IBIS
international journal of avian science

Ibis (2024) doi: 10.1111/ibi.13343

Estimating sampling biases in citizen science datasets

LOUIS J. BACKSTROM,*[1,2,3,4] COREY T. CALLAGHAN,[5] HANNAH WORTHINGTON,[3,4]
RICHARD A. FULLER[1,2] & ALISON JOHNSTON[3,4]
[1] *Centre for Biodiversity and Conservation Science, The University of Queensland, Brisbane, Queensland, Australia*
[2] *School of the Environment, The University of Queensland, Brisbane, Queensland, Australia*
[3] *Centre for Research into Ecological and Environmental Modelling, University of St Andrews, St Andrews, UK*
[4] *School of Mathematics and Statistics, University of St Andrews, St Andrews, UK*
[5] *Department of Wildlife Ecology and Conservation, Fort Lauderdale Research and Education Center, University of Florida, Davie, Florida, 33314-7719, USA*

The rise of citizen science (also called community science) has led to vast quantities of species observation data collected by members of the public. Citizen science data tend to be unevenly distributed across space and time, but the treatment of sampling bias varies between studies, and interactions between different biases are often overlooked. We present a method for conceptualizing and estimating spatial and temporal sampling biases, and interactions between them. We use this method to estimate sampling biases in an example ornithological citizen science dataset from eBird in Brisbane City, Australia. We then explore the effects of these sampling biases on subsequent model inference of population trends, using both a simulation study and an application of the same trend models to the Brisbane eBird dataset. We find varying levels of sampling bias in the Brisbane eBird dataset across temporal and spatial scales, and evidence for interactions between biases. Several of the sampling biases we identified differ from those described in the literature for other datasets, with protected areas being undersampled in the city, and only limited seasonal sampling bias. We demonstrate variable performance of trend models under different sampling bias scenarios, with more complex biases being associated with typically poorer trend estimates. Sampling biases are important to consider when analysing ecological datasets, and analysts can use this method to ensure that any biologically relevant sampling biases are detected and given due consideration during analysis. With appropriate model specification, the effects of sampling biases can be reduced to yield reliable information about biodiversity.

Keywords: community science, eBird, population trends, spatial–temporal bias.

Effective monitoring of biodiversity is critical in addressing the global biodiversity crisis. One increasingly popular approach to biodiversity monitoring is citizen science, which leverages data collected by volunteers across broad spatial, temporal and taxonomic extents. Citizen science data have already advanced our understanding of numerous important ecological questions and there is potential for many further contributions (Dickinson *et al.* 2010, Brown & Williams 2019). However, concerns over the quality of data produced by citizen science projects remain a significant barrier to their use in science (Burgess *et al.* 2017). One class of data quality challenges are sampling biases, which can negatively affect inferences drawn from citizen science data, including species distribution models (e.g. Steen *et al.* 2019, Johnston *et al.* 2020), analyses of population trends (e.g. Boersch-Supan *et al.* 2019, Fink *et al.* 2023) and phenological studies (e.g. Courter *et al.* 2013, Steiner *et al.* 2022). In this study, we focus on the

*Corresponding author.
Email: ljb38@st-andrews.ac.uk
Twitter: @BackstromLouis

© 2024 The Author(s). *Ibis* published by John Wiley & Sons Ltd on behalf of British Ornithologists' Union.
This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

effect of sampling biases on population trend models, but also discuss their effect on model inference more broadly.

Ecological citizen science projects usually collect observations of wild organisms. These data contain four fundamental pieces of information: what was observed, where was it observed, when was it observed and by whom (Isaac & Pocock 2015). Citizen science data can therefore be considered across three dimensions: spatial (where), temporal (when) and taxonomic (what), and with sampling across each of these dimensions influenced by observer behaviour (whom) (August *et al.* 2020, Di Cecco *et al.* 2021, Johnston *et al.* 2023). Data volumes tend to increase with decreasing project structure, meaning that less-structured citizen science projects (e.g. iNaturalist, eBird) can receive large amounts of data with the caveat that observations are more likely to be heterogeneously sampled across space, time and taxonomy (Isaac *et al.* 2014, Kelling *et al.* 2019), an issue referred to as 'sampling bias'. Citizen science datasets therefore tend to be a non-probability sample of biodiversity (Boyd *et al.* 2023), rather than a representative sample in which the data have the same characteristics as a given target population, so they are best suited for statistical inference (Boyd *et al.* 2024). Nevertheless, 'sampling bias' need not be interpreted negatively, nor inherently avoided (Boakes *et al.* 2016). Rather, the effect of any given sampling bias is contingent on its role in obscuring the answer to a particular research question (Callaghan *et al.* 2019).

