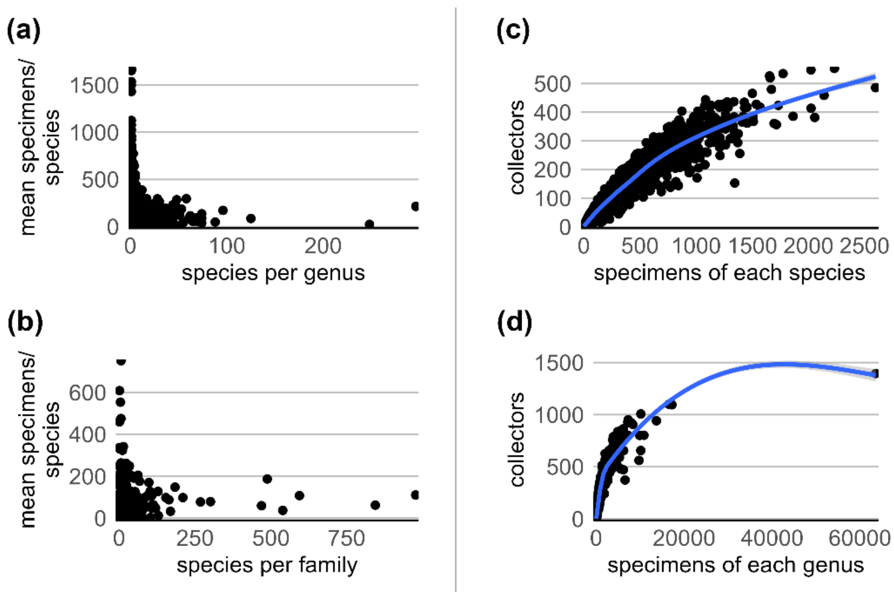

Figure 6. The plots show the collection depth (average number of specimens per species) for each (a) genus and (b) family. The scatter plots in the right pane (panels c & d) show the relationship between the number of specimens per species and the number of collectors who collected these species of each (c) species and (d) genus.

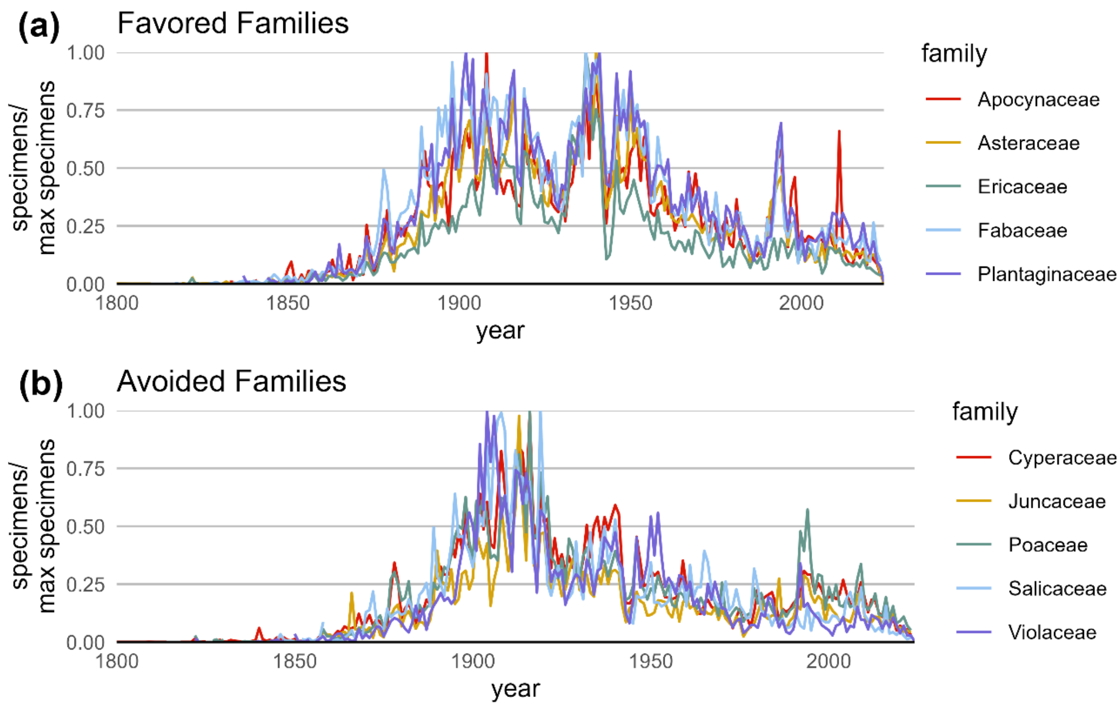

Figure 7. The annual variation in collection intensity for a subset of families collected by (a) more people than expected (favored families) and (b) less people than expected (avoided families). The vertical axes are adjusted to show variation in collection intensity for each family on the same scale where 1 represents the maximum number of specimens collected in a given year for each family.

