Examining Collection Biases Across Different Taxonomic Groups: Understanding How Biases Can Compare Across Herbarium Datasets

Jordan Williams & Katelin D. Pearson*

Department of Biological Science, Florida State University, FL

https://doi.org/10.33697/ajur.2019.005

*Student:jaw16e@my.fsu.edu**
Mentor: kpearson@bio.fsu.edu

ABSTRACT
Specimen-based data are an invaluable resource for an increasing diversity of scientific fields, including global change biology, ecology, evolution, and genetics; however, certain analyses of these data may be limited by the non-random nature of collecting activity. Geographic, temporal, and trait-based collecting biases may consequently affect the understanding of species' distributions, obviating the need to determine what biases exist and how they may impact further analyses. Trait-based biases were examined in herbarium specimen records of two abundant and diverse families (Asteraceae and Fabaceae) in a well-collected and digitized region (California) by comparing geographic-bias-adjusted simulations of random collections to actual collecting patterns. Collecting biases were fairly similar between families for a number of traits, such as a strong bias against collecting introduced species, while seasonal collecting biases showed a peak in activity in the Spring for both families. However, while there was only a dip in the fall for Asteraceae, Fabaceae were seriously under-collected for the majority of the year. These results demonstrate that significant collecting biases exist and may differ depending on the dataset, highlighting the importance of understanding the dataset and potentially accounting for its sampling limitations.

KEYWORDS
Biodiversity; Natural History Collection; Sampling Bias; Biodiversity Specimens; iDigBio; Botanical Databases; Plant Traits

INTRODUCTION

Herbaria, collections of dried and pressed plant specimens, are excellent sources of botanical, ecological, environmental, and other data, helping us understand the changes species, and even ecosystems, have experienced over time. Herbarium specimens have enabled studies of species distributions, pollution, climate change, and even disease spread.[1, 2] The digitization of herbarium specimens has provided researchers with large datasets that are easily accessible; the median number of specimens being used in papers is notably higher when scientists use digitized databases.[1] However, despite all the benefits they provide, herbarium specimen data may contain biases, since collectors rarely collect specimens in a truly random fashion, resulting in over- and under-collecting of certain types of species, in certain locations, or at certain times. Collectors were historically focused on the discovery of new species and their distribution in particular areas, and this non-random collection pattern has thus become problematic only recently. Indeed, only after researchers began to use these collections for studies other than taxonomy or distribution, did legitimate concerns about several types of collection biases start to arise since, as mentioned above, they can drastically affect the results of any studies based on herbarium specimens.[1-3]

For instance, there are geographical biases;[3-11] taxonomic biases that occur when certain types of plants are collected over others (*e.g.*, for a specific research project);[12, 13, 14] temporal biases[9, 10, 13, 15-19] phenological biases that occur based on when the plants flower;[20] and biases based on individual morphology of the plant such as a particularly tall or oddly-shaped specimen. As these biases have come to light and the purpose of herbaria evolved, increasing attention is being paid to how these biases occur and possible methods to correct or account for them.[3, 21-23] As we try to integrate these old collections into current studies, aside from taxonomy and distribution, it is becoming more clear that in order to achieve accurate results, we need to rethink the way we approach specimen collection. Historically, a sterile specimen without flowers or fruits would have not been considered worthwhile to collect because they could be nearly impossible to identify. So while herbaria are excellent reflections on reproductive phenology, there is little reference for vegetative traits.

Many of the studies listed above have identified biases in specimen-based datasets with limited taxonomic scopes (*i.e.*, using few species), while even fewer identify how these biases compare between different datasets.[21] In this study, trait-based biases were compared in datasets of two different taxonomic groups, the sunflower family (Asteraceae) and bean family (Fabaceae), in the U.S. state of California. These plant families are widespread in the state (There are around 1,400 known

Asteraceae species and 700 Fabaceae species present), and their morphological, geographical, and taxonomic diversity allow us to test for multiple trait-based biases. Furthermore, active herbarium specimen digitization in California has made large amounts of previously hard to access specimen data available, enabling more comprehensive analyses of biases on a larger scale. Understanding the similarities and differences in collecting biases shared among different institutions may inform future analysis using these data.