Different types of sampling bias have differing effects on data analysis. Examples of spatial sampling biases include a disproportionate amount of data collected near roads (Petersen *et al.* 2021) and major settlements (Botts *et al.* 2011). Estimates of spatially varying ecological processes are particularly susceptible to impacts of spatial biases (Johnston *et al.* 2020). Temporal sampling biases are many and varied, and include the rapid rise in the rate of observations submitted to citizen science projects over the last decades (Isaac & Pocock 2015) and the well-documented increase in sampling effort at the weekend compared with weekdays (Courter *et al.* 2013). Temporal biases typically affect temporally oriented applications of observation data, such as population trend analyses (Horns *et al.* 2018) or studies of phenology (Steiner *et al.* 2022). Taxonomic sampling biases emerge when certain taxa are preferentially recorded relative to others, either because of observer behaviours or differences in detectability among taxa (Ward 2014, Johnston *et al.* 2023). These biases frequently impact applications at the community level, where interactions or comparisons between species are of most interest (Buckland & Johnston 2017). In this paper, we focus on estimating spatial and temporal sampling biases and leave estimation of taxonomic biases as an area for future research.

Sampling biases also exist across scales within each of the three dimensions and can interact with each other. For example, temporal heterogeneity in sampling effort (the cumulative sampling contributions of all observers) may manifest hourly (e.g. a preference by users to submit bird sightings in the morning; Kelling *et al.* 2015) or seasonally (e.g. more sampling effort during migration; La Sorte & Somveille 2019). Interactions between sampling biases may occur across multiple scales within the same dimension (e.g. seasonal patterns in the magnitude and timing of daily recording peaks – a temporal–temporal interaction), or between multiple dimensions (e.g. seasonal changes in the spatial distribution of sampling effort – a temporal–spatial interaction). Interactions between different dimensions of sampling bias are an increasingly studied phenomenon (e.g. Meyer *et al.* 2016, Pescott *et al.* 2019, Bowler *et al.* 2022), but the effects of interacting sampling biases on model inference are still not fully understood. Additionally, current approaches to estimating biases vary, with no clear best practice adopted in the literature. Previous studies have tended to either not estimate sampling biases (with authors assuming any biases to be adequately controlled during modelling), or have done so in various ways, for example by modelling the effect of spatial or temporal variables on sampling effort (Shirey *et al.* 2021, Tang *et al.* 2021, Bowler *et al.* 2022), or by simple enumeration of the number of observations across space or time (Hughes *et al.* 2021, Ver Hoef *et al.* 2021). If sampling biases are not properly explored and accounted for, analysts risk producing unreliable results.

The solutions available for alleviating problematic sampling biases will depend on the specific goals of a study, and several solutions have been proposed at both the pre-analysis and analysis stages. For pre-analysis assessment of sampling biases, one recent approach (Boyd *et al.* 2022) has

1474919x, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/ibi.13343 by University Of St Andrews University, Wiley Online Library on [16/07/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

been to provide a structured tool for analysts to gauge the 'Risk-Of-Bias' in their chosen dataset(s) across ecologically relevant dimensions and scales. Another approach (Schmill *et al.* 2014) uses a set of statistical tools to assess sampling bias. This approach is similar to the method we present in this paper but differs in being (1) more mathematically complex than our method and (2) only designed to address single sampling biases, rather than multiple interacting biases simultaneously. For analytical mitigation of sampling biases, various modelling approaches have been developed that attempt to either implicitly or explicitly control for sampling biases (see, for example, Boyd *et al.* 2024). For example, the Double Machine Learning trend model of Fink *et al.* (2023) reduces the effect of interacting sampling biases on model inference of population trends by explicitly modelling the changing propensity of samples to be from different environments in different years.

