Large-bodied birds are over-represented in unstructured citizen science data

Corey T. Callaghan[1,2,3,*], Alistair G. B. Poore[2], Max Hofmann[1,3], Christopher Roberts[2], Henrique M. Pereira[1,3]

[1]German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Puschstr. 4, 04103 Leipzig, Germany
[2]Ecology & Evolution Research Centre; School of Biological, Earth and Environmental Sciences; UNSW Sydney; Sydney, NSW
[3]Institute of Biology, Martin Luther University Halle-Wittenberg, Am Kirchtor 1, 06108 Halle (Saale), Germany

*Corresponding author:
email: corey.callaghan@idiv.de

Note: This is a pre-print and the final published version of this manuscript can be found here: https://www.nature.com/articles/s41598-021-98584-7

ABSTRACT

Citizen science platforms are quickly accumulating hundreds of millions of biodiversity observations around the world annually. Quantifying and correcting for the biases in citizen science datasets remains an important first step before these data are used to address ecological questions and monitor biodiversity. One source of potential bias among datasets is the difference between those citizen science programs that have unstructured protocols and those that have semi-structured or structured protocols for submitting observations. To quantify biases in an unstructured citizen science platform, we contrasted bird observations from the iNaturalist platform with that from a semi-structured citizen science platform — eBird — for the continental United States. We tested whether four traits of species (color, flock size, body size, and commonness) predicted if a species was under- or over-represented in the unstructured dataset compared with the semi-structured dataset. We found strong evidence that large-bodied birds were over-represented in the unstructured citizen science dataset; moderate evidence that common species were over-represented in the unstructured dataset; moderate evidence that species in large groups were over-represented; and no evidence that colorful species were over-represented in unstructured citizen science data. Our results suggest that biases exist in unstructured citizen science data when compared with semi-structured data, likely as a result of the detectability of a species and the inherent recording process. Importantly, in programs like iNaturalist the detectability process is two-fold — first, an individual needs to be detected, and second, it needs to be photographed, which is likely easier for many large-bodied species. Our results indicate that caution is warranted when using unstructured citizen science data in ecological modelling, and highlight body size as a fundamental trait that can be used as a covariate for modelling opportunistic species occurrence records, representing the detectability or identifiability in unstructured citizen science datasets. Future research in this space should continue to focus on quantifying and documenting biases in citizen science data, and expand our research by including structured citizen science data to understand how biases differ among unstructured, semi-structured, and structured citizen science platforms.

Keywords: citizen science; biases; opportunistic data; presence-only data, species occurrence data, eBird; iNaturalist; species traits; detectability

INTRODUCTION

Citizen science, or community science, — the involvement of volunteers in scientific endeavors — is increasingly seen as a cost-effective method for biodiversity monitoring and research. Accordingly, the quantity and diversity of citizen science projects in the ecological and environmental sciences is increasing [1]. Such projects are quickly accumulating hundreds of millions of biodiversity observations around the world annually [2,3] expanding the spatial and temporal scope of our understanding in ecology, conservation, and natural resource management [4,5]. Citizen science projects vary widely in their scope, design, and intent [6,7,8]. Projects can range from unstructured (e.g., little training needed to participate and contribute opportunistic/incidental observations) to semi-structured (e.g., with minimal workflows and guidelines, but additional data collected with each observation can be included) to structured (e.g., prescribed sampling in space and time by mostly trained and experienced volunteers). The level of structure consequently influences the overall data quality of a particular project [9,10].

Data quality from citizen science projects has been questioned [11, 12], and such concerns can act as a barrier to the widespread use of citizen science data in ecology and conservation [13]. These concerns arise because citizen science data can be biased temporally, spatially, and/or taxonomically. Temporally, many citizen science datasets are biased because participants frequently sample on weekends [14] or disproportionately during specific times of the year such as spring migration for birds [15], or during specific times of day, such as the morning period when birds are most active. Spatially, there is often a disproportionate number of sightings from areas with large human populations [16], areas with more accessibility [17], regions with high biodiversity that attract observers [18], and regions of the world with higher socioeconomic status [19].

Taxonomic biases also exist as some taxa receive orders of magnitude more citizen science observations than other taxa, evidenced by the fact that birds represent a disproportionate amount of data in the Global Biodiversity Information Facility [2]. Even within citizen science projects focused on specific taxa, there can be strong taxonomic biases towards particularly charismatic species or those that are readily identified [20, 21, 22, 23].

Despite potential biases in citizen science datasets, contrasts of data from volunteer participants to those contributed by more structured datasets have shown that citizen science programs can provide reliable data [12, 24]. For example, one case study found that mark-recapture models of whale sharks are reliable whether using sightings reported by the public or by experienced researchers [25], and another case study found that unstructured data performs comparably with structured data in identifying and monitoring invasive plant species [26]. When analyzed appropriately, citizen science data can further our understanding of many facets of biodiversity, including estimating species distributions [27, 28, 29], managing habitat for conservation [30], estimating species richness [31], monitoring pollination services [32], and quantifying population trends [33, 34]. In the above examples, highlighting the potential uses of citizen science data, statistical solutions to account for known biases and noise in citizen science data are used [3, 35, 36].

In addition to being an excellent resource for scientists to better understand ecological questions, citizen science projects are beneficial for society by encouraging increased engagement of the general public with science [37, 38]. Many citizen science projects provide learning opportunities for their volunteers. For example, participants in citizen science projects have increased their knowledge about invasive weeds [39, 40, 41], increased their knowledge of bird biology and behavior

[42], and even enhanced their conservation awareness and sense of place [42, 43]. The ecological advances derived from citizen science data, combined with the important role it plays in community engagement with science, suggests that citizen science data will continue to play an important role in ecological and conservation research in the future [2, 4, 38, 44]. However, what motivates volunteers to participate in science, and contribute observations, has important implications for the quality of the data obtained [45], particularly if there are biases towards certain species, places, or times of sampling.

To ensure the continued and expanded use of citizen science data in ecology and conservation, it is important to document and understand the different biases present in citizen science datasets. Importantly, the degree of bias in a particular dataset will be influenced by the level of structure of that citizen science project. For example, unstructured projects (e.g., iNaturalist, www.inaturalist.org) or semi-structured projects (e.g., eBird, www.ebird.org) will generally be more spatially biased than structured projects that have pre-defined spatial sampling locations (e.g., Breeding Bird Surveys). Or, a citizen science project that collects incidental presence-only data, such as iNaturalist, is likely more susceptible to individual observer preferences compared with a semi-structured or structured project that requires all species encountered to be recorded by the observers. Charismatic species [21] can be over-represented in citizen science data because observers are more likely to record species that they, or society, consider more interesting or relevant [46]. Similarly, rare species are more likely to be the subject of conservation monitoring or more likely to be actively searched for by amateur naturalists [47, 48] and thus can be over-represented in biodiversity datasets. In contrast, in some citizen science projects, abundant species can form a disproportionate number of records (e.g., [49]) because species' abundance and ease of identification can lead to an increase in the number of records by casual observers [50]. Inherently linked with observer preferences are issues of differences in species detectability [50], and the ease of making the observations. Therefore, species traits (e.g., body size, color, flock size) may have an additive effect, influencing both the detectability of a species [51, 52, 53], and in turn, the likelihood of a species being submitted to an unstructured citizen science database.

Quantifying biases in citizen science datasets can help (1) researchers using these data to better account for biases when drawing ecological conclusions, (2) the design and implementation of future citizen science projects, and (3) understand what species or regions may need data collection from professional scientists by understanding the 'limits' of citizen science projects [19]. Here, we quantify biases in bird observation data from an unstructured, citizen science project — iNaturalist — with that from a semi-structured one — eBird. We restricted our comparison to birds because (1) birds are among the most popular taxa with the non-scientific public, ensuring large sample sizes in both citizen science projects, and (2) data on the species traits that may influence the likelihood of unstructured observations are readily available for birds. We assessed the over-representation or under-representation of bird species' observations in the unstructured citizen science project compared to the semi-structured project (see Figure 1). We then tested the following predictions: that (1) more colorful species; (2) larger species; (3) species with the 'least concern' IUCN status; and (4) more gregarious species (i.e., with larger flock sizes) are over-represented in the unstructured citizen science dataset (iNaturalist) in contrast to the semi-structured citizen science dataset (eBird). Our analysis highlights the importance of considering species' traits when using citizen science data in ecological research.