METHODS AND PROCEDURES

Specimen data for California Asteraceae and Fabaceae was downloaded using the iDigBio portal (idigbio.org). Data cleaning consisted of removing any erroneous (*e.g.*, non-plant, outside California) specimens, standardization of taxonomic names using the Taxonomic Name Resolution Service, and removing duplicate collections (*i.e.*, specimens of the same species collected in the same county on the same day). Records not classified to the species level and records of species with fewer than 50 specimens were excluded, as they were not considered to have enough specimen information to be accurately tested. The resulting dataset consisted of 151,035 specimens of 612 Asteraceae species and 78,744 specimens of 276 Fabaceae species.

Random datasets of Asteraceae and Fabaceae collections were simulated separately using a Monte Carlo approach similar to that of Schmidt-Lebuhn *et al.* 2013 that accounted for nonrandom spatial sampling (geographic biases). Briefly, simulated specimens of a given species from both families were "collected" at a frequency directly proportional to the level of collecting activity of actual specimens in the counties in which the species was collected in the actual dataset. Essentially, our simulated data sets imitated the levels of actual collection efforts, and we generated thousands of these simulations to mimic conditions if collection was truly a random effort. For each dataset (Asteraceae or Fabaceae), 10,000 simulated datasets were created. To compare the random simulations to the actual collections, the collection numbers from all simulations for a specific species were averaged together; this average was then compared to the observed number of specimens for the species in actual herbaria.

Data for growth habit (herb, shrub, or tree), generation time (annual or perennial), plant height, nativity (introduced or native), flowering period (greater than 3 months), and elevation (less than 100 meters) of each species were determined using the Flora of North America website (efloras.org), the USDA PLANTS database, Calflora.org, and other reputable sources. It should be noted that for growth habit tests, due to the relatively low number of tree species we examined in the Fabaceae family, we combined shrubs and trees into one group, woody plants. This group was then compared to Asteraceae shrub species. We also determined which species flower each month of the year for both families, and used this data to determine if there was a collection bias towards certain months. The number of all species collected in a particular month was averaged (total number of specimens divided by number of species). This means there is some overlap in the data as numerous plants flower longer than one month.

The percent difference was calculated between the average number of specimens generated by the simulation within the group with a given trait to the actual number of specimens with the trait in the actual dataset. A negative percent difference indicated that the bias exists against specimens with the trait, and a positive percent difference indicates that a bias exists towards specimens with the trait. The simulations for each species were used to create random expectations in the form of histograms. We then compared the simulated expectations to relative to the number of actual specimens in the database (see the dashed line in Figure 1).

RESULTS

As expected, the species tested exhibited a variety of traits. This made it possible to test for multiple biases and to compare biases between families.

Both native Asteraceae and Fabaceae species were over-collected relative to an expectation of random collection. Specimens of native Asteraceae were ten percent more abundant in the actual dataset than the simulated dataset (**Figure 1**), and specimens of native Fabaceae were nine percent more abundant. In contrast, introduced species were under-collected by 23% in Asteraceae and 75% in Fabaceae.

For the growth habits of specimens, forb/herb species in both families were over-collected (Asteraceae: five percent; Fabaceae: ten percent). In comparison, shrub specimens in Asteraceae were over-collected by 11%, while woody specimens in Fabaceae were under-collected by 32%.

The pereniality of species was examined, and while annuals were over-collected by nine percent in Asteraceae and under-collected by 14% in Fabaceae, perennials were over-collected in both Asteraceae and Fabaceae by ten percent and 14% respectively.

