BIAS IN OPEN BIODIVERSITY DATA: METHODOLOGICAL IMPLICATIONS FOR CONSERVATION DECISION-MAKING

Jelena Đukić[1]*,
[0009-0000-2115-7283]

Daniela Cvetković[2]
[0000-0001-7921-2222]

[1]Student,
Singidunum University,
Belgrade, Serbia

[2]Singidunum University,
Belgrade, Serbia

Abstract:

Modern nature conservation increasingly relies on open biodiversity data, especially on species occurrence records from digital platforms, collections, monitoring programs, and citizen science. These datasets can cover large areas and long time periods. However, their analytical value depends not only on data volume, but also on representativeness, metadata quality, and the way they were collected. One of the main problems is bias, because some areas, taxonomic groups, time periods, and observer types generate far more records than others. As a result, open data can give a distorted picture of biodiversity and affect models, trend estimates, and conservation priorities. This paper reviews the main forms of bias and links them with their effects on analytical results and decision-making. It concludes that open biodiversity data are highly valuable for modern nature conservation, but their analytical and practical value depends on careful methodological use and a clear understanding of their limits.

Keywords:

Open Data, Data Bias, Citizen Science, Nature Conservation, Biodiversity.

INTRODUCTION

Nature conservation depends on reliable, high-quality data on species, their distribution, and temporal dynamics. However, classical institutional monitoring is often limited by a lack of resources and weaker technical and organizational capacities [1]. Consequently, publicly available biodiversity data have become an important source of information for planning and implementing conservation measures [2]–[3].

The development of global repositories and aggregators, such as GBIF and OBIS, as well as many platforms that support citizen science projects, such as iNaturalist and eBird, has led to a rapid growth in data collection and reporting. At the same time, smartphones, with better cameras, more stable internet access, and GPS functions, have improved communication between volunteers and researchers [4]–[5]. Still, the way these data are collected often reflects site accessibility, participant movement, and interest in certain species more than actual ecological conditions on-the-ground [6]–[8].

Correspondence:
Jelena Đukić

e-mail:
jelena.djukic.23@singimail.rs

DOI: 10.15308/Sinteza-2026-570-576

These uneven patterns are not limited to single platforms or local datasets. They reflect broader patterns of biodiversity knowledge, including strong taxonomic unevenness and more general gaps in ecological data [9]–[10].

For that reason, this paper looks at open biodiversity data as both a data-related and a methodological problem. If some areas or taxa are recorded much more often than others, analyses can give a wrong picture of distribution, threat status, or trends [11]. This issue is important because it shows that data science approaches in ecology depend on data availability as well as on a clear understanding of how the data were generated and how bias can be identified and reduced [4].

Accordingly, the central research question of this paper is: how do the main forms of bias in open biodiversity data affect analytical outputs and, consequently, decision-making in nature conservation? The contribution of this paper is threefold: (1) it systematizes the main forms of bias in open biodiversity data, (2) connects them with typical analytical distortions, and (3) summarizes methodological approaches that can improve the reliability of their use in conservation practice.

2. METHODOLOGICAL APPROACH

This paper is based on a narrative literature review focused on recent studies dealing with open biodiversity data, citizen science, data bias, and methodological approaches for bias mitigation. Particular attention was given to papers published in 2024–2026, supplemented by earlier foundational works on FAIR principles, biodiversity data standards, and modelling approaches relevant to detection and sampling bias. The reviewed literature was selected based on its relevance to these questions: how bias arises in open biodiversity data, how it affects analytical results, and which methodological approaches can reduce or better account for it.

3. OPEN BIODIVERSITY DATA AND THEIR ANALYTICAL VALUE

Open biodiversity data include publicly available digital records on species, their occurrence, distribution, and associated ecological, taxonomic, and spatiotemporal attributes. These records can come from different sources, including field research, monitoring programs, museum and other collections, and observations made by citizens in citizen science projects [4]. Their importance lies in their broad spatial and temporal coverage. Thus, they are used to track changes in species distribution and abundance, to support macroecological analyses, and to identify conservation priorities [2], [12]. Their added value lies in their potential to be combined, searched, and reused in new studies, which makes them an important source of data for secondary analysis and broader scientific use [2], [4], [13].