Summary of Results

We identified nearly 10,000 collectors who have made important contributions to our understanding of plant biodiversity in the northeastern United States. We confirmed that a few mega-collectors contributed a disproportionately large share of these collections. Our analysis reveals many novel ways in which the collection efforts by thousands of less-prolific collectors have greatly expanded the temporal, spatial, and taxonomic dimensions of NHCs.

We assert that overall bias in collections across space, time, and taxa, is strongly impacted by predictable collection norms that are the result of the shared collector practices of many collectors rather than by stochastic biases of individual collectors (Fig. 8). The predictability of these biases provides an opportunity to address them more thoughtfully in biodiversity models that depend on these data. Specifically, we identified five collection norms common to the practices of all collectors: they tend to collect a.) more species rather than multiple specimens of the same species; b.) about 10 specimens per locality during their lifetime; c.) from localities sampled by other collectors; d.) during the peak growing season in spring and summer when climates are more favorable and photosynthetic rates and reproduction are generally higher; e.) species from smaller genera and families; and f.) particular species that are available outside of peak collecting months (i.e., when climates are less favorable for plant growth. We also identified that some collections norms have changed through time with collectors avoiding several taxonomically complex taxa during some decades.

In contrast to the collections norms detailed above, we also identified several divergences between the collector practices of more- versus less-prolific collectors. Specifically, more-prolific collectors i.) collected largely during fewer months; ii.) had stronger affinities to certain localities; and iii.) were not active in several large regions sampled by less-prolific collectors (e.g., the state of New York, USA).

A summary of our findings is presented in Fig. 8, where we outline the collector practices and resulting collection biases we have identified in the context of three key dimensions of bias: taxon, time, and space. We include two previously identified temporal collection biases, the decline in overall collections that was first presented by Prather *et al.* (2004) and the decline in collections during World Wars I and II identified by Daru *et al.* (2018). We also include the overall spatial clustering of collections, which was first defined by Myers *et al.* (2016).

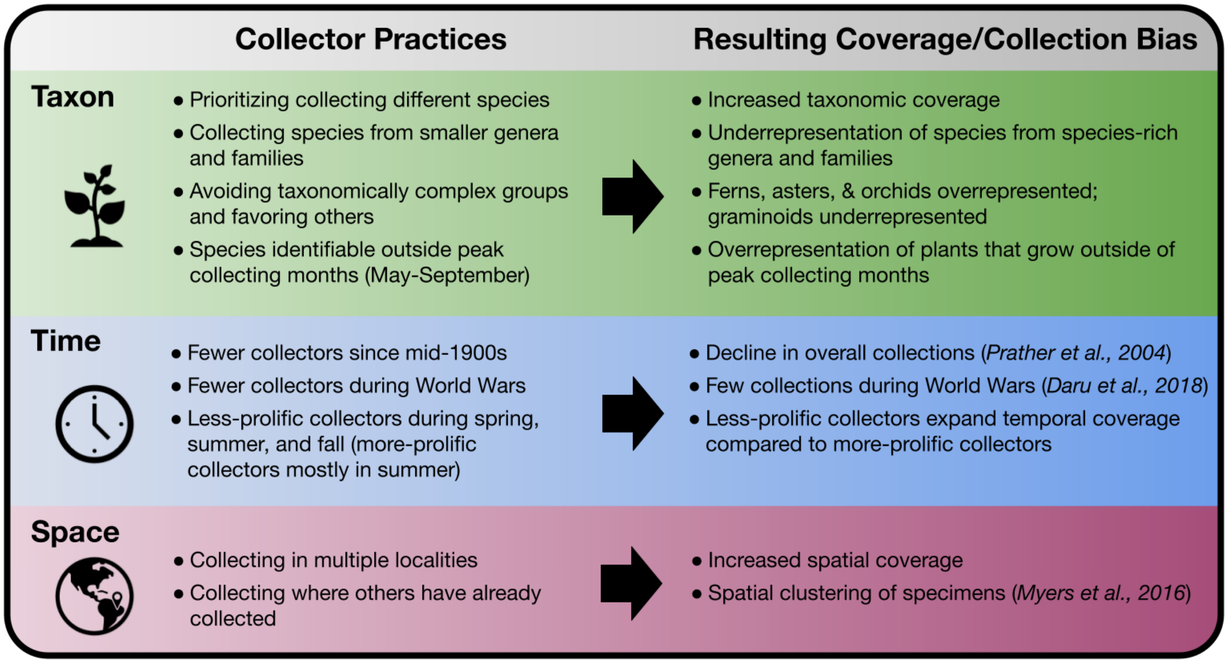

Figure 8. This graphic describes the collector practices that have shaped overall the overall coverage and collection bias in natural history collections along three dimensions: taxon, time, and space.