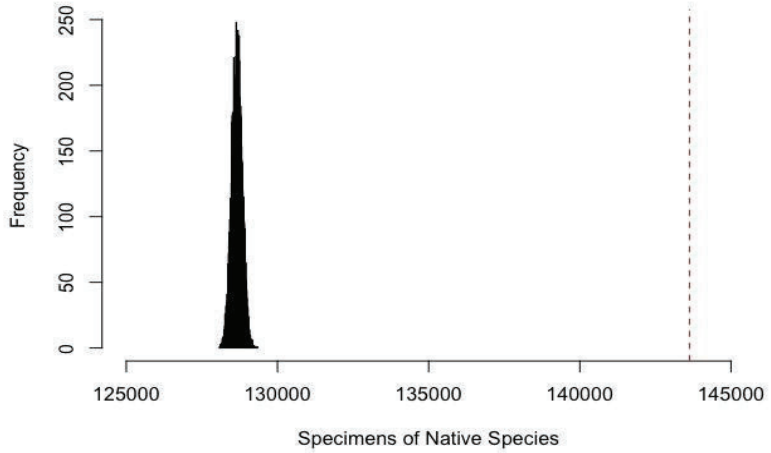

Figure 1. Histogram demonstrating the bias toward collecting native species of Asteraceae in California. The distribution on the left represents the number of simulated datasets with a given number of specimens with this trait, while the red dashed line on the right indicates the average of the actual collected specimens.

In addition, species with a maximum elevation of 100 m or less were also found to be over-collected in both families, with a percent difference of 22% in Asteraceae and 14% in Fabaceae.

In contrast, species with a flowering period of more than three months were considered, and these species were found to be over-collected by 14% in Asteraceae, but seriously under-collected by 30% in Fabaceae (**Table 1; Figure 2**).

Trait	Family	# of Species	Actual Value	Simulated Value	Percent Difference
Native	AST	547	143,630	128,662	11
	FAB	245	73,006	66,497	9
Introduced	AST	50	13,680	17,309	-23
	FAB	30	5,426	11,930	-75
Wooded	AST	89	22,592	20,221	11
	FAB	43	9,270	2,750	-32
Forb/Herb	AST	406	107,875	102,413	5
	FAB	233	69,474	65,996	10
Elevation <100 m	AST	285	92,099	73,615	22
	FAB	152	58,693	51,0304	14
Flowering >3 months	AST	292	85,577	74,158	14
	FAB	70	16,051	21,753	-30
Annual	AST	209	63,094	57,571	9
	FAB	82	24,699	28,356	-14
Perennial	AST	343	83,327	75,383	10
	FAB	170	49,212	42,827	14

Table 1. Results for test of collecting bias in specimens of Asteraceae and Fabaceae in California. Table shows trait tested, number of species for each trait, actual number of specimens, number of specimens expected if collecting effort lacked bias, and percent difference between actual and simulated number of specimens.

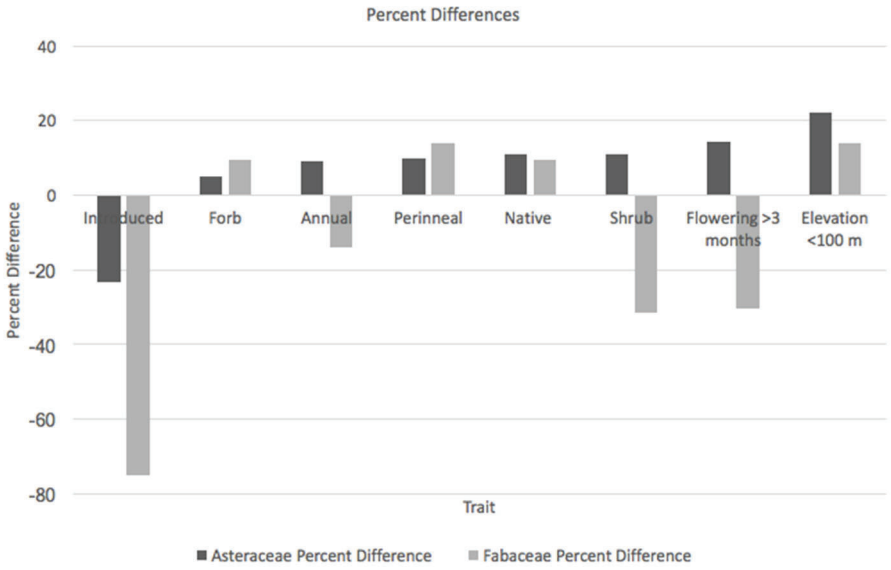

Figure 2: Percent difference totals.
This bar plot shows the calculated percent difference values compared to each other. Here you can see how collecting biases vary between taxonomic groups. While each family does exhibit the same types of biases, in many cases it is under-collected in one family and over-collected in the other.