Here, we first present a novel method for estimating spatial and temporal sampling biases, and interactions between the two, in citizen science datasets. Second, we develop simulations to demonstrate the effect of interacting sampling biases on model inference, using habitat-specific models of population trends as an example of an ecological metric of interest. Finally, we apply the same procedures from the simulation study to real-world data from the Brisbane eBird dataset to illustrate the importance of considering interacting sampling biases in estimating population trends in birds.

METHODS

Datasets and study area

eBird is the world's largest contributory citizen science project (Chandler *et al.* 2017); as of January 2024, the dataset contains 1.53 billion bird occurrence records from 116 million checklists (a list of birds observed on a single visit to a site). eBird protocols allow for checklists to be denoted as either 'complete' or 'incomplete', dependent on whether or not the list contains all species identified by the observer during that observation period. Most of the eBird dataset comprises complete checklists and can therefore be interpreted in a standard detection–nondetection format. Such checklist-style sampling protocols are common in both citizen science and professional monitoring schemes, particularly for birds and butterflies, and the methods presented here are therefore broadly applicable to a range of datasets.

We chose Brisbane City (Queensland, Australia) as the study area for the analyses of sampling biases and population trends (a map is provided in Supporting Online Information Fig. S1). We chose Brisbane as the city has a large birding community, and more than 100 000 eBird checklists have been recorded across the city; we also expected *a priori* that there would be some level of sampling bias present across various spatial and temporal scales, but did not make any explicit predictions about any of these. We classified broad habitat types for mainland Brisbane City (i.e. excluding islands and marine areas) using several spatial datasets from the Queensland Government. We first used the Regional Ecosystems dataset (Queensland Herbarium 2022) to classify the city into five distinct habitat types (Non-Remnant, Dry Forest, Wet Forest, Wetland and Estuary; see Supporting Online Information Table S1). We identified additional Wetland and Estuary areas using the Wetland Areas dataset (Queensland Herbarium 2019), and added a sixth habitat type, Built Up, by superimposing areas from the Built Up Areas dataset (Queensland Department of Resources 2022). We derived Protected Areas from the Collaborative Australian Protected Areas Database (Commonwealth of Australia 2021), and Elevation from a 1-second Digital Elevation Model (Gallant *et al.* 2011).

We explored sampling biases and population trends within the eBird Basic Dataset (EBD, which contains individual records of species observations) and the eBird Sampling Dataset (ESD, which contains overall records of sampling events – checklists). We imported the December 2021 release of these two datasets (Cornell Lab of Ornithology 2021) into R 4.2.0 (R Core Team 2022), and filtered the data to the Brisbane area (Supporting Online Information Fig. S1). Additionally, we applied the following standard filters: complete checklists only, protocol either Stationary or Travelling, checklist durations of 5–300 min, distances travelled of 0–8 km, and year between 2012 and 2021 inclusive. We also removed duplicate copies of shared checklists (i.e. checklists from different people who were birdwatching together) and summarized all observations in the EBD at the species level. We combined the EBD and ESD to create a zero-filled detection–nondetection dataset for 157 species in the city with at least 1000 detections.

© 2024 The Author(s). *Ibis* published by John Wiley & Sons Ltd on behalf of British Ornithologists' Union.

1474919x, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/ibi.13343 by University Of St Andrews University, Wiley Online Library on [16/07/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Nondetections were inferred in complete checklists in which the species was not reported. After filtering, there were 61 417 checklists in the Brisbane eBird dataset, a 39% reduction from the full dataset, mostly due to the removal of incomplete checklists and duplicate copies of shared checklists. Spatial statistics were calculated within a 200-m buffer around the point location associated with each checklist. We found that our results were not sensitive to changing the buffer size.