The literature therefore distinguishes between primary and secondary biodiversity data. Primary data are direct records of organisms tied to a specific place and time. Secondary data are created when such records are taken from different and heterogeneous sources, often with uneven metadata quality, differences in documentation, and different forms of bias. Because of that, their analytical value must be assessed with caution [2], [4], [12].

To be usable, such data need to be structured, standardized, and described in a way that allows them to be linked with other datasets. In this context, the FAIR principles are important, as well as standards and tools such as the Darwin Core standard (DwC) and DMP-Tool, which make it easier to harmonize records from different sources and integrate them into broader repositories and combined datasets [4], [14]–[15].

Data bias is an important methodological challenge because it reduces representativeness and therefore limits the analytical value of open biodiversity data [11].

4. MAIN FORMS OF BIAS IN OPEN DATA

In this paper, bias in open biodiversity data is considered across four interrelated dimensions: spatial, taxonomic, temporal, and observer-related bias.

Spatial bias appears when data are unevenly distributed across space, that is, when some areas are recorded much more intensively than others [7], [16]. This unevenness often follows patterns of human movement and site accessibility, so urban areas, parks, and protected sites are often better documented than remote areas [8], [11]. This can make better-sampled areas appear more important for biodiversity than they actually are, while poorly documented areas remain underestimated in analyses and conservation planning [7], [11].

Taxonomic bias means that not all groups of organisms are equally represented in biodiversity databases. Some groups are recorded much more often because they are more visible, easier to identify, or attract more attention from researchers and users of digital platforms. As a result, the number of records for a group

does not always reflect how common it is in nature, but also how likely people are to observe and report it [5]–[6], [9].

Díaz-Calafat and colleagues [5] showed, using insects as an example, that both academic and citizen science data have strong taxonomic bias, with some orders being recorded and reported much more often than others. A similar pattern at a broader level was shown by García-Roselló and colleagues [7], who reported that European biodiversity data are much more complete for vertebrates and vascular plants than for many other taxonomic groups. Such unevenness reduces the ecological representativeness of databases [7]. At a broader level, Troudet and colleagues [9] showed that taxonomic bias in large global biodiversity databases is strongly linked to social preferences for some groups of organisms. This again shows that the distribution of records does not reflect only biological reality, but also human attention and interest [9].

Temporal bias refers to the fact that observations are not evenly distributed through time, but appear more often on certain days, in certain seasons, or during some parts of the year [5]. Opportunistic data therefore often reflect not only the real dynamics of nature, but also the rhythm of observer activity. Some periods are recorded more intensively because observers are more active, conduct fieldwork more frequently, or detect organisms more easily [5]. Consequently, temporal variation in data volume does not necessarily reflect real changes in species distribution or activity. They may also result from changes in observation frequency and recording practices [5], [8].

Observer bias is linked to the characteristics of the observers themselves. Using iNaturalist as an example, Grady and colleagues [8] showed that different user groups have different spatiotemporal patterns of reporting observations, and therefore have different effects on database structure.

Table 1. Main forms of bias in open biodiversity data

Type of bias	How it arises	Typical consequence	Possible response
Spatial	More recording in accessible, urban, or tourist areas	False picture of absence or low representation in poorly sampled areas	Targeted additional sampling and modelling that takes bias into account
Taxonomic	More frequent recording of attractive or easily recognized groups	Underestimation of less visible taxa and distorted comparisons between groups	Combining sources and checking taxonomic coverage
Temporal	Seasonality, weekends, changes in user activity, and field conditions	Unstable trend estimates and mixing of biological signal with sampling signal	Using temporal metadata and interpreting trends with caution
Observer	Differences in observer experience, motivation, and spatial habits	Heterogeneous recording probability and uneven observation quality	Validation, observer-behavior models, and transparent presentation of uncertainty