When comparing the number of overall specimens collected each month (Figure 3), we found each family had a distinct trend. While species were over-collected during every month in Asteraceae, only April and May flowering species were over-collected in Fabaceae; all the other months showed under-collection. Asteraceae specimens were less over-collected during the late summer and early fall months (Aug-Nov) and were most over-collected in December and February.

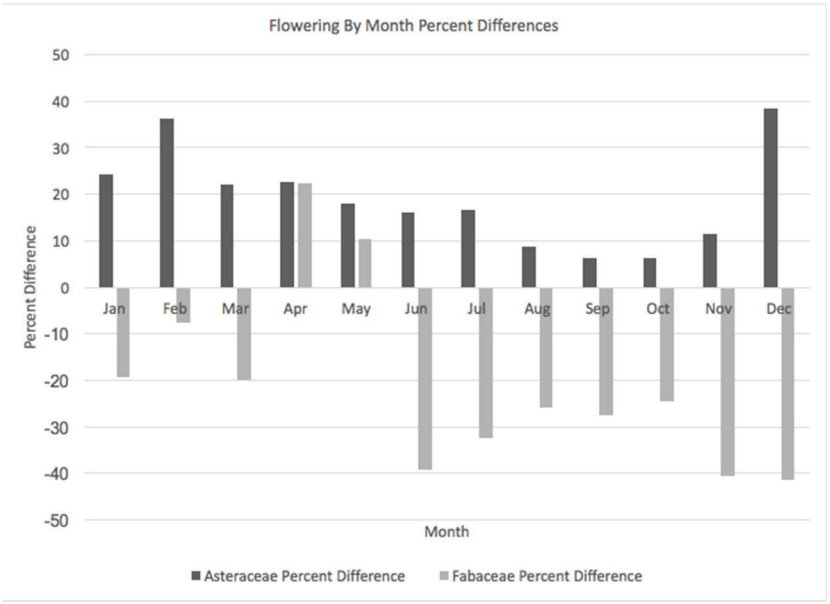

Figure 3. Percent Difference of specimens collected per month.
This graph shows the percent differences for each month of the year for California Asteraceae and Fabaceae species. Here you can see that Asteraceae experiences an overall over-collection bias, while Fabaceae experiences serious under-collection bias.

DISCUSSION

In the dataset of Asteraceae and Fabaceae herbarium specimens from California, significant biases towards collecting native, annual, perennial, forb, and shrub species were discovered. The results are also comparable to those of Daru *et. al* (2017), who conducted similar tests on vascular plants from Australia, South Africa, and New England. They used the same idea of comparing percent differences between actual collections and sets of randomly simulated datasets. However, unlike our comparisons between families, they tested differences within one family across multiple regions. It is important to consider, however, that at this point in time, not all herbaria have been digitized. The effort is still relatively new, and only a fraction of all existing collections are available online. This means that our dataset is not a perfect representation of species range, habitat, and traits. For the results discussed below it was assumed that the data well-represents the species studied.

It was predicted that introduced species would be over-collected, due to a stigma that the plants are more exotic or in high demand. However, this hypothesis was not supported, and introduced specimens in both families were under-collected more than was expected by chance (23% Asteraceae; 75% Fabaceae). This indicates that collectors are more likely to choose to collect native species, likely they seem more critical to study or have proven more useful in the types of studies herbarium data is typically used for. Alternatively, introduced species may be less abundant, especially at range edges. Historically, their distributions started much smaller and there was not much value in collecting them at the time. These conclusions are further supported by the evidence of over-collection of native species (11% Asteraceae; 9% Fabaceae). At their foundation, herbariums were mainly used for taxonomic purposes so it makes sense that collectors would focus on collecting the native plants for identification. The bias may be so pronounced in Fabaceae due to the fact that the sample size was small; compared to Asteraceae, and even small deviations from the mean would appear as high percent differences. Alternatively, the technique used to create the randomized collection data may be at fault. To determine how many specimens from each species should be collected for the unbiased randomized results, the number of counties the species was found in was considered; however, this did not account for how abundant the species actually were within the counties or the size of the county itself. For example, a species that appears in many counties may only be found a narrow range of disturbed habitats within those counties.