METHODS

We made comparisons between iNaturalist (www.inaturalist.org) — an unstructured citizen science project — and eBird (www.ebird.org) — a semi-structured citizen science project [15, 54].

iNaturalist citizen science data. iNaturalist is a multi-taxon citizen science project hosted by the California Academy of Sciences. It is an unstructured citizen science project where volunteers contribute opportunistic photos or sound recordings through a smart-phone or web-portal. Photos are then identified to the lowest possible taxonomic resolution using a community identification process, and once two users, or more than two-thirds, confirm the species-level identification of an organism it is considered "research grade". Observations that are research grade are then uploaded to the Global Biodiversity Information Facility. We downloaded iNaturalist observations from the Global Biodiversity Information Facility for the contiguous United States [55] for the period from January 2010 to May 2019, on December 3rd, 2019. For more details on the iNaturalist methodology, see here: https://www.inaturalist.org/pages/getting+started.

eBird citizen science data. eBird is one of the most successful citizen science projects in the world, with > 1 billion bird observations globally. It was launched in 2002 by the Cornell Lab of Ornithology and focuses on collecting reliable data on the distributions and relative abundance of birds throughout the world [54]. It is a semi-structured project where volunteers submit 'checklists' of birds seen and/or heard on birding outings, following different protocols (e.g., stationary, incidental, or travelling). An important component of eBird that differentiates it from unstructured data collection is that users are required to indicate whether the checklist is 'complete' – meaning they included all species they were able to identify during that birding outing. When using complete checklists only in an analysis, a user can infer non-detections in the dataset for any species not recorded. Observers can submit checklists at any time and place of their choosing, and for any duration and distance travelled. Most non-incidental checklists additionally include the duration and distance travelled while birding. Filters are set — based on spatiotemporal coordinates — which restrict the species and their associated counts that can be submitted without approval from a regional expert reviewer [56]. We used the eBird basic dataset (version ebd_May-2019) and restricted our analysis to data from the contiguous United States for the period from January 2010 to May 2019. We also restricted the data used to those of the best 'quality' by excluding incomplete checklists, checklists that were incidental or historical, which travelled >5 km, lasted <5 min, and lasted >240 min, minimizing the leverage of outliers on analyses [57, 58].

Filtering and aggregating the citizen science datasets. Although both datasets are global in scope, we restricted our analysis to the contiguous United States as both of these citizen science projects initiated in the United States, and thus the data are most numerous from there. For comparisons, we aggregated data at the state level. This was done to account for differences that may exist throughout the entirety of the United States including differences in user behavior and the species pools that differ geographically. We used the eBird Clements taxonomy (version 2018) and all species from iNaturalist were matched with this taxonomy. A total of 1,030 species was initially collated from the eBird checklists, but many of these only occurred once or a few times — possibly representing misidentifications that had not yet been fixed by local reviewers or escaped and exotic birds which are incorporated in the eBird dataset but not considered part of the local avifauna or of interest to our analysis here. Although, these could represent scarce and uncommon species in a state as well, albeit these are rarely sampled by iNaturalist. To account for these biases, we removed species that were on <1% of eBird checklists for a given state; trimming the eBird observations to the 'core' suite of species that occur in a state (*sensu* [57]). After trimming the species and harmonizing the taxonomy with iNaturalist, there were 507 species remaining which were considered in our main analyses presented throughout the results. Although our results here are presented using the 1% cutoff level, we tested the sensitivity of this cutoff level and found comparable results across 0, 0.5, 1, and 1.5% cutoffs. For each state, the eBird and iNaturalist data were summarized by calculating the total number of observations in that state for every species. Using these aggregated data, we conducted preliminary comparisons of the unstructured and semi-structured datasets by quantifying the relationship between the number of eBird and iNaturalist observations at the state level, and at the species level.

Species-specific over- or under-representation in iNaturalist

Our first analytical step was to model the log-log linear relationship between the total number of observations in iNaturalist and total number of observations in eBird for a species (Figure 1). This linear model was repeated separately for each state, where the response variable was log-transformed number of iNaturalist observations and the predictor variable was log-transformed number of eBird observations. Each model fitted was stratified by state, to account for inherent differences among states that were not of interest in our particular analysis, such as (1) the number of observers in a state, (2) the different relative abundance of a species throughout the United States, and (3) any other intrinsic differences that might exist among states that was not of interest in our analysis. A species with a high (i.e., positive) residual would be over-represented in iNaturalist relative to eBird, whereas a species with a low (i.e., negative) residual would be under-represented in iNaturalist relative to eBird (Figure 1). Then we took the residuals from these models and used these as the response variables in our subsequent analyses of species characteristics (see below).

Species-specific trait data

We tested whether four predictor variables (see Figure 1) would explain the over- or under-representation of bird species in the unstructured citizen science data. For each species, we used a proxy for their commonness/abundance, categorized according to IUCN status, taken from HBW BirdLife international checklist version 3 (http://datazone.birdlife.org/species/taxonomy). This variable was treated as an ordinal variable in our models (see below) and encompassed Least Concern, Vulnerable, and Near Threatened species. The three species recorded as endangered were removed from this analysis due to a lack of power at this level with so few observations. For each species we used the continuous predictor variables of (1) body size; (2) color; and (3) average flock size. Body sizes (adult body mass in grams) were taken from the amniote life history database compiled by Myhrvold et al. [59] and were log-transformed to meet normality assumptions. Color was taken from Dale et al. 2015 [60] and was extracted as RGB values for six patches per species (upper breast, lower breast, crown, forehead, nape, throat). To define a continuum of color where the brightest/most colorful (and likely most detectable species based on color) had the highest value we combined both the 'distance from brown' and the 'brightness' of a species for the data from Dale et al. 2015 [60]. Distance from brown was defined as the maximum Euclidian distance in the cubic RGB color space from brown (R = 102, B = 68, G = 0) from any of the six patches on a species, regardless of sex. Brightness was defined as the maximum relative luminance (i.e., 0.2126R + 0.7152G + 0.0722B) from any of the six patches on a species, regardless of sex. These two variables were combined and scaled from 0 to 1 for all species in Dale et al. 2015 [60] and this value was used as our measure of color. Calculations were done in “Lab” space, an approximately perceptually uniform color space standardized by the Commission Internationale d'Eclairage. Exploratory analyses showed similar results with HSV color space. Flock size — an approximation of the gregariousness of a species — was taken from eBird as the average number of reported individuals among all checklists where a species was reported, across all data. We acknowledge that the number of a species reported on an eBird checklist likely encompasses both the gregariousness of a species as well as the density of a species in an area, as birders can travel through multiple territories.

Statistical analysis

We used mixed effects models to examine the effects of species traits on the relative bias between our unstructured and semi-structured citizen science datasets. The response variable was the residuals from a log-log linear model fit between the eBird observations and the iNaturalist observations for a given species (described above), the predictor variables were the respective traits, and the random effect was state. First, we ran a global model where all traits were included as predictor variables: log10-transformed body size, log10-transformed flock size, color, and IUCN status treated as an ordinal variable. Second, to confirm the results of this global model, we ran four separate models – one for each trait as listed above – because there was much missing data for species’ traits. This approach allowed us to test the relationship of a predictor given the other predictor variables (i.e., all predictors against the response variable simultaneously) as well as the independent relationships (i.e., each predictor separately against the response variable).

Data analysis and availability

All analyses were carried out in R software [61] and relied heavily on the tidyverse workflow [62]. Mixed-effects models were fitted using the lme4 package [63] and p-values were extracted using the lmerTest package [64]. Data and code to reproduce these analyses are available here: https://doi.org/10.5281/zenodo.5509770.

RESULTS

A total of 255,727,592 eBird and 1,107,224 iNaturalist observations were used in our analysis. At the state level, the number of eBird checklists and the number of iNaturalist observations were strongly correlated (Figure 2a; $R^2 = 0.58$, p-value < 0.001). Similarly, at the species level, the total number of iNaturalist observations and eBird observations for a given species was strongly correlated (Figure 2b; $R^2 = 0.9$), and for both datasets the number of observations per species was positively-skewed (Figure S1). We also found that the percent of eBird checklists a species was found on and the percent of total iNaturalist observations a species comprised was strongly correlated among states (Figure S2), suggesting that species are sampled to a proportionally similar extent in unstructured and semi-structured citizen science projects.