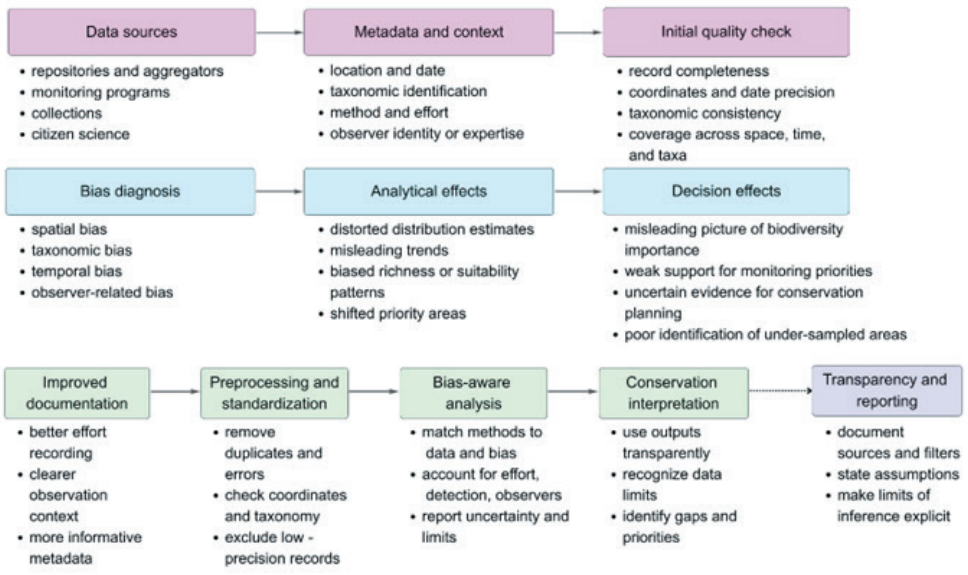

Figure 1. Simplified workflow showing how bias in open biodiversity data can propagate from record generation and documentation to analytical results, interpretation, and conservation decisions, together with the main stages at which bias can be recognized, reduced, and transparently reported

Guilbault and colleagues [16] further demonstrated that many approaches for correcting spatial bias assume that all observers behave in roughly the same way, which is not the case in practice. Their study distinguishes between behaviour types such as "explorers," who search for new and less visited locations, and "followers," who more often record observations at already known and frequently visited sites. This distinction matters because the effectiveness of spatial bias correction depends on how observers move and where they choose to record observations [16].

A brief overview of the main forms of bias, their consequences, and possible responses is shown in Table 1.

The relationships between data collection, bias, analytical effects, and conservation interpretation can be summarized as a simple workflow, shown in Figure 1.

5. HOW BIAS AFFECTS ANALYSES AND DECISION-MAKING

Bias in open biodiversity data can strongly affect analytical results and, with that, the conclusions derived from them. Bowler and colleagues [11] treat gaps and bias in biodiversity data as a problem that makes it much harder to draw reliable conclusions about species patterns and trends. The problem becomes particularly critical when the factors that influence sampling and data availability, such as site accessibility, urbanization, or human population density, overlap with the ecological and threat-related factors that influence the species themselves [11].

One of the most important consequences of bias is the systematic distortion of species distribution and representation. Díaz-Calafat and colleagues [5] demonstrated that insect records from the Iberian Peninsula are affected by strong taxonomic, spatial, temporal, and ecological biases, with some groups receiving much more attention than others. In such conditions, differences in the number of records between taxa cannot be interpreted as a direct sign of their true representation, because these numbers are also shaped by how visible organisms are, by observer interest, and by established recording habits [5], [9]. As a result, analyses may overestimate the importance of well-documented groups while underestimating less visible or under-recorded taxa.

A similar effect occurs in spatial analyses. García-Roselló and colleagues [7] showed that European biodiversity data are not evenly distributed, and that there are large differences between regions and taxonomic groups in coverage. When such data are used for spatial analyses, species distribution models, habitat suitability estimates, or the identification of priority conservation areas, the result may be systematically shifted toward well-studied areas. This means that areas with many records may appear more important or species-rich than they actually are, while poorly studied areas may be incorrectly interpreted as less valuable when they are simply under-sampled [7], [16].

Bias also strongly affects analyses of change over time. Díaz-Calafat and colleagues [5] identified seasonal and calendar effects in data collection patterns, while Grady and colleagues [8] showed that iNaturalist users have different spatiotemporal recording habits depending on their level and type of participation. This means that changes in the number of records over time do not necessarily reflect changes in abundance, activity, or distribution of a species. They may also result from variations in observer behaviour, availability, motivation, or recording routine [5], [8]. In that sense, a trend analysis may suggest that a species is spreading, declining, or changing its phenology, even though part of that pattern is driven by changes in observation effort.

Open biodiversity data are increasingly used, cited, or recommended in environmental impact assessment documents and related regulatory procedures. As a result, the effects of bias are no longer limited to scientific interpretation, but can also influence practical regulatory and consulting decisions, for example when documenting the presence or absence of species in the area of an urban development project [17].