We also examined possible collection biases in species' growth habits. Plants classified as forb/herb were predicted to be collected more than shrubs because forb/herb species may be easier to collect, lacking woody structures. While the results supported the over-collection of forb/herbs in both families (five percent Asteraceae, ten percent Fabaceae), shrubs, too, were over-collected in Asteraceae by 11%. This inconsistency suggests that this pattern is ungeneralizable among taxonomic groups. However, the dataset of Fabaceae herb species was nearly five times as large as the woody species dataset (233 versus 43 species), meaning it could have produced an artificially large percent difference value. To avoid similar error in the future, it may be beneficial to compare families of the same size to make comparisons more accurate.

Looking at the elevation of species, it was predicted that specimens with a minimum elevation below 100 meters would be over-collected because they are easier for collectors to access. As expected, the simulations indicated an over-collection bias in both families. While 100 meters is not necessarily a very high altitude, these results suggest that as altitude increases, it gets harder to collect. In the future it would be beneficial to compare alpine species to sea level species to see if the trend remains.

Comparing annual versus perennial species, annuals appear typically more weed-like and may seem less desirable to collectors compared to the showier perennial plants. As such, annuals were expected to be more under-collected compared to perennials. In Asteraceae, annual species were over-collected by nine percent while perennials were over-collected by ten percent; for Fabaceae, annuals were under-collected by 14% while perennials were over-collected by 14%. In both families, there was a greater collection of perennials compared to annuals. It is a small percent difference in Asteraceae, however, which may indicate that the Fabaceae dataset was much smaller and more influenced by small deviations. Daru *et. al* 2017 also found a similar trend, with annual species being over-collected in both South Africa and New England.

It was predicted that species that flower longer than three months would be collected more frequently due to their availability. Collectors are more likely to collect specimens while they are in bloom, and species with a longer flowering period may thus be collected more often than a species that flowers for shorter periods. The results were mixed, as Asteraceae was over-collected by 14%, but Fabaceae specimens were under-collected by nearly 30%. This could be due to the availability of the families. Asteraceae is a more widespread family than Fabaceae, meaning that collectors might be over-collecting them compared to Fabaceae simply because there are more specimens and species to collect. This may also be due to flowering Asteraceae species standing out more in comparison to Fabaceae species, *i.e.*, with larger or more vibrant flower heads.

Finally, whether there was a bias toward collecting species that flowered in certain months of the year was examined. While there was a clear over-collecting bias in each month, it was much greater in some versus others (Figure 3). In Asteraceae, species that flower during the summer months are less over-collected than those that flower during winter months, which have percent

differences more than double those of the summer months. This could be due to conditions collectors are more likely to go out in. For instance, when the weather is neither excessively hot or extremely cold, collectors might be more willing to venture out and collect specimens. This could also be a result of the number of specimens that flower during certain months.

Alternatively, this could be a result of the availability of collectors. Over-collection of Asteraceae decreases drastically in August, September, and October, when most schools have begun again and many collectors may be teaching. Collection activity increases again in December, perhaps over Christmas Break when collectors may be able to go back out into the field. The number remains high in the spring months, likely due to the fact that at this time many species have begun to flower, making collectors more interested obtaining specimens.