Across the 507 species included in our analyses (Table S1), we showed that larger species were more likely to be over-represented in the unstructured citizen science dataset, and this was true in most states, as illustrated by the empirical comparison (Figure 3a). The empirical comparison also showed over-representation of flock size in the unstructured dataset, although some states showed a negative relationship indicating the possibility that this trait varies in space (Figure 3b).

There was no discernible pattern in the relationship between color and over- or under-representation in iNaturalist data (Figure 3c), and there was some evidence that Least Concern species were over-represented in the iNaturalist data (Figure 3d).

In contrast to our empirical comparisons (Figure 3), our mixed effects multiple regression linear model (N=3986) with state as a random effect (Figure 4) found strong evidence that body size (parameter estimate=0.049; 95% CI=0.023, 0.073) and flock size (parameter estimate=0.051; 95% CI=0.034, 0.069) were over-represented in iNaturalist compared with eBird; moderate evidence that common species were over-represented (parameter estimate=0.027; 95% CI=-0.003, 0.058); and no evidence that color influenced the over- or under-representation of a species in iNaturalist (parameter estimate=-0.008; 95% CI=-0.064, 0.048). The patterns found in the multiple regression model qualitatively matched that of the individual trait models, where more observations were included in some instances (see Table S2).

DISCUSSION

We compared two popular citizen science platforms throughout the continental United States and found that there was strong agreement between the relative number of observations of a species in iNaturalist and eBird, albeit there were about 200 times more observations in eBird than iNaturalist. This suggests that species are observed at similar rates in both citizen science projects — i.e., the inherent processes driving observation in both unstructured and semi-structured citizen science projects are similar. Nevertheless, in support of our predictions (Figure 1) we found strong evidence that large-bodied birds are over-represented in the unstructured citizen science dataset compared with the semi-structured dataset. We also found moderate evidence that common species were over-represented in the unstructured data, and weak evidence that species in large flocks were over-represented. In contrast to our prediction, however, we found no evidence that brightly-colored species were over-represented in unstructured citizen science data.

Our finding that large-bodied birds were over-represented in an unstructured citizen science dataset is probably because larger-bodied birds are more detectable [53, 65]. This confirms previous research which has shown that smaller-bodied taxa are under-represented in citizen science data [66, 67, 68], but this may not be the case for some taxa such as mammals [69]. However, it is difficult to know whether this is an inherent preference shown by users of the unstructured citizen science data, or if this comes about as part of the recording process (e.g., species' detectability; [50]). Species detectability is complex and can be linked to species' mobility or habitat preferences of the species themselves; for example, large-bodied wading birds generally occurring in open wetlands are more easily detected than small-bodied songbirds generally occurring in dense forest understory.

Related to detectability, an important distinction between iNaturalist and eBird is how identifications are made. For an observer to make a record in iNaturalist, usually a photograph is uploaded (although sound recordings are also accepted). Because a photograph is needed, the detectability process is two-fold — first, it needs to be detected, and second, it needs to be photographed, which is likely easier for many large-bodied birds. Longer lenses, often restricted to serious photographers, may be needed to photograph smaller-bodied birds whereas smartphones can usually capture a sufficient image of a larger-bodied bird. In contrast to iNaturalist, in eBird, a lot of identifications are made acoustically, and identification can sometimes also use contextual clues such as behavior or habitat of the bird — often difficult to capture in a photograph. Most traits analyzed here are related to visual encounter/identification, thus likely explaining the differences found between the unstructured iNaturalist and the semi-structured eBird data. To illustrate this difference, in New York state, the most under-represented species in iNaturalist (i.e., with the lowest residuals) are Marsh Wren, American Crow, Warbling Vireo, Least Flycatcher, Willow Flycatcher – all species that are identified largely acoustically. In contrast, the most over-represented species in iNaturalist (i.e., with the highest residuals) are House Sparrow, American Robin, Palm Warbler, Northern Mockingbird – all species that are easy to visually see and thus detect and photograph (Table S1). Therefore, the bias towards large-bodied birds in the unstructured data is probably a result of detectability and the ability to capture a photograph [53]. Photographs can also be uploaded to eBird, and a further test of this hypothesis could interrogate the species in eBird which have photographs uploaded. This process is similar in insects, for example, which are generally small, but larger insects (e.g., butterflies) are both easier to observe, photograph, and identify — making it likely that the biases we found in birds generalize to insects as well. Indeed, a study of bugs and beetles found that smaller species are typically less represented in citizen science data [68]. Importantly, because this body size bias is systematic, it is likely easier to model as we know that this data is not missing at random (e.g., [70]) and thus body size can be included in various modelling processes when using unstructured citizen science data (e.g., [67]).

Similar to body size, we found that birds which occur in larger groups (i.e., flocks) are over-represented in the unstructured dataset. This, again, may be inherently linked to the recording process, rather than a specific bias or preference of the observers themselves. This is because common birds, that occur in large flocks, are more likely to be seen and thus submitted to the unstructured citizen science data [65]. A larger flock will likely also provide more opportunities to capture a photograph than when observing a single individual, as has been illustrated in the detectability of animals from aerial surveys by professionals [71].

One explanation for the least concern birds being somewhat over-represented in iNaturalist is user behavior — eBird data are more likely to be derived from avid birdwatchers (e.g., those that search out uncommon birds and keep serious lists) compared with iNaturalist data which may be derived from more recreational birdwatchers that focus on 'backyard' species. The types of participants, and their motivations, of iNaturalist and eBird are therefore likely very different as has generally been shown among citizen science projects (e.g., [72]). Participants submitting observations to eBird are likely better at identifying birds than those submitting to iNaturalist and can also rely on acoustic and contextual clues to make identifications, as discussed above. Importantly, our analysis focused on only unstructured versus semi-structured data, but future work should expand this comparison to include structured datasets (e.g., breeding bird surveys) to understand if the biases found here also exist when compared with more structured datasets. For example, there may be a skew in eBird data towards rare birds when compared to standardized surveys (e.g., breeding bird surveys) resulting from birders preferentially adding rare and uncommon species. Such a finding would further highlight the divergence in behavior between the users of iNaturalist and eBird.

The lack of signal of the colorfulness of a species in predicting over-representation in iNaturalist could suggest that iNaturalist users are not limited by ‘attractiveness/aesthetics’ but mostly by detectability, as discussed above (Figure 4). Alternatively, the lack of a signal here could be a result of the comparison being between a semi-structured and an unstructured dataset – i.e., both eBird and iNaturalist are skewed towards more colorful species, and a comparison with a structured dataset will help test this hypothesis. Quantifying the influence of color on detectability remains a challenge (e.g., [73]). In contrast to our results, others have demonstrated a clear preference of ‘color’ by the general public in online google searches of birds [74]. However, the role of aesthetics, or color, by the public may be complex as illustrated by one study which found that only blue and yellow were significant in determining bird ‘beauty’ [75]. In other taxa, more colorful insect species are more commonly reported [68], as well as more patterned and morphologically interesting species. This may suggest, at least in the case of insects, that contributors are selecting subjects based on their visual aesthetics, not just their detectability. The discrepancies between our results and that of [68] suggest that the influence of traits may vary between different taxa, making it important to explore these relationships for a range of organisms rather than extrapolating the results of birds, or bugs and beetles, to other groups.

While citizen science data are undoubtedly valuable for ecology and conservation [4, 76, 77], there remain limits to the use of citizen science datasets [13, 78]. The ability to sample remote regions, for example, will likely remain a limitation in citizen science data, and this has been well-recognized [17]. Quantifying the limits of citizen science datasets for use in ecology and conservation remains an important step for the future widespread use of citizen science data in ecology and conservation. Data-integration — where noisy citizen science data are integrated with professionally-curated datasets — will likely be increasingly important in the future use of citizen science data [79, 80]. By knowing the biases present in citizen science data, experts can preferentially generate data that maximize the integration process, for example by collecting data from remote regions. Further, professional scientists should use limited funding to target species that are likely to be under-represented in citizen science datasets — i.e., rare, small-bodied, species.