Estimating sampling biases

We estimated sampling biases by comparing the observed distribution of sampling effort across a dimension (for single biases) or dimensions (for interactions between two biases) with the distribution expected under a null hypothesis of representative sampling. We enumerated sampling effort simply by adding up the number of eBird checklists; we explored the effect of incorporating other measures of survey effort (e.g. duration and distance travelled) but found very little difference in our results. By comparing the null and observed distributions of sampling effort, a metric of sampling bias can be derived, representing the proportion of the observed distribution that would need to be redistributed for the data to conform to the expected distribution (see below). Sampling biases were explored at multiple scales in two dimensions: temporal and spatial, with temporal–temporal, spatial–spatial and temporal–spatial interactions also explored.

This method of estimating sampling biases makes an explicit assumption that data are 'Missing at Random', with all variables affecting whether data are included in a sample or not (i.e. the sampling biases) being known and accounted for (Bhaskaran & Smeeth 2014). This assumption simplifies the modelling process but is rarely met in practical scenarios, because not all variables influencing sampling are known or measurable (i.e. data are Missing not at Random; Boyd *et al.* 2024). Here, sampling bias refers primarily to the discrepancies of the variable of interest (i.e. sampling) and not merely to the spatiotemporal variables used as proxies in our analysis. Although we do employ a range of variables (e.g. habitat classes) to approximate and understand the sampling dynamics, the true focus of sampling bias remains on the sampling process itself. Furthermore, as we increase the number of relevant covariates, we will describe more of the variation in sampling bias, and the impact of this violation will be reduced.

Estimating sampling biases with the Hoover Index

Our method of estimating sampling bias is a modification of the Hoover Index (Hoover 1941), which was first described in economic demography as a means of describing inequality in a dataset. The metric ranges from zero to one (theoretically), with lower values corresponding to lower inequality (in our case, lower sampling bias). Put simply, it describes the deviation of a given discrete probability distribution from the uniform distribution, and is defined by the following function:

$$H = \frac{\sum_{i=1}^{n} |x_i - \overline{x}|}{2\sum_{i=1}^{n} x_i}$$

where i is the discrete category within a set of n distinct levels (e.g. habitat types), x_i is the number of samples assigned to category i, and $\overline{x}$ is the average number of samples per category. H is bounded between 0 (perfect equality, i.e. a uniform distribution of samples across categories) and an upper bound defined by $1 - \frac{1}{n}$ (maximum inequality, i.e. all samples within one category). This metric describes the proportion of data (in our case, checklists) that would need to be redistributed for the data to fit a uniform distribution across categories.

In many instances an 'unbiased' (or representative) dataset would not be expected to have a uniform distribution. For example, the habitat types in Brisbane do not have equal areas, as different habitats cover different proportions of land in the city. If sampling were evenly distributed across the landscape, one would not expect a uniform distribution of sampling frequencies across habitats but rather frequencies corresponding to the relative areas of the different habitat types. Therefore, we have adapted the Hoover Index to enable the expected distribution to be an uneven distribution of sampling across categories. Our modified Hoover Index has the following form:

$$H' = \frac{\sum_{i=1}^{n} |x_i - y_i|}{2\sum_{i=1}^{n} x_i}$$

where y_i is the number of samples within category i that would be expected under unbiased or representative sampling. In this case, H' is bounded by

© 2024 The Author(s). *Ibis* published by John Wiley & Sons Ltd on behalf of British Ornithologists' Union.

1474919x, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/ibi.13343 by University Of St Andrews University, Wiley Online Library on [16/07/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

0 (the sample proportions align perfectly with the expected proportions) and 1 (the sample proportions do not at all align with the expected proportions, i.e. all samples belong to categories where the expected sampling frequency is zero).