For Fabaceae, the results were drastically different from the results of the Asteraceae test. As seen in Figure 4, there was a serious bias towards under-collection in almost every month of the year. Under-collection appears fairly constant throughout most months with the exception of two months in the spring, April and May, where there is an over-collection bias. This could be because of the availability of collectors during these months, as speculated for the California Asteraceae, or the result of the availability of the species. The peak seen in the spring could be due to spring bloom, as many species flower during this time and it is seen as a prime opportunity for collection. There is also notable under-collection during hotter months of the year, which could indicate collectors are more likely to go out when the weather conditions are more favorable. Daru *et. al* 2017 found a similar over-collection bias during the spring and summer months, which they conclude is due to collectors' desire to showcase species during these peak seasons. Due to the expense of collecting trips, it is logical to want to collect specimens that are blooming in order to best represent and identify the species. Much like how there was under-collection during hot months in California, there was serious under-collection in winter months in New Zealand, where winters are harsh and unfavorable. As hypothesized, they noticed that collection efforts increased during times of the year when schools are on vacation and during major holidays. However, they point out that tests like these fail to account for critical parts of a specimen's life cycle, such as buds and fruit maturation.[21] The more likely explanation is that the months in which collection efforts to showcase species in bloom occur line up with weather considered more agreeable.

In another recent study similar to ours, a number of additional different traits with collecting biases were highlighted and would be worth testing further in our dataset and other herbaria (Daru *et. al*, 2017). For instance, this study suggested an over-collection in roadside specimens and under-collection of threatened species, and also revealed that a large number of specimens were collected by a few major collectors, inferring that the individual biases the collectors held would influence the collections themselves. In future, it would be thus beneficial to see if these same types of biases exist within our study between families.

CONCLUSIONS

Our results demonstrate that there are collecting biases in herbarium data. Every test conducted supported a bias, either to under collect or over collect. This experiment was however, limited to studying the traits and species of just one state, California. The biases observed here may not be the same as in another state or country. These data do, however, serve as a good starting point for further, more in depth experiments to create actual correction methods for biases that can be used worldwide.

ACKNOWLEDGEMENTS

The authors thank Dr. Austin Mast at the Florida State University for oversight of this project, and Madeline Funaro for assistance with literature research. The authors also thank the Undergraduate Research Opportunity Program at the Florida State University for supporting the research of Jordan Williams. The authors thank Schmidt-Lebuhn for developing the method and sharing the code that was adapted for this study to simulate randomly collected specimen datasets. Katelin D. Pearson was partially supported through iDigBio, which is funded by a grant from the National Science Foundation's Advancing Digitization of Biodiversity Collections Program (award number 1547229). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

1. Lavoie, C. (2013). Biological collections in an ever changing world: Herbaria as tools for biogeographical and environmental studies. *Perspectives in Plant Ecology, Evolution and Systematics, 15*(1), 68-76. *doi:10.1016/j.ppees.2012.10.002*

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

3. Graham, C.H., Ferrier, S., Huettman, F., Moritz, C., Townsend Peterson, A., 2004. New developments in museum-based informatics and applications in biodiversity analysis. *Trends Ecol. Evol.* 19, 497–503

4. Hortal, J., Lobo, J. & Jiménez-Valverde, A. (2007). Limitations of Biodiversity Databases: Case Study on Seed-Plant Diversity in Tenerife, Canary Islands. *Conservation biology: the journal of the Society for Conservation Biology,* 21, 853-63. *10.1111/j.1523-1739.2007.00686.x*

5. Kadmon, R., Farber, O. & Danin, A. (2004). Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. *Ecological Applications* **14**, 401–413

6. Parnell, J. A. N., Simpson, D. A., Moat, J., Kirkup, D. W., Chantaranothai, P., Boyce, P. C., Bygrave, P., Dransfield, S., Jebb, M. H. P., Macklin, J., Meade, C., Middleton, D. J., Muasya, A. M., Prajaksood, A., Pendry, C. A., Pooma, R., Suddee, S. & Wilkin, P. (2003). Plant collecting spread and densities: Their potential impact on biogeographical studies in Thailand. *Journal of Biogeography* **30**, 193–209

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

8. Schulman, L., Toivonen, T., Ruokolainen, K., 2007. Analysing botanical collecting effort in Amazonia and correcting for it in species range estimation. *J. Biogeogr.* 34, 1388–1399

9. Soberón, J. M., Llorente, J. B. & Onate, L. (2000). The use of specimen-label databases for conservation purposes: an example using Mexican papilionid and pierid butterflies. *Biodiversity and Conservation* **9**, 1441–1466