Ultimately, citizen science data will continue to perform, at least in part, a substantial role in the future of ecology and conservation research [44]. Understanding, documenting, and quantifying the biases associated with these data remains an important first step before the widespread use of these data in answering ecological questions and biodiversity monitoring [5]. Our results highlight that for birds, semi-structured eBird out-samples unstructured iNaturalist data, but the number of observations recorded per species are strongly correlated between the two platforms. When looking at the differences in this relationship, it is clear that biases exist, likely due to the biases in the unstructured iNaturalist data. We note that we compared the unstructured dataset to a semi-structured dataset, and the semi-structured dataset does not necessarily represent the "truth". The biases found here, could also be present when comparing a semi-structured dataset to true density or abundance of birds in the landscape. To better understand these differences, future research in this space should continue to focus on quantifying and documenting biases in citizen science data, and understanding how these biases change from unstructured to semi-structured to structured citizen science platforms. Nevertheless, our results demonstrate the importance of using species-specific traits, when modelling citizen science datasets [27, 29, 52, 81, 82, 83,84].

ACKNOWLEDGEMENTS

We thank the countless contributors to both eBird and iNaturalist who contribute their observations as well as the Cornell Lab of Ornithology and the California Academy of Sciences to their commitment of making citizen science data open access. CTC, HMP, and MH acknowledge funding of iDiv via the German Research Foundation (DFG FZT 118). CTC was supported by a Marie Skłodowska-Curie Individual Fellowship (No 891052).

AUTHOR CONTRIBUTIONS

CTC conceived and led the study with input from all authors. CTC performed the analyses with input from all authors. CTC wrote the first version of the manuscript and all authors contributed to editing the manuscript.

COMPETING INTERESTS

The author(s) declare no competing interests.

REFERENCES

1. Pocock, M. J., Tweddle, J. C., Savage, J., Robinson, L. D. & Roy, H. E. The diversity and evolution of ecological and environmental citizen science. *PLoS One* **12**, e0172579 (2017).

2. Chandler, M. *et al.* Contribution of citizen science towards international biodiversity monitoring. *Biological Conservation* **213**, 280–294 (2017).

3. Chandler, M. *et al.* Involving citizen scientists in biodiversity observation. in *The GEO handbook on biodiversity observation networks* 211–237 (Springer, Cham, 2017).

4. McKinley, D. C. *et al.* Citizen science can improve conservation science, natural resource management, and environmental protection. *Biological Conservation* **208**, 15–28 (2017).

5. Pereira, H. M. *et al.* Monitoring essential biodiversity variables at the species level. in *The GEO handbook on biodiversity observation networks* 79–105 (Springer, Cham, 2017).

6. Wiggins, A. & Crowston, K. From conservation to crowdsourcing: A typology of citizen science. in *2011 44th hawaii international conference on system sciences* 1–10 (IEEE, 2011).

7. Haklay, M. Citizen science and volunteered geographic information: Overview and typology of participation. in *Crowdsourcing geographic knowledge* 105–122 (Springer, 2013).

8. Kelling, S. *et al.* Using semistructured surveys to improve citizen science data for monitoring biodiversity. *BioScience* **69**, 170–179 (2019).

9. Welvaert, M. & Caley, P. Citizen surveillance for environmental monitoring: Combining the efforts of citizen science and crowdsourcing in a quantitative data framework. *SpringerPlus* **5**, 1890 (2016).

10. Callaghan, C. T., Rowley, J. J., Cornwell, W. K., Poore, A. G. & Major, R. E. Improving big citizen science data: Moving beyond haphazard sampling. *PLoS Biology* **17**, e3000357 (2019).

11. Bonter, D. N. & Cooper, C. B. Data validation in citizen science: A case study from project FeederWatch. *Frontiers in Ecology and the Environment* **10**, 305–307 (2012).

12. Kosmala, M., Wiggins, A., Swanson, A. & Simmons, B. Assessing data quality in citizen science. *Frontiers in Ecology and the Environment* **14**, 551–560 (2016).

13. Burgess, H. K. *et al.* The science of citizen science: Exploring barriers to use as a primary research tool. *Biological Conservation* **208**, 113–120 (2017).

14. Courter, J. R., Johnson, R. J., Stuyck, C. M., Lang, B. A. & Kaiser, E. W. Weekend bias in citizen science data reporting: Implications for phenology studies. *International journal of biometeorology* **57**, 715–720 (2013).

15. Sullivan, B. L. *et al.* The eBird enterprise: An integrated approach to development and application of citizen science. *Biological Conservation* **169**, 31–40 (2014).

16. Kelling, S. *et al.* Can observation skills of citizen scientists be estimated using species accumulation curves? *PloS one* **10**, e0139600 (2015).

17. Tiago, P., Ceia-Hasse, A., Marques, T. A., Capinha, C. & Pereira, H. M. Spatial distribution of citizen science casuistic observations for different taxonomic groups. *Scientific reports* **7**, 1–9 (2017).

18. Geldmann, J. *et al.* What determines spatial bias in citizen science? Exploring four recording schemes with different proficiency requirements. *Diversity and Distributions* **22**, 1139–1149 (2016).

19. Callaghan, C. T. *et al.* Three frontiers for the future of biodiversity research using citizen science data. *BioScience* **71**, 55–63 (2021).

20. Ward, D. F. Understanding sampling and taxonomic biases recorded by citizen scientists. *Journal of insect conservation* **18**, 753–756 (2014).

21. Troudet, J., Grandcolas, P., Blin, A., Vignes-Lebbe, R. & Legendre, F. Taxonomic bias in biodiversity data and societal preferences. *Scientific reports* **7**, 1–14 (2017).

22. Martín-López, B., Montes, C., Ramírez, L. & Benayas, J. What drives policy decision-making related to species conservation? *Biological Conservation* **142**, 1370–1380 (2009).

23. Boakes, E. H. *et al.* Distorted views of biodiversity: Spatial and temporal bias in species occurrence data. *PLoS Biol* **8**, e1000385 (2010).

24. Aceves-Bueno, E. *et al.* The accuracy of citizen science data: A quantitative review. *Bulletin of the Ecological Society of America* **98**, 278–290 (2017).

25. Davies, T. K., Stevens, G., Meekan, M. G., Struve, J. & Rowcliffe, J. M. Can citizen science monitor whale-shark aggregations? Investigating bias in mark–recapture modelling using identification photographs sourced from the public. *Wildlife Research* **39**, 696–704 (2013).

26. Crall, A. W. *et al.* Assessing citizen science data quality: An invasive species case study. *Conservation Letters* **4**, 433–442 (2011).

27. Strien, A. J. van, Swaay, C. A. van & Termaat, T. Opportunistic citizen science data of animal species produce reliable estimates of distribution trends if analysed with occupancy models. *Journal of Applied Ecology* **50**, 1450–1458 (2013).

28. Johnston, A., Moran, N., Musgrove, A., Fink, D. & Baillie, S. R. Estimating species distributions from spatially biased citizen science data. *Ecological Modelling* **422**, 108927 (2020).

29. Tiago, P., Pereira, H. M. & Capinha, C. Using citizen science data to estimate climatic niches and species distributions. *Basic and Applied Ecology* **20**, 75–85 (2017).

30. Sullivan, B. L. *et al.* Using open access observational data for conservation action: A case study for birds. *Biological Conservation* **208**, 5–14 (2017).

31. Callaghan, C. T. *et al.* Citizen science data accurately predicts expert-derived species richness at a continental scale when sampling thresholds are met. *Biodiversity and Conservation* **29**, 1323–1337 (2020).

32. Birkin, L. & Goulson, D. Using citizen science to monitor pollination services. *Ecological Entomology* **40**, 3–11 (2015).

33. Delaney, D. G., Sperling, C. D., Adams, C. S. & Leung, B. Marine invasive species: Validation of citizen science and implications for national monitoring networks. *Biological Invasions* **10**, 117–128 (2008).

34. Schultz, C. B., Brown, L. M., Pelton, E. & Crone, E. E. Citizen science monitoring demonstrates dramatic declines of monarch butterflies in western north america. *Biological Conservation* **214**, 343–346 (2017).