Discussion

Taxonomic bias: prioritizing greater species diversity and the underrepresentation of large, complex taxa

We found that botanists in the northeastern US prioritized collecting more species versus collecting multiple specimens of the same species. Although this tendency has been viewed as problematic in biology (Lewin, 1982; May, 2004), we assert that such collecting has contributed considerably to expanding taxonomic coverage represented in NHCs and improving our understanding of species diversity and distributions (Alba *et al.*, 2021). Despite this tendency, however, collectors do not sample species randomly: many collect the same taxa while avoiding others (Fig. 6). For instance, the brightly colored pink lady's slipper orchid (*Cypripedium acaule* Aiton) was collected by many people (444 people) whereas hairy panicgrasses (*Dichanthelium acuminatum* (Sw.) Gould & C.A.Clark) was collected by relatively few (153 people). This collection norm affects our attempts to model biodiversity owing to the gap between taxon diversity and abundance information recorded in NHCs versus their actual diversities and abundances in nature (Elith & Leathwick, 2007; Gomes *et al.*, 2018). This pattern mirrors the collection norm whereby collectors tend to collect ten specimens per locality and suggests that collectors travelled to different localities to collect new species rather than comprehensively collecting at a single locality.

Taxonomic collection norms have likely contributed to the overrepresentation of less species-rich taxa with distinctive morphologies (e.g., *Lobelia, Polystichum,* and *Dryopteris*) in herbaria relative to larger taxa that are often taxonomically challenging (e.g., *Carex, Crataegus,* and *Salix*). Specimens from many large taxa were collected by fewer people than expected, suggesting these were mainly collected by botanists with specialized taxonomic interests. In the northeastern US, such specialist-prone taxa include genera like *Sphagnum* (peat mosses), *Dichanthelium* (rosette grasses), *Salix* (willows), and *Crataegus* (hawthorns), and families like Poaceae, Cyperaceae, and Juncaceae (collectively, the graminoids). These groups often require microscopic examination to distinguish subtle differences necessary for accurate species identification and often can only be identified with reproductive features at specific maturation stages (FNA Editorial Committee, eds., 1993+). Further complicating species identification and delimitation are their complex evolutionary histories, including infrageneric hybridization (Ennos *et al.*, 2005). We hypothesize that this taxonomic bias in collections is often driven by the perceived taxonomic complexity and difficulty to identify species within such groups (for discussions of taxonomic complexity, see Ennos *et al.*, 2005; Karbstein *et al.*, 2024). This collection norm suggests that the most diverse groups, which are likely in greatest need of study, are woefully underrepresented in NHCs.

We also identified clear trends in shifting taxonomic collection norms through time, a pattern that has received little attention. We observed that taxonomic biases have apparently shifted, with certain taxa being favored and others apparently avoided across different generations of botanists. For example, in the northeastern US, many collectors in the 1930s avoided families like Poaceae, Cyperaceae, Juncaceae, and Sphagnaceae. We hypothesize that collectors from the Citizens Conservation Corps, many of whom lacked formal botanical training, may have avoided families they perceived as more complex. In other words, we hypothesize that collectors are less prone to collect what they don’t know. This has significant implications for comparing temporal trends between taxa; variations in historical collection intensity may affect apparent changes in characteristics such as species distribution modeling (Franklin & Miller, 2009) and phenology (Miller-Rushing *et al.*, 2008). Therefore, understanding the overall temporal distribution of collections is crucial for appreciating how record availability—and the uncertainty in these data—changes over time.

Spatial bias: less-prolific collectors contribute unique spatial coverage with more-random spatial sampling