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

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

12. Ahrends, A., Rahbek, C., Bulling, M.T., Burgess, N.D., Platts, P.J., Lovett, J.C., Wilkins Kindemba, V., Owen, N., Ntemi Sallu, A., Marshall, A.R., Mhoro, B.E., Fanning, E., Marchant, R., 2011. Conservation and the botanist effect. *Biol. Conserv.* 144, 131–140

13. Roberts, D. L., Taylor, L. and Joppa, L. N. (2016), Threatened or Data Deficient: assessing the conservation status of poorly known species. *Diversity Distrib.,* 22: 558–565. *doi:10.1111/ddi.12418*

14. Urmi, E., Schnyder, N., 2000. Bias in taxon frequency estimates with special reference to rare bryophytes in Switzerland. *Lindbergia* 25, 89–100

15. Bean, W. T., Stafford, R. and Brashares, J. S. (2012), The effects of small sample size and sample bias on threshold selection and accuracy assessment of species distribution models. *Ecography*, 35: 250–258. *doi:10.1111/j.1600-0587.2011.06545.x*

16. Delisle, F., Lavoie, C., Jean, M., Lachance, D., 2003. Reconstructing the spread of inva- sive plants: taking into account biases associated with herbarium specimens. *J. Biogeogr.* 30, 1033–1042

17. Hedenäs, L., Bisang, I., Tehler, A., Hamnede, M., Jaederfelt, K., Odelvik, G., 2002. A herbarium-based method for estimates of temporal frequency changes: mosses in Sweden. *Biol. Conserv.* 105, 321–331

18. Hofmann, H., Urmi, E., Bisang, I., Müller, N., Küchler, M., Schnyder, N., Schubiger, C., 2007. Retrospective assessment of frequency changes in Swiss bryophytes over the last two centuries. *Lindbergia* 32, 18–32

19. Rich, T.C.G., 2006. Floristic changes in vascular plants in the British Isles: geographi- cal and temporal variation in botanical activity 1836–1988. *Bot. J. Linn. Soc.* 152, 303–330

20. Schmidt-Lebuhn, A. N., Knerr, N. J., Kessler, M. 2013. Non-geographic collecting biases in herbarium specimens of Australian daisies (Asteraceae). *Biodiversity and Conservation.* 22: 905-919

21. Daru, Barnabas & Park, Daniel & Primack, Richard & Willis, Charles & Barrington, David & Whitfeld, Timothy & G. Seidler, Tristram & W. Sweeney, Patrick & R. Foster, David & M. Ellison, Aaron & Davis, Charles. (2018). Widespread sampling biases in herbaria revealed from large-scale digitization. *New Phytologist*, 217, 2: 939-955, *10.1111/nph.14855*

22. McCarthy, M. A. (1998). Identifying declining and threatened species with museum data. *Biological Conservation* **83**, 9–17.

23. Wolf, A., W. R. L. Anderegg, S. J. Ryan, and J. Christensen. 2011. Robust detection of plant species distribution shifts under biased sampling regimes. *Ecosphere* 2(10):115. *doi:10.1890/ES11-00162.1*

ABOUT STUDENT AUTHORS

Jordan Williams is a junior student in biology at the Florida State University who has been involved in research since her freshman year. She is eager to pursue botanical research upon her graduation, and her other scientific interests include genetics and marine biology.

PRESS SUMMARY

Biodiversity specimens are preserved organisms that can be used to study a variety of subjects, from the introduction of invasive species to climate change. Data about these specimens (*e.g.*, location, time, habitat) are aggregated in databases and used for analyses at larger scales than was previously possible. Many analyses assume that these data were randomly collected, which is important for accurate statistical results, but this is not always the case. Collectors are often biased in the way they collect specimens. This study examined how species' traits influence their collection and discovered that different collection biases exist in different groups of organisms. These results are important because the existence of collecting biases imply that trends determined from these data could be shaped by patterns of specimen collection rather than actual biological processes. Knowing what biases exist can help scientists understand how to account for them in their research.