35. Bird, T. J. *et al.* Statistical solutions for error and bias in global citizen science datasets. *Biological Conservation* **173**, 144–154 (2014).

36. Isaac, N. J., Strien, A. J. van, August, T. A., Zeeuw, M. P. de & Roy, D. B. Statistics for citizen science: Extracting signals of change from noisy ecological data. *Methods in Ecology and Evolution* **5**, 1052–1060 (2014).

37. Dickinson, J. L. *et al.* The current state of citizen science as a tool for ecological research and public engagement. *Frontiers in Ecology and the Environment* **10**, 291–297 (2012).

38. Bonney, R. *et al.* Next steps for citizen science. *Science* **343**, 1436–1437 (2014).

39. Jordan, R. C., Gray, S. A., Howe, D. V., Brooks, W. R. & Ehrenfeld, J. G. Knowledge gain and behavioral change in citizen-science programs. *Conservation biology* **25**, 1148–1154 (2011).

40. Crall, A. W. *et al.* The impacts of an invasive species citizen science training program on participant attitudes, behavior, and science literacy. *Public Understanding of Science* **22**, 745–764 (2013).

41. Jordan, R. C., Ballard, H. L. & Phillips, T. B. Key issues and new approaches for evaluating citizen-science learning outcomes. *Frontiers in Ecology and the Environment* **10**, 307–309 (2012).

42. Evans, C. *et al.* The neighborhood nestwatch program: Participant outcomes of a citizen-science ecological research project. *Conservation Biology* **19**, 589–594 (2005).

43. Haywood, B. K., Parrish, J. K. & Dolliver, J. Place-based and data-rich citizen science as a precursor for conservation action. *Conservation Biology* **30**, 476–486 (2016).

44. Pocock, M. J. *et al.* A vision for global biodiversity monitoring with citizen science. in *Advances in ecological research* vol. 59 169–223 (Elsevier, 2018).

45. Tiago, P., Gouveia, M. J., Capinha, C., Santos-Reis, M. & Pereira, H. M. The influence of motivational factors on the frequency of participation in citizen science activities. *Nature Conservation* **18**, 61 (2017).

46. Isaac, N. J. & Pocock, M. J. Bias and information in biological records. *Biological Journal of the Linnean Society* **115**, 522–531 (2015).

47. Angulo, E. & Courchamp, F. Rare species are valued big time. *PloS one* **4**, e5215 (2009).

48. Booth, J. E., Gaston, K. J., Evans, K. L. & Armsworth, P. R. The value of species rarity in biodiversity recreation: A birdwatching example. *Biological Conservation* **144**, 2728–2732 (2011).

49. Rowley, J. J. *et al.* FrogID: Citizen scientists provide validated biodiversity data on frogs of australia. *Herpetological Conservation and Biology* **14**, 155–70 (2019).

50. Boakes, E. H. *et al.* Patterns of contribution to citizen science biodiversity projects increase understanding of volunteers recording behaviour. *Scientific reports* **6**, 33051 (2016).

51. Garrard, G. E., McCarthy, M. A., Williams, N. S., Bekessy, S. A. & Wintle, B. A. A general model of detectability using species traits. *Methods in Ecology and Evolution* **4**, 45–52 (2013).

52. Denis, T. *et al.* Biological traits, rather than environment, shape detection curves of large vertebrates in neotropical rainforests. *Ecological Applications* **27**, 1564–1577 (2017).

53. Sólymos, P., Matsuoka, S. M., Stralberg, D., Barker, N. K. & Bayne, E. M. Phylogeny and species traits predict bird detectability. *Ecography* **41**, 1595–1603 (2018).

54. Wood, C., Sullivan, B., Iliff, M., Fink, D. & Kelling, S. eBird: Engaging birders in science and conservation. *PLoS Biol* **9**, e1001220 (2011).

55. GBIF.org (3rd December 2019). GBIF occurrence download. https://www.doi.org/10.15468/dl.lpwczr

56. Gilfedder, M. *et al.* Brokering trust in citizen science. *Society & natural resources* **32**, 292–302 (2019).

57. Callaghan, C., Lyons, M., Martin, J., Major, R. & Kingsford, R. Assessing the reliability of avian biodiversity measures of urban greenspaces using eBird citizen science data. *Avian Conservation and Ecology* **12**, (2017).

58. Johnston, A. *et al.* Best practices for making reliable inferences from citizen science data: Case study using eBird to estimate species distributions. *BioRxiv* 574392 (2019).

59. Myhrvold, N. P. *et al.* An amniote life-history database to perform comparative analyses with birds, mammals, and reptiles: Ecological archives E096-269. *Ecology* **96**, 3109–3109 (2015).

60. Dale, J., Dey, C. J., Delhey, K., Kempenaers, B. & Valcu, M. The effects of life history and sexual selection on male and female plumage colouration. *Nature* **527**, 367–370 (2015).

61. R Core Team. *R: A language and environment for statistical computing*. (R Foundation for Statistical Computing, 2020).

62. Wickham, H. *et al.* Welcome to the tidyverse. *Journal of Open Source Software* **4**, 1686 (2019).

63. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* **67**, 1–48 (2015).

64. Kuznetsova, A., Brockhoff, P. B. & Christensen, R. H. B. lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software* **82**, 1–26 (2017).

65. Johnston, A. *et al.* Species traits explain variation in detectability of UK birds. *Bird Study* **61**, 340–350 (2014).

66. Steen, V. A., Elphick, C. S. & Tingley, M. W. An evaluation of stringent filtering to improve species distribution models from citizen science data. *Diversity and Distributions* **25**, 1857–1869 (2019).

67. Henckel, L., Bradter, U., Jönsson, M., Isaac, N. J. & Snäll, T. Assessing the usefulness of citizen science data for habitat suitability modelling: Opportunistic reporting versus sampling based on a systematic protocol. *Diversity and Distributions* **26**, 1276–1290 (2020).

68. Caley, P., Welvaert, M. & Barry, S. C. Crowd surveillance: Estimating citizen science reporting probabilities for insects of biosecurity concern. *Journal of Pest Science* **93**, 543–550 (2020).

69. Périquet, S., Roxburgh, L., Roux, A. le & Collinson, W. J. Testing the value of citizen science for roadkill studies: A case study from south africa. *Frontiers in Ecology and Evolution* **6**, 15 (2018).

70. Nakagawa, S. & Freckleton, R. P. Model averaging, missing data and multiple imputation: A case study for behavioural ecology. *Behavioral Ecology and Sociobiology* **65**, 103–116 (2011).

71. Schlossberg, S., Chase, M. & Griffin, C. Using species traits to predict detectability of animals on aerial surveys. *Ecological Applications* **28**, 106–118 (2018).

72. Aristeidou, M., Scanlon, E. & Sharples, M. Profiles of engagement in online communities of citizen science participation. *Computers in Human Behavior* 74, 246–256 (2017).

73. Troscianko, J., Skelhorn, J. & Stevens, M. Quantifying camouflage: How to predict detectability from appearance. *BMC evolutionary biology* **17**, 1–13 (2017).

74. Schuetz, J. G. & Johnston, A. Characterizing the cultural niches of North American birds. *Proceedings of the National Academy of Sciences* 22, 10868–10873 (2019).

75. Lišková, S. & Frynta, D. What determines bird beauty in human eyes? *Anthrozoös* 26, 27–41 (2013).

76. Tulloch, A. I., Possingham, H. P., Joseph, L. N., Szabo, J. & Martin, T. G. Realising the full potential of citizen science monitoring programs. *Biological Conservation* **165**, 128–138 (2013).

77. Kobori, H. *et al.* Citizen science: A new approach to advance ecology, education, and conservation. *Ecological research* **31**, 1–19 (2016).

78. Callaghan, C. T., Poore, A. G., Major, R. E., Rowley, J. J. & Cornwell, W. K. Optimizing future biodiversity sampling by citizen scientists. *Proceedings of the Royal Society B* **286**, 20191487 (2019).

79. Pacifici, K. *et al.* Integrating multiple data sources in species distribution modeling: A framework for data fusion. *Ecology* **98**, 840–850 (2017).