In the European context, Moersberger and colleagues [1] point out that current monitoring systems still suffer from taxonomic, spatial, and temporal gaps and biases. Under such conditions, non-representative data affect the assessment of species and habitat status, the setting of monitoring priorities, and the selection of conservation measures [1].

6. POSSIBILITIES FOR REDUCING BIAS

Bias in biodiversity data cannot be fully eliminated [11], but it can be recognized, assessed, and reduced with appropriate methodological and statistical approaches [2], [16], [18]. This is particularly important because neither citizen science data nor academic records are free from different forms of bias, so uncritical use of both source types can lead to unreliable conclusions [5]. Approaches to addressing this issue can be identified at several interconnected levels, from data col-

lection and documentation, through data preprocessing, to modelling of the observation process and improved monitoring design.

The first level of mitigation concerns how data are collected and described. Better documentation of data facilitates the assessment of how biases arise and their incorporation into the analysis [4], [13]–[15]. For that reason, it is important to record not only species presence, but also information on observation effort, time, location, sampling method, observer identity and expertise, as well as information on what, where, and when observations are recorded, and what is not recorded [19]–[20]. Arazy and Malkinson [19] especially stress that part of observer-related bias can be reduced by "semi-structuring" unstructured citizen science data, that is, by adding basic information on user behaviour and observation context. Such metadata do not remove bias by themselves, but they make it more visible and analytically usable.

The second level of mitigation concerns preprocessing, that is, the filtering and standardization of data prior to statistical analysis. In practice, this includes removing duplicates, checking coordinates, harmonizing taxonomy, excluding records lacking sufficient spatial or temporal precision, and, where possible, standardizing observation effort [4], [13]–[15], [18], [20]. Johnston and colleagues [20] show that careful data selection and cleaning can significantly improve species distribution estimates, especially when combined with information on observation effort. In a similar way, Bowler and colleagues [11], working from a missing-data framework, point out that subsampling, weighting, and imputation can be useful, but that their effectiveness depends on how well the mechanisms underlying uneven data availability are understood. Thus, bias cannot be addressed through a single universal filter, but requires procedures tailored to its specific source.

The third level of mitigation concerns modelling the observation process itself. For biodiversity data, it is not sufficient to know only where a species is present. It is also necessary to account for the probability of detection and reporting. Accordingly, occupancy models are important because they distinguish true species presence from detection probability [21]–[22]. Van Strien and colleagues [21] showed that open biodiversity data can produce more reliable estimates of distribution trends when analyzed using models that account for variation in observation effort, incomplete reporting, and differences in detection. Schmidt and colleagues [22] further show that differences between observers can be an important source of detection heterogeneity and should therefore be included in the analysis. Therefore, models should, whenever possible, take into account observer characteristics, as well as seasonal, annual, and other factors affecting detection probability [19]–[20], [22].

Reducing spatial bias in species distribution modelling is a particular challenge, especially when presence-only and opportunistic data are used [7], [16], [23]–[25]. Presence-only data indicate where a species has been recorded, but not where it has been reliably searched for and not detected. Opportunistic data are typically collected without a predefined sampling plan, and therefore often reflect patterns of human movement and observation rather than evenly surveyed space [24]–[25]. In such cases, the literature recommends careful use of spatial filters, appropriate selection of background points, and other correction methods aimed at separating ecological signals from sampling bias [16], [23], [25]. Baker and colleagues [23] show that corrections for spatial sampling bias are not always automatically useful and that their effects should be evaluated by comparing models with and without correction. Fithian and Elith [25], as well as Mäkinen and colleagues [24], show that combining different data types can improve results compared to relying on a single source, as integrated models better account for heterogeneous sampling and improve predictions across the full species range.

It is also important to stress that reducing bias is not only a statistical problem, but also a monitoring design problem [12], [19], [22]. Bias can be partially reduced by directing observers toward poorly covered areas, collecting more information on their preferences and recording patterns, providing additional training, and strengthening validation procedures [18]–[19], [22]. Accordingly, information on observer behaviour and differences between observers is not merely an addition to the data, but an important component of the analytical framework [19], [22]. This approach links data collection with data analysis and creates better conditions for methodologically sound bias reduction [12], [19], [22].