Interactions between sampling biases across two dimensions can also be calculated. Because our modified Hoover Index requires that $\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$, then when dealing with interactions across two dimensions one must first account for the bias present in one dimension when producing the expected distribution of the other dimension. This is the same calculation that is conducted to estimate expected values for a χ^2 contingency table under the assumption of independence between the two dimensions of sampling bias. For example, if eBird checklists in Brisbane were evenly distributed across both habitats (spatial) and years (temporal), one would expect an equal number of checklists in a given habitat across all years. As an illustration, wetlands cover 6% of land in Brisbane, we have data from 10 full years and there are 61 417 checklists, so we would expect 368 checklists in wetlands in each year under a null hypothesis of even sampling across years. However, as the number of checklists across years is not evenly distributed (there is temporal sampling bias), one must adjust the expected number of checklists in each habitat to reflect this temporal sampling bias while retaining the appropriate expectation for the proportion of checklists for each habitat (e.g. wetlands should have 6% of checklists in each year, so 22 checklists in 2012, 32 in 2013, and so on, rising to 735 in 2022). The sum of expected checklists across habitats in a given year is equal to the observed total number of checklists (across all habitats) in that year. The same logic applies for other interactions (whether temporal–temporal, spatial–spatial or temporal–spatial). Mathematically, the Hoover Index (and its modifications) could be extended to continuous probability distributions, but for simplicity all biases presented here are explored along discrete axes.

We constructed confidence intervals for our estimates of sampling bias using a nonparametric bootstrap on the checklist locations and recalculating the modified Hoover Index with each bootstrap resample. We repeated this 1000 times, and the limits of the 95% confidence intervals were defined as the 2.5% and 97.5% centiles of the bootstrap distribution of modified Hoover Indices. We also tested the significance of the deviation from expected number of checklists for each individual category. This was done by calculating the number of checklists observed and expected both within and outside the category (or interaction of categories across two dimensions). We then used a binomial distribution to test (in isolation) whether each observed proportion was significantly different from the null hypothesis of expected proportion under unbiased sampling.

Assessing the impacts of sampling biases in simulations

We conducted a series of simulations to explore the effects of sampling biases on model outcomes. In these simulations, species exist at different occupancy rates and have different trends in occupancy in each of two habitats. We simulated sampling bias across habitats (each habitat is not sampled in proportion to its area) and additionally applied temporal trends that altered the sampling bias in each habitat over time. The goal was to assess the impact of sampling bias across time (temporal bias), habitats (spatial bias) and changes over time (spatial–temporal interactions), on our inference about trends in species occupancy (temporal ecological patterns). We chose to explore trends at the per-habitat level rather than at an aggregate (e.g. landscape-wide) level, as accurate habitat-specific trends are a prerequisite for accurate aggregate trends when using data that are spatially biased with respect to habitat.

We created a simple virtual environment, comprising two spatially implicit habitats (referred to as 'yellow' and 'blue') of equal area, and one species that initially occupied the habitats at occupancy rates of 70% and 30%, respectively. For simplicity, detection probability was fixed at 100%. Each simulation spanned 10 years, with checklists simulated for 365 days in each year ($1 \leq t \leq 3650$) and no intra-annual patterns in occupancy. We simulated six scenarios of sampling bias to explore situations that might occur in real-world data (Table 1). To reflect the common increase in citizen science data, for all scenarios we simulated a linear increase in sampling effort over time (temporal bias), such that year 1 had 1000 expected checklists and every year thereafter an additional 1000. For each scenario, the 55 000 checklists per simulation ($1000 + 2000 + \ldots + 10\,000$) were randomly allocated to a date and habitat according to

© 2024 The Author(s). Ibis published by John Wiley & Sons Ltd on behalf of British Ornithologists' Union.

1474919x, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/ibi.13343 by University Of St Andrews University, Wiley Online Library on [16/07/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 1. Simulation parameters for each of the six scenarios of sampling bias tested.

			Change in occupancy (r)		Sampling rate at $t = 1$		Sampling rate at $t = 3650$	
Scenario number	Occupancy description	Sampling description	Yellow	Blue	Yellow	Blue	Yellow	Blue
1	Stable	Even No change over time	0.000	0.000	0.50	0.50	0.50	0.50
2	Decreasing blue	Even No change over time	0.000	−0.001	0.50	0.50	0.50	0.50
3	Stable	Biased No change over time	0.000	0.000	0.25	0.75	0.25	0.75
4	Stable	Biased Change from blue to yellow	0.000	0.000	0.25	0.75	0.75	0.25
5	Decreasing blue	Biased No change over time	0.000	−0.001	0.25	0.75	0.25	0.75
6	Decreasing blue	Biased Change from blue to yellow	0.000	−0.001	0.25	0.75	0.75	0.25