80. Robinson, O. J. *et al.* Integrating citizen science data with expert surveys increases accuracy and spatial extent of species distribution models. *Diversity and Distributions* **26**, 976–986 (2020).

81. Strien, A. J. van, Termaat, T., Groenendijk, D., Mensing, V. & Kery, M. Site-occupancy models may offer new opportunities for dragonfly monitoring based on daily species lists. *Basic and Applied Ecology* **11**, 495–503 (2010).

82. Van der Wal, R. *et al.* Mapping species distributions: A comparison of skilled naturalist and lay citizen science recording. *Ambio* **44**, 584–600 (2015).

83. Dennis, E. B., Morgan, B. J., Brereton, T. M., Roy, D. B. & Fox, R. Using citizen science butterfly counts to predict species population trends. *Conservation biology* **31**, 1350–1361 (2017).

84. Stoudt, S., Goldstein, B. R. & De Valpine, P. Identifying charismatic bird species and traits with community science data. bioRxiv. https://doi.org/10.1101/2021.06.05.446577

FIGURES

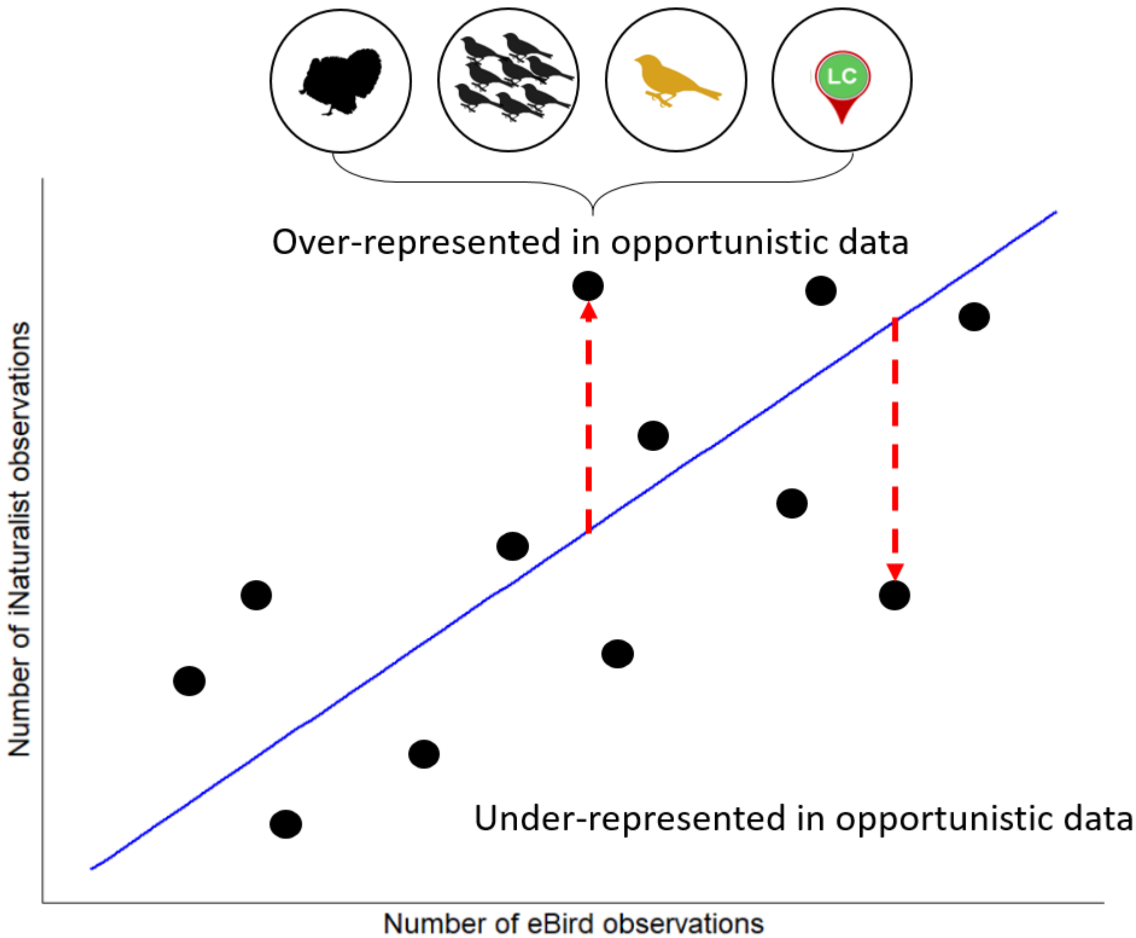

Figure 1. A conceptual figure depicting the methods used in our analysis. We used the residual from the relationship between the number of eBird observations (i.e., semi-structured citizen science observations) and iNaturalist observations (i.e., unstructured citizen science observations) to quantify the over- or under-representation of a species in unstructured citizen science data. We predicted that species which were over-represented in unstructured iNaturalist data would be larger in size, occur more frequently in large flocks, be brighter in color, and be categorized as Least Concern IUCN status (a proxy for commonness).

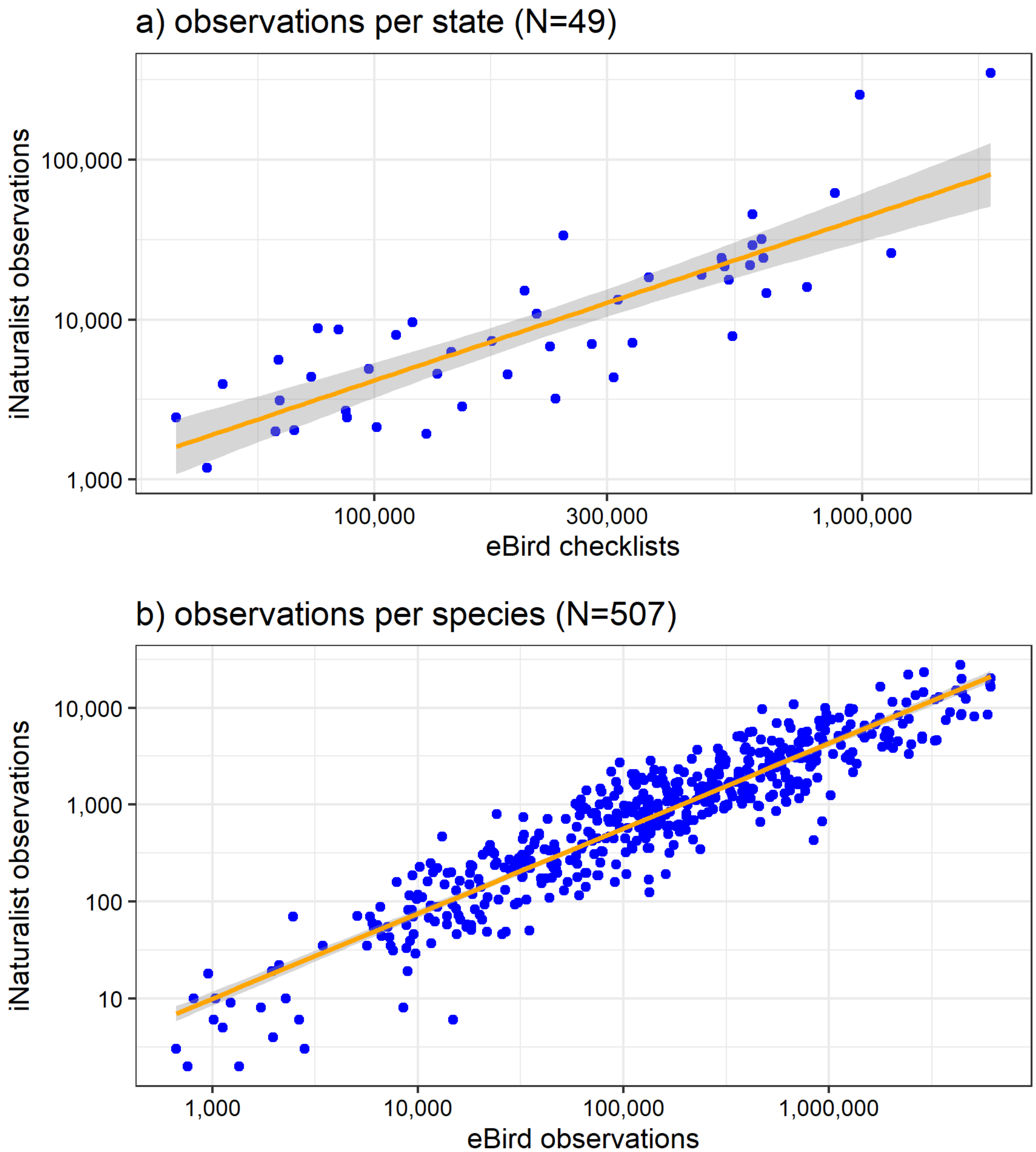

Figure 2. a) The relationship between the total number of eBird checklists and total number of iNaturalist observations for 49 states, including the District of Columbia. There was strong evidence that these variables were correlated (R^2=0.58, p-value <0.001) suggesting that sampling between datasets is correlated among states. b) The relationship between the number of observations for a species from eBird (x-axis) and the number of observations for a species from iNaturalist (y-axis) for only eBird species which were found on >1% of eBird checklists.