Overall, reducing bias in biodiversity data requires a combination of approaches rather than reliance on a single method or data type [4], [11]–[12], [16], [18]–[25]. The goal is not to produce a "perfect" dataset, but to make the limitations of existing data more visible, interpretable, and more methodologically controlled. This is a key condition for their more reliable use in ecology and nature conservation.

7. CONCLUSION

Open biodiversity data are an important resource for modern ecology and nature conservation because they provide broader spatial and temporal coverage than can often be achieved through classical monitoring alone. Their value, however, depends not only on data volume and accessibility, but also on how well these data represent ecological reality.

This paper demonstrates that spatial, taxonomic, temporal, and observer-related bias can significantly influence species distribution estimates, trend analyses, distribution models, and the identification of conservation priorities. For that reason, bias should not be treated solely as a technical issue of data quality, but as a methodological issue affecting the entire chain from data collection to interpretation and decision-making.

The reviewed literature further shows that biases cannot be fully removed, but can be recognized, assessed, and mitigated through a combination of improved metadata, preprocessing, observation-process modelling, and enhanced monitoring design. The key implication is that open biodiversity data should not be treated as inherently neutral evidence. Their analytical and practical value depends on whether their limitations are explicitly identified and methodologically addressed prior to their use in conservation planning.

Taken together, more reliable biodiversity decisions require both continued data collection and a deeper understanding of how existing data are produced, filtered, modelled, and interpreted.

REFERENCES

[1] H. Moersberger et al., "Biodiversity monitoring in Europe: User and policy needs," Conservation Letters, vol. 17, no. 5, p. e13038, 2024, doi: 10.1111/conl.13038.

[2] L. J. Backstrom, C. T. Callaghan, N. P. Leseberg, C. Sanderson, R. A. Fuller, and J. E. M. Watson, "Assessing adequacy of citizen science datasets for biodiversity monitoring," Ecology and Evolution, vol. 14, no. 2, p. e10857, 2024, doi: 10.1002/ece3.10857.

[3] S. Pringle et al., "Opportunities and challenges for monitoring terrestrial biodiversity in the robotics age," Nature Ecology & Evolution, vol. 9, pp. 1031–1042, 2025, doi: 10.1038/s41559-025-02704-9.

[4] N. Marques et al., "Retrieving biodiversity data from multiple sources: making secondary data standardised and accessible," Biodiversity Data Journal, vol. 12, p. e133775, 2024, doi: 10.3897/BDJ.12.e133775.

[5] J. Díaz-Calafat, S. Jaume-Ramis, K. Soacha, A. Álvarez, and J. Piera, "Revealing biases in insect observations: A comparative analysis between academic and citizen science data," PLOS ONE, vol. 19, no. 7, p. e0305757, 2024, doi: 10.1371/journal.pone.0305757.

[6] E. J. Carlen et al., "A framework for contextualizing social-ecological biases in contributory science data," People and Nature, vol. 6, no. 2, pp. 377-390, 2024, doi: 10.1002/pan3.10592.

[7] E. García-Roselló et al., "Geographic biases undermine environmental representativeness of European biodiversity data," Science of the Total Environment, vol. 996, p. 180178, 2025, doi: 10.1016/j.scitotenv.2025.180178.

[8] E. L. Grady, C. J. Campbell, C. T. Callaghan, and R. P. Guralnick, "iNaturalist Users Exhibit Distinct Spatiotemporal Sampling Preferences, with Implications for Biodiversity Science and Project Planning," Citizen Science: Theory and Practice, vol. 11, no. 1, pp. 1-15, 2026, doi: 10.5334/cstp.868.

[9] J. Troudet, P. Grandcolas, A. Blin, R. Vignes-Lebbe, and F. Legendre, "Taxonomic bias in biodiversity data and societal preferences," Scientific Reports, vol. 7, p. 9132, 2017, doi: 10.1038/s41598-017-09084-6.

[10] J. Hortal, F. de Bello, J. A. F. Diniz-Filho, T. M. Lewinsohn, J. M. Lobo, and R. J. Ladle, "Seven shortfalls that beset large-scale knowledge of biodiversity," Annual Review of Ecology, Evolution, and Systematics, vol. 46, pp. 523–549, 2015, doi: 10.1146/annurev-ecolsys-112414-054400.