We identified an important divergent collection practice between more- and less-prolific collectors whereby less-prolific collectors contribute unique spatial coverage versus collections by more-prolific collectors (see Fig. 4). These less-prolific collectors enhance sampling near commonly collected localities (Fig. 4) and act as the backbone for entire regions where more-prolific collectors have not collected (Fig. 5b). For example, less-prolific collectors greatly improve spatial coverage in large portions of New York State, western Massachusetts, and near major universities (e.g., Rutgers University and Cornell University), likely highlighting the impact of student collectors on overall spatial sampling in the Northeast (see Fig. 5b). Thus, the cumulative spatial coverage by more-prolific collectors is considerably lower than that of less-prolific collectors, indicating that the collections made by the latter more accurately reflect plant diversity across different regions. It is important to note, however, that although there has been extensive herbarium digitization in the northeastern US, digitization is still ongoing and some of the spatial patterns apparent in currently available GBIF data will inevitably change as more data becomes available. In particular, several large collections in the region (e.g., the New York State Museum and the Pennsylvania State University Herbarium) do not publish their specimen data to major biodiversity aggregators like GBIF, potentially contributing to the comparably low specimen density in areas like upstate New York (Fig. 5a). Continued efforts to digitize 'silent' herbaria (Zhigila, Schmidt *et al.*, 2025) and make their data digitally accessible are necessary for understanding how data from NHCs can be leveraged for studying global biodiversity. Interestingly, the spatial bias of less-prolific collectors does not differ significantly from the overall spatial bias in herbaria. However, these collections are still biased with respect to random sampling. This suggests that while less-prolific collectors do not exhibit the same preference for specific collection sites as more-prolific collectors, they also tend to revisit locations where collections have previously been made. Despite this spatial collection norm, the increased spatial coverage provided by less-prolific collectors has greatly improved the overall spatial sampling in herbaria. This increased spatial coverage has helped facilitate the recent application of herbarium data to disciplines that rely on extensive sampling; for example, ecology (Meineke *et al.*, 2019a; Heberling, 2022); invasion biology (Crawford & Hoagland 2009; Schmidt *et al.,* 2023), species distribution modeling (Daru *et al.*, 2021), environmental science (Carbone *et al.*, 2023; Jakovljević *et al.*, 2024), and conservation biology (Schatz, 2002).

Finally, the broad spatial sampling by numerous less-prolific collectors that we identified reflects patterns also observed with contemporary iNaturalist data, where contributions by millions of community scientists greatly extend spatial sampling beyond what is captured in herbaria (Eckert *et al.*, 2024; also see Daru & Rodriguez, 2023). This similarity indicates that the spatial biases of community scientists align more closely with those of less-prolific collectors than with the more-prolific collectors who contributed heavily to overall spatial biases in collections. Similarly, several studies have demonstrated that small, regional collections provide unique temporal and spatial coverage not represented in larger collections (Glon *et al.*, 2017; Monfils *et al.*, 2020; Marsico *et al.*, 2020). We hypothesize that the expanded coverage of smaller herbaria is driven by the efforts of less-prolific collectors who also provide unique temporal and spatial coverage not captured by more-prolific collectors.

Temporal bias: variability driven by collector activity

The substantial declines in collections over the past 75 years is consistent with trends observed in other regional floras (Prather *et al.*, 2004; Daru *et al.*, 2018) and is strongly correlated with declines in the number of active collectors. This suggests that while more-prolific collectors may heavily influence the interannual intensity of collections at certain times (Bebber *et al.*, 2012; Daru *et al.*, 2018), the overall trends are primarily driven by fluctuations in the number of all active collectors.

Notably, the reduction in annual collections coincided with the years when the US was involved in World Wars I (1917–1920) and II (1941–1946). During the two world wars, millions of men were conscripted for military service and at the same time millions of women, students, and older Americans entered the workforce (Witt, 1942; Wilcock, 1957) and would have been largely unable to collect plants.. Following decreased collections during World War I, the spike in collections and active collectors from 1932 through 1941 corresponds with US government efforts to reduce unemployment and support environmental projects during the Great Depression (1929–1939; Salmond, 1967). During this period, the government employed thousands of citizens—primarily young men aged 18 to 25—for projects focusing on environmental improvements (e.g., in the Civilian Conservation Corps; Salmond, 1967). A key objective of these initiatives was to produce local species inventories, documented through "complete herbaria," to aid in land planning and protection (Department of the Interior, 1936). Since these projects often targeted similar habitats—primarily forested areas—many inventories likely covered areas with similar species composition in the northeastern US. Consequently, despite the spikes in collections, active collectors, and collection locations during this time, the number of species collected during this period did not increase substantially. Once World War II began and people from the same demographic were heavily drafted into WWII, all metrics once again quickly declined. This highlights how major socio-political events affecting significant population segments can directly impact NHCs by reducing the pool of available collectors. Similar impacts of socio-political events on NHCs were recently documented in collection requests for multiomic sampling, which plummeted during the global COVID pandemic (Davis *et al.*, 2024).

We identified that less-prolific collectors increased overall sampling at the start and end of the primary growing season (late spring and early autumn), which diverges from collections by more-prolific collectors whose activity during these periods markedly decreases. The intensity of sampling during these off-peak periods is crucial for improving the accuracy of phenological estimates (Miller-Rushing *et al.*, 2008) and understanding the impact of anthropogenic climate change on early- and late-season species (Kudo & Ida, 2013; Park *et al.*, 2023). We hypothesize that the increased sampling by less-prolific collectors at the beginning and end of the growing season (i.e., April–May and Septemer–October) might be related to student collections in university botany classes during the academic year (typically September–May).