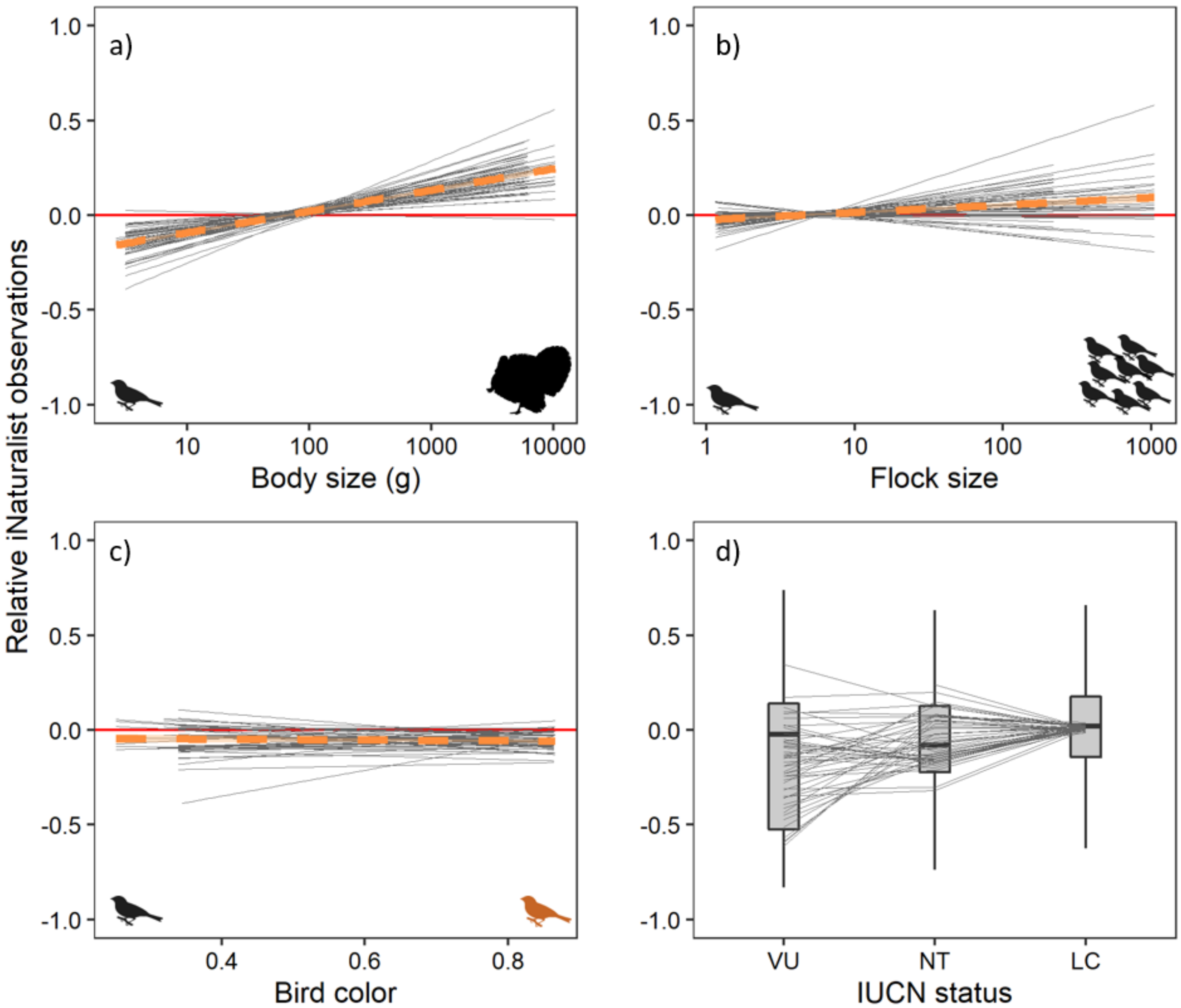

Figure 3. The relationship between a) body size of a species, b) flock size, c) color and d) commonness and the residuals of a linear model fit between iNaturalist and eBird observations (see Figure 1). These results demonstrate that there is a strong bias of body size in iNaturalist compared with eBird. Positive values on the y-axis mean over-represented in iNaturalist and negative values on the y-axis mean under-represented in iNaturalist. Body size and flock size are represented on a log10 scale. Each line represents a state (N=49). For a-c), the overall relationship pooling states is represented by the orange fitted line and 95% confidence interval.

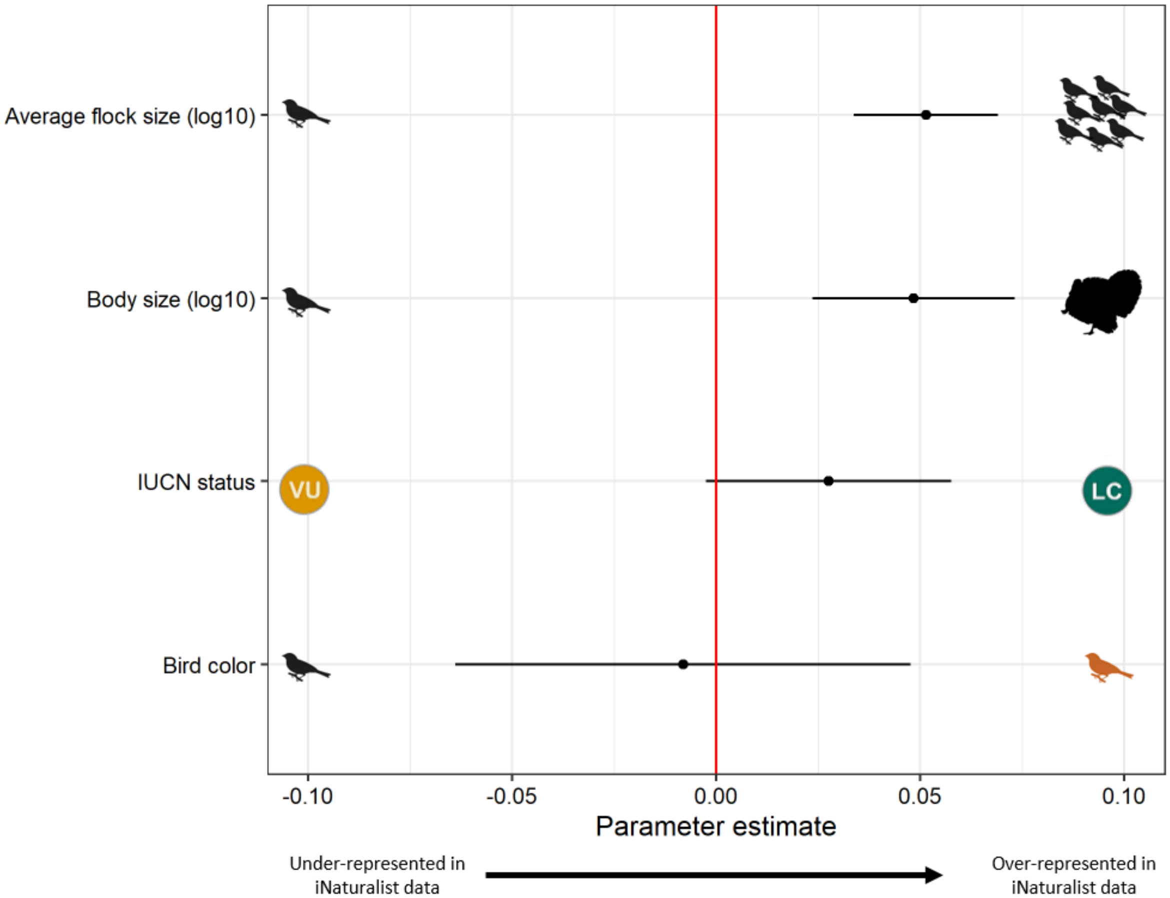

Figure 4. Results of a linear mixed effect model where all four variables were considered simultaneously, and state was a random effect. Strong support was found for body size and flock size (their 95% confidence interval does not overlap 0), whereas moderate support was found for IUCN status, and no support was found for color.