[11] D. E. Bowler, R. J. Boyd, C. T. Callaghan, R. A. Robinson, N. J. B. Isaac, and M. J. O. Pocock, "Treating gaps and biases in biodiversity data as a missing data problem," Biological Reviews, vol. 100, pp. 50-67, 2025, doi: 10.1111/brv.13127.

[12] A. Gonzalez et al., "From data to decisions: Toward a Biodiversity Monitoring Standards Framework," Proceedings of the National Academy of Sciences of the United States of America, vol. 123, no. 10, p. e2519347123, 2026, doi: 10.1073/pnas.2519347123.

[13] D. Roberts et al., "A framework for publishing primary biodiversity data," BMC Bioinformatics, vol. 12, Suppl. 15, p. I1, 2011, doi: 10.1186/1471-2105-12-S15-I1.

[14] M. D. Wilkinson et al., "The FAIR Guiding Principles for scientific data management and stewardship," Scientific Data, vol. 3, p. 160018, 2016, doi: 10.1038/sdata.2016.18.

[15] J. Wieczorek et al., "Darwin Core: An Evolving Community-Developed Biodiversity Data Standard," PLOS ONE, vol. 7, no. 1, p. e29715, 2012, doi: 10.1371/journal.pone.0029715.

[16] E. Guilbault, P. Somervuo, and I. W. Renner, "Explorers vs. followers: A behavioural approach to spatial bias correction in species distribution modelling," Ecological Modelling, vol. 510, p. 111311, 2025, doi: 10.1016/j.ecolmodel.2025.111311.

[17] C. T. Callaghan, C. Winnebald, B. Smith, B. M. Mason, and L. López-Hoffman, "Citizen science as a valuable tool for environmental review," Frontiers in Ecology and the Environment, vol. 23, no. 1, p. e2808, 2025, doi: 10.1002/fee.2808.

[18] I. E. Vessio, J. S. MacIvor, N. Persaud, C. Xia, and A. Filazzola, "Improving data reliability in community science projects with post-validation criteria," Journal for Nature Conservation, vol. 87, p. 127001, 2025, doi: 10.1016/j.jnc.2025.127001.

[19] O. Arazy and D. Malkinson, "A Framework of Observer-Based Biases in Citizen Science Biodiversity Monitoring: Semi-Structuring Unstructured Biodiversity Monitoring Protocols," Frontiers in Ecology and Evolution, vol. 9, p. 693602, 2021, doi: 10.3389/fevo.2021.693602.

[20] A. Johnston, W. M. Hochachka, M. E. Strimas-Mackey, V. Ruiz Gutierrez, O. J. Robinson, E. T. Miller, T. Auer, S. T. Kelling, and D. Fink, "Analytical guidelines to increase the value of community science data: An example using eBird data to estimate species distributions," Diversity and Distributions, vol. 27, no. 7, pp. 1265–1277, 2021, doi: 10.1111/ddi.13271.

[21] A. J. van Strien, C. A. M. van Swaay, T. Termaat, and V. Devictor, "Opportunistic citizen science data of animal species produce reliable estimates of distribution trends if analysed with occupancy models," Journal of Applied Ecology, vol. 50, no. 6, pp. 1450–1458, 2013, doi: 10.1111/1365-2664.12158.

[22] B. R. Schmidt, S. S. Cruickshank, C. Bühler, and A. Bergamini, "Observers are a key source of detection heterogeneity and biased occupancy estimates in species monitoring," Biological Conservation, vol. 283, p. 110102, 2023, doi: 10.1016/j.biocon.2023.110102.

[23] D. J. Baker, I. M. D. Maclean, and K. J. Gaston, "Effective strategies for correcting spatial sampling bias in species distribution models without independent test data," Diversity and Distributions, vol. 30, no. 3, p. e13802, 2024, doi: 10.1111/ddi.13802.

[24] J. Mäkinen, C. Merow, and W. Jetz, "Integrated species distribution models to account for sampling biases and improve range-wide occurrence predictions," Global Ecology and Biogeography, vol. 33, no. 3, pp. 356–370, 2024, doi: 10.1111/geb.13792.

[25] W. Fithian and J. Elith, "Bias correction in species distribution models: pooling survey and collection data for multiple species," Methods in Ecology and Evolution, vol. 6, no. 4, pp. 424–438, 2015, doi: 10.1111/2041-210X.12242.