Surprisingly, although 90% of specimens are collected in the northeastern US between May and September, species collected outside the peak months are disproportionately represented among the most abundant species in herbaria. These include many evergreen (e.g., *Polystichum acrostichoides* (Michx.) Schott and *Dryopteris marginalis* (L.) A.Gray), woody (e.g., *Vaccinium corymbosum* L. and *Acer rubrum* L.), and early-flowering species (e.g., *Viola sororia* Willd. and *Arisaema triphyllum* (L.) Schott), as well as species with winter-available flowers or fruits (e.g., *Ilex verticillata* (L.) A.Gray and *Hamamelis virginiana* L.). We hypothesize this overrepresentation is driven by collectors' familiarity with these species, which are more accessible and—in some cases—more identifiable outside of peak collection months when fewer species are available.

Exceptions to the norms: unique collector practices contribute overall bias

Despite the similar collector practices we identified, we emphasize that understanding how some collectors diverged from these norms is important for understanding overall collection bias in NHCs. For example, the most prolific collector in our dataset, R. L. Schaeffer, Jr., collected

50,287 specimens from only 195 localities—far fewer than expected based on our model. He collected, almost exclusively, in the vicinity of Allentown, PA where Schaeffer taught botany at Muhlenberg College from 1954-1983 ('R. L. Schaeffer Obituary', 2001). His singular efforts had an outsized impact on overall spatial bias in the northeastern US with his collections being the main driver of the high collection density in eastern PA, one of the most collection-dense areas in the northeastern US. Furthermore, the expansive taxonomic coverage and high collection depth of Schaeffer's specimens provides a rich documentation of the flora of eastern Pennsylvania over nearly a half century that can be leveraged for a diversity of collections-based investigations (e.g., Meineke *et al.*, 2019b). This highlights how integrating historical information about collectors (especially mega-collectors like Schaeffer) can help explain the more stochastic processes in biodiversity data and can illuminate important datasets better characterizing species and ecosystem responses to anthropogenic pressures.

Conclusion

Our findings reveal how our understanding of biodiversity is founded on the cumulative effort of thousands of people, many of whom have made small but impactful contributions to natural history collections (NHCs). The cumulative spatial, temporal, and taxonomic practices of all collectors give rise to the overall biases in collections. It is crucial that we identify and categorize these collector practices to better understand the drivers of overall collection bias in NHCs and begin developing tools to address them. We have identified numerous predictable collection norms that appear to have shaped overall bias in NHCs. The predictability of these biases provides an exciting and promising opportunity to begin incorporating statistical tools to address collection biases in biodiversity models. These results can also be leveraged to guide future collection efforts that can minimize gaps in collections and reduce bias in NHCs moving forward. We highlight that collector practices—even by those who collected only a small number of specimens—have vastly expanded the coverage of NHCs and we assert that continued collections of all sizes are crucial for continuing to expand the coverage of NHCs and further increasing their utility for understanding biodiversity in the face of global change.

Acknowledgements

The authors thank Aaron Ellison for guidance on statistical analysis and comments on an earlier draft of this manuscript. We also thank Jonnathan Kennedy for providing updates to the Harvard Index of Botanists, Nawal Shrestha for support with coding and analysis, and the rest of the Davis lab for input throughout the project. RJS and KES acknowledge support from the National

Science Foundation (Graduate Research Fellowship Program) and the Harvard University Herbaria. CCD acknowledges funding from Harvard University and from National Science Foundation funding grants: DEB 1754584, EF1208835, DEB 2101884, DEB 1802209, and MRA 2105903. LS acknowledges support from USDA Hatch project 1026539, NSF-DEB 1601101, and School of Environmental and Biological Sciences to Chrysler Herbarium at Rutgers University.

Author Contribution

RJS, CCD, and LS conceptualized the study. RJS and CCD developed the methodology, RJS and KES led the data curation, and RJS completed the investigations and formal analysis. RJS led data visualization with support from CCD, LS, and KES. RJS and CCD led writing with input and support from LS and KES.

Data Availability Statement

The data generated during this study are available in the supporting information of this manuscript. Tables S2, S3, and all code created for this study are available on the Harvard Dataverse (https://doi.org/10.7910/DVN/OJCODH).

Conflict of Interest Statement

CCD declares that he is supported by LVMH Research and Dior Science, a company involved in the research and development of cosmetic products based on floral extracts. He also serves as a member of Dior's Age Reverse Board.