SUPPLEMENTARY FIGURES

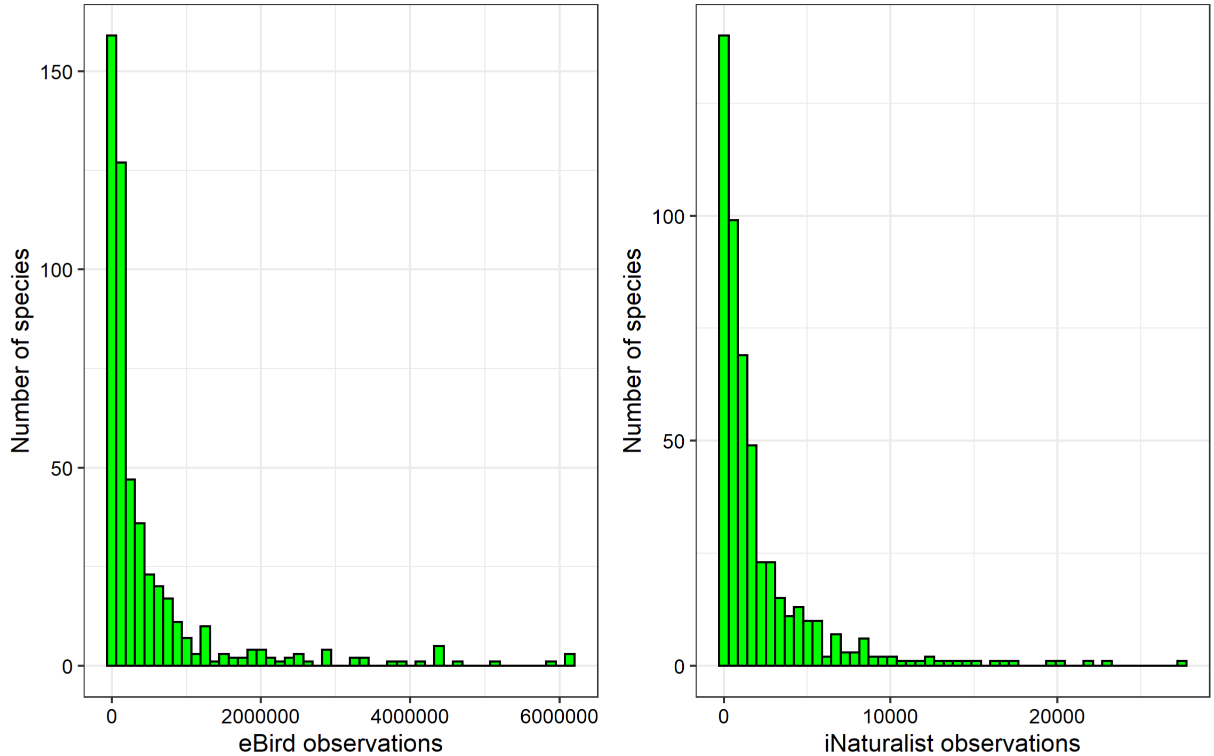

Figure S1. Histograms of the number of observations for a species from both eBird and iNaturalist citizen science projects.

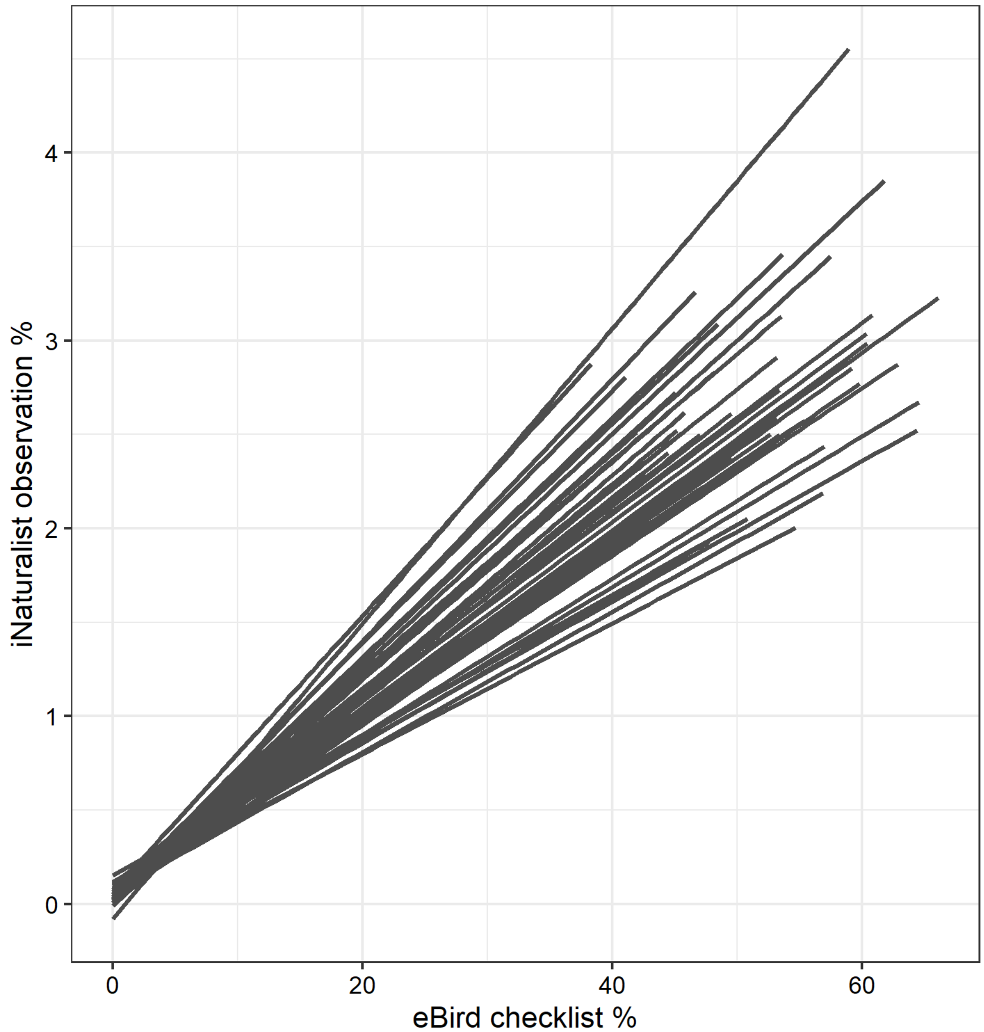

Figure S2. Among states (each line represents a state; N=49) we found that the percent of eBird checklists a species was found on and the percent of all iNaturalist observations a species comprised was strongly correlated.

Table S1. Uploaded separately. Raw data used for modelling, including the residual difference between iNaturalist and eBird, stratified by state.

Table S2. Results of single regression models, where each trait was treated separately, and consequently had different sample sizes in the model fit. Each model was fit with the residuals used as the response variable, the specific trait as the predictor variable, where body size and flock size were log10-transformed and IUCN was treated as an ordinal variable, and state was a random effect. This analysis was performed to confirm the results of the multiple regression mixed effects analysis presented in the main results (Figure 4).

	estimate	t	p-value	Number of obs
Body size	0.11	31.59	<0.001	7743
Color	-0.01	-0.413	0.68	4542
Flock size	0.033	6.118	<0.001	8702
IUCN status	0.078	7.73	<0.001	7629