References

Alba C, Levy R, Hufft R. **2021**. Combining botanical collections and ecological data to better describe plant community diversity (J-Z Wan, Ed.). *PLOS ONE* **16**: e0244982.

Bebber DP, Carine MA, Davidse G, Harris DJ, Haston EM, Penn MG, Cafferty S, Wood JRI, Scotland RW. **2012**. Big hitting collectors make massive and disproportionate contribution to the discovery of plant species. *Proceedings of the Royal Society B: Biological Sciences* **279**: 2269–2274.

Carbone MS, Ayers TJ, Ebert CH, Munson SM, Schuur EAG, Richardson AD. **2023**. Atmospheric Radiocarbon for the Period 1910-2021 Recorded by Annual Plants. *Radiocarbon* **65**: 357–374.

CBD. **2022**. *Decision adopted by the conference of the parties to the convention on biological diversity 15/4. Kunming-montreal global biodiversity framework*. Montreal, Canada.

Crawford PHC, Hoagland BW. **2009**. Can herbarium records be used to map alien species invasion and native species expansion over the past 100 years? *Journal of Biogeography* **36**: 651–661.

CT DEEP. **2023**. Northeastern States Town Boundary Set. *Connecticut Department of Energy & Environmental Protection*.

Daru BH, Davies TJ, Willis CG, Meineke EK, Ronk A, Zobel M, Pärtel M, Antonelli A, Davis CC. **2021**. Widespread homogenization of plant communities in the Anthropocene. *Nature Communications* **12**: 6983.

Daru BH, Park DS, Primack RB, Willis CG, Barrington DS, Whitfeld TJS, Seidler TG, Sweeney PW, Foster DR, Ellison AM, *et al.* 2018. Widespread sampling biases in herbaria revealed from large-scale digitization. *New Phytologist* **217**: 939–955.

Daru BH, Rodriguez J. **2023**. Mass production of unvouchered records fails to represent global biodiversity patterns. *Nature Ecology & Evolution* **7**: 816–831.

Davis CC. **2023**. The herbarium of the future. *Trends in Ecology & Evolution* **38**: 412–423.

Davis CC. **2024**. Collections are truly priceless. *Science* **383**: 1035–1035.

Davis CC, Sessa E, Paton A, Antonelli A, Teisher JK. **2024**. Guidelines for the effective and ethical sampling of herbaria. *Nature Ecology & Evolution*.

Department of the Interior. **1936**. Annual Report of the Department of the Interior 1936.

Eckert I, Bruneau A, Metsger DA, Joly S, Dickinson TA, Pollock LJ. **2024**. Herbarium collections remain essential in the age of community science. *Nature Communications* **15**: 7586.

Elith J, Leathwick J. **2007**. Predicting species distributions from museum and herbarium records using multiresponse models fitted with multivariate adaptive regression splines. *Diversity and Distributions* **13**: 265–275.

Ennos R, French G, Hollingsworth P. **2005**. Conserving taxonomic complexity. *Trends in Ecology & Evolution* **20**: 164–168.

Forest Inventory and Analysis. **2023**. *U.S. Department of Agriculture, Forest Service, Research & Development*.

Flora of North America Editorial Committee, eds. 1993. Flora of North America (FNA). *Flora of North America*.

Franklin J, Miller JA. **2009**. *Mapping species distributions: spatial inference and prediction*. Cambridge ; New York: Cambridge University Press.

Funk VA. **2003**. 100 Uses for an Herbarium: well at least 72. *American Society of Plant Taxonomists Newsletter* **17**: 17–19.

GBIF.org. 2024. GBIF Occurrence Download. Available at: https://doi.org/10.15468/dl.rndw9f. (Accessed: 28 August 2024).

Glon HE, Heumann BW, Carter JR, Bartek JM, Monfils AK. **2017**. The contribution of small collections to species distribution modelling: A case study from Fuireneae (Cyperaceae). *Ecological Informatics* **42**: 67–78.

Gomes VHF, IJff SD, Raes N, Amaral IL, Salomão RP, De Souza Coelho L, De Almeida Matos FD, Castilho CV, De Andrade Lima Filho D, López DC, *et al.* 2018. Species Distribution Modelling: Contrasting presence-only models with plot abundance data. *Scientific Reports* **8**: 1003.

Groom Q, Bräuchler C, Cubey R, Dillen M, Huybrechts P, Kearney N, Klazenga N, Leachman S, Paul DL, Rogers H, *et al.* 2022. The disambiguation of people names in biological collections. *Biodiversity Data Journal* **10**: e86089.

Harvard University Herbaria. **2024**. Harvard Index of Botanists. *Harvard University Herbaria & Libraries*.

Heberling JM. **2022**. Herbaria as Big Data Sources of Plant Traits. *International Journal of Plant Sciences* **183**: 87–118.

Hedrick BP, Heberling JM, Meineke EK, Turner KG, Grassa CJ, Park DS, Kennedy J, Clarke JA, Cook JA, Blackburn DC, *et al.* 2020. Digitization and the Future of Natural History Collections. *BioScience* **70**: 243–251.

Jakovljević K, Mišljenović T, Van Der Ent A, Baker AJM, Invernón VR, Echevarria G. **2024**. “Mining” the herbarium for hyperaccumulators: Discoveries of nickel and zinc (hyper)accumulation in the genus *NOCCAEA* (Brassicaceae) through X-ray fluorescence herbarium scanning. *Ecological Research* **39**: 450–459.

Johnson KR, Owens IFP, the Global Collection Group. **2023**. A global approach for natural history museum collections. *Science* **379**: 1192–1194.

Karbstein K, Kösters L, Hodač L, Hofmann M, Hörandl E, Tomasello S, Wagner ND, Emerson BC, Albach DC, Scheu S, *et al.* 2024. Species delimitation 4.0: integrative taxonomy meets artificial intelligence. *Trends in Ecology & Evolution* **39**: 771–784.

Kozlov MV, Sokolova IV, Zverev V, Zvereva EL. **2021**. Changes in plant collection practices from the 16th to 21st centuries: implications for the use of herbarium specimens in global change research. *Annals of Botany* **127**: 865–873.

Kudo G, Ida TY. **2013**. Early onset of spring increases the phenological mismatch between plants and pollinators. *Ecology* **94**: 2311–2320.

Lendemer J, Thiers B, Monfils AK, Zaspel J, Ellwood ER, Bentley A, LeVan K, Bates J, Jennings D, Contreras D, *et al.* 2020. The Extended Specimen Network: A Strategy to Enhance US Biodiversity Collections, Promote Research and Education. *BioScience* **70**: 23–30.

Lewin R. **1982**. Biology Is Not Postage Stamp Collecting: Ernst Mayr, the eminent Harvard evolutionist, explains why he thinks some physical scientists have a problem with evolution. *Science* **216**: 718–720.

Mancini M, Barber A, Block TA, Skema C. **2019**. Mid-Atlantic megalopolis georeferencing guidelines.

Marín-Rodulfo M, Rondinel-Mendoza KV, Martín-Girela I, Cañadas EM, Lorite J. **2024**. Old meets new: Innovative and evolving uses of herbaria over time as revealed by a literature review. *PLANTS, PEOPLE, PLANET* **6**: 1261–1271.

Marsico TD, Krimmel ER, Carter JR, Gillespie EL, Lowe PD, McCauley R, Morris AB, Nelson G, Smith M, Soteropoulos DL, *et al.* 2020. Small herbaria contribute unique biogeographic records to county, locality, and temporal scales. *American Journal of Botany* **107**: 1577–1587.

May RM. **2004**. Tomorrow's taxonomy: collecting new species in the field will remain the rate–limiting step (HCJ Godfray and S Knapp, Eds.). *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* **359**: 733–734.

Meineke EK, Classen AT, Sanders NJ, Jonathan Davies T. **2019a**. Herbarium specimens reveal increasing herbivory over the past century (A Iler, Ed.). *Journal of Ecology* **107**: 105–117.

Meineke EK, Davies TJ, Daru BH, Davis CC. **2019b**. Biological collections for understanding biodiversity in the Anthropocene. *Philosophical Transactions of the Royal Society B: Biological Sciences* **374**: 20170386.

Miller-Rushing AJ, Inouye DW, Primack RB. **2008**. How well do first flowering dates measure plant responses to climate change? The effects of population size and sampling frequency. *Journal of Ecology* **96**: 1289–1296.

Monfils AK, Krimmel ER, Bates JM, Bauer JE, Belitz MW, Cahill BC, Caywood AM, Cobb NS, Colby JB, Ellis SA, *et al.* 2020. Regional Collections Are an Essential Component of Biodiversity Research Infrastructure. *BioScience* **70**: 1045–1047.

Myers N, Mittermeier RA, Mittermeier CG, Da Fonseca GAB, Kent J. 2000. Biodiversity hotspots for conservation priorities. *Nature* 403: 853–858.

Park DS, Davis CC. **2017**. Implications and alternatives of assigning climate data to geographical centroids. *Journal of Biogeography* **44**: 2188–2198.

Park DS, Xie Y, Ellison AM, Lyra GM, Davis CC. **2023**. Complex climate-mediated effects of urbanization on plant reproductive phenology and frost risk. *New Phytologist* **239**: 2153–2165.

Pebesma E. **2018**. Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal* **10**: 439.

Pebesma E, Bivand R. **2023**. *Spatial Data Science: With Applications in R*. New York: Chapman and Hall/CRC.

PennDOT. **2024**. PennDOT Open Data. *Pennsylvania Department of Transportation*.

Perring FH, Walters SM (Eds.). **1962**. *Atlas of the British flora*. London: Thomas Nelson & Sons.

Prather LA, Alvarez-Fuentes O, Mayfield MH, Ferguson CJ. **2004**. The Decline of Plant Collecting in the United States: A Threat to the Infrastructure of Biodiversity Studies. *Systematic Botany* **29**: 15–28.

Preston CD. **2013**. Following the BSBI's lead: the influence of the *Atlas of the British flora* , 1962–2012. *New Journal of Botany* **3**: 2–14.

Obituary for Robert L. Schaeffer (Aged 83). **2001**. *The Morning Call*: 28.

Rudis VA. **2003**. *Comprehensive Regional Resource Assessments and Multipurpose Uses of Forest Inventory and Analysis Data, 1976 to 2001: A Review*. Asheville, North Carolina: USDA Forest Service, Southern Research Station.

Salmond JA. **1967**. *The Civilian Conservation Corps, 1933-1942; a New Deal case study*. Durham, North Carolina: Duke University Press.

Schatz GE. **2002**. Taxonomy and Herbaria in Service of Plant Conservation: Lessons from Madagascar's Endemic Families. *Annals of the Missouri Botanical Garden* **89**: 145.

Schmidt RJ, King MR, Aronson MFJ, Struwe L. **2023**. Hidden cargo: The impact of historical shipping trade on the recent-past and contemporary non-native flora of northeastern United States. *American Journal of Botany* **110**: e16224.

Schorn C, Weber E, Bernardos R, Hopkins C, Davis C. **2016**. The New England Vascular Plants Project: 295,000 specimens and counting. *Rhodora* **118**: 324–325.

Shorthouse DP. **2024**. Bionomia. *Bionomia*.

Silge J, Robinson D. **2016**. tidytext: Text Mining and Analysis Using Tidy Data Principles in R. *The Journal of Open Source Software* **1**: 37.

Sweeney PW, Starly B, Morris PJ, Xu Y, Jones A, Radhakrishnan S, Grassa CJ, Davis CC. **2018**. Large–scale digitization of herbarium specimens: Development and usage of an automated, high–throughput conveyor system. *TAXON* **67**: 165–178.

United States Census Bureau. 2024. Cartographic Boundary Files: States: 1 : 500,000 (national). Available at: https://www.census.gov/geographies/mapping-files/time-series/geo/cartographic-boundary.html. (Accessed: 6 September 2024).

Webster MS (Ed.). **2017**. *The extended specimen: emerging frontiers in collections-based Ornithological Research*. Boca Raton London New York: CRC Press, Taylor & Francis Group.

Weeks A, Collins E, Majors T, Murrell Z, Paul D, Sheik M, Shorthouse D, Zeringue-Krosnick S. **2024**. Workshop Report: Supporting inclusive and sustainable collections-based research infrastructure for systematics (SISRIS). *Research Ideas and Outcomes* **10**: e126532.

Whelan A. **1948**. Of People and Places. *The Bridgeport Sunday Post*: 23.

Wilcock RC. **1957**. The Secondary Labor Force and the Measurement of Unemployment. In: The Measurement and Behavior of Unemployment. National Bureau for Economic Research, 167–210.

Witt B. **1942**. Labor in Transition to a War Economy. In: Hanna HS, ed. Monthly Labor Review. Bureau of Labor Statistics.

Zhigila D, Schmidt RJ, Thiers B, Abdul S, Abdullahi S, AbdulRahaman A, Aigbokhan E, Ajibade G, Ajikah L, Akomaye F, *et al.* 2025. Biodiversity science is improved when silent herbaria speak.

Supporting Information

Table S1 Herbaria whose specimens were used for this study, indicating the institution code, institution name, and the number of specimens from each herbarium that were used in this study.

Table S2 Total specimens used in this study after data cleaning, georeferencing, and collector disambiguation.

Table S3 A table containing the DarwinCore recordedBy strings from gbif, the unique identifier representing each collector, and the number of specimens, species, and localities in which each person collected plants.

Table S4 The one hundred most frequently collected species in the northeastern US.

Table S5 The number of specimens, species, and collectors that collected taxa mentioned in the text of the manuscript.

Fig. S1 A flowchart showing the data cleaning process including the number of specimens removed at each step.

Fig. S2 A boxplot showing the difference in number of specimens of each species related to whether the species has been collected only during peak collection months (May, June, July, August, and September) or also collected in non-peak months.