Two habitats (yellow/blue) were simulated, each with a specified sampling rate (probability of a checklist being allocated to that habitat), which changed linearly throughout the simulation period for two scenarios (4 and 6). Sampling rates are provided for the start ($t = 1$) and end ($t = 3650$) of the simulation period and may be inferred for any intermediate point in time. The occupancy rates of each habitat changed through time according to the exponential decay function $\psi(t) = \psi_1(1+r)^t; -1 < r < 0$ where $\psi(t)$ is the occupancy of the habitat at time t, ψ_1 is the initial occupancy of the habitat and r is the rate of decline.

weightings specified by the temporal bias, spatial bias and spatiotemporal bias interaction in that scenario (Table 1). For each checklist, the occurrence of the species (present or absent) was determined by randomly generating a presence or absence from a Bernoulli distribution with a probability equal to the occupancy rate for that habitat at that point in time. Occurrences within a given habitat were uncorrelated and independent within a simulation run. For example, a site with constant occupancy rate of 0.5 would, with no temporal correlation, have on average half of the visits recording species present and half recording species absent. We simulated population declines using the exponential decay function $\psi(t) = \psi_1(1+r)^t$, where $\psi(t)$ is the occupancy rate of the habitat at time t, ψ_1 is the initial occupancy rate, and r is the rate of population change, which in all our scenarios was a decline $(-1 < r < 0)$.

For each simulation, we randomly split the dataset into two equal halves for use in model fitting and predicting (we chose an even split because our simulated datasets were not size-limited). We fit three Generalized Linear Models to one half, with presence/absence as the response variable, using a binomial error distribution with a logit link function. We could run these simple models instead of occupancy models, as we assumed that detectability was perfect in this simulation. The three models were: (1) a basic model, with only time as an explanatory variable, (2) an additive model, with time and habitat as explanatory variables, and (3) an interactive model, with time and habitat as explanatory variables, and an interaction term fit between them. We then used each model to predict occupancy values for the other half of the data. We ran each of the six scenarios 100 times and summarized the results of each.

Assessing the impacts of sampling biases in real-world data

We applied the same modelling procedure above to real-world data, using the Brisbane City eBird dataset. Each checklist in the dataset was allocated to one of the six habitat types defined above (Datasets and Study Area). As in the simulation study, three different Generalized Linear Models of varying complexity (basic, additive and interactive) were fit to the data for each of the 157 species with at least 1000 detections. These models were then used to estimate the reporting rate of each species in each habitat type over the study period; reporting rate is the proportion of complete checklists that report a species, which is a product of species occupancy and species detectability. The overall trend in species reporting rate

© 2024 The Author(s). *Ibis* published by John Wiley & Sons Ltd on behalf of British Ornithologists' Union.

1474919x, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/ibi.13343 by University Of St Andrews University, Wiley Online Library on [16/07/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

in each habitat from the beginning to the end of the study period was determined for each model, allowing for comparisons between models. As above, we explored trends at the per-habitat level, as accurate habitat-specific trends are a prerequisite for accurate aggregate trends when analysing spatially biased data. Unlike the simulation study, where the truth is defined *a priori*, the real reporting rate at any point in time for any given species is unknown and hence the relevant quantities of interest for this study were the differences between the three models, rather than any estimates of model accuracy relative to the 'truth'. In using this model construction, we assumed that detectability does not vary substantially over the spatial or temporal extents of our data (Newson *et al.* 2013), and that changes to reporting rate are a consequence of changes in the ecological occupancy rate over the survey period (rather than detectability). Based on our knowledge of the Brisbane system and the spatial and temporal data extents, these are reasonable assumptions for most species, but we acknowledge that this is unlikely to be true for all species in the dataset. Furthermore, for other applications, detectability may change substantially, especially over seasonal, or longer, time frames, which could affect the observation process of both citizen science data and structured surveys. We assumed perfect detectability as it allowed us to explore basic differences in results across our simple models under various scenarios of sampling bias. In reality, however, detectability may also vary across some of these axes of bias (e.g. time or habitat), which may lead to even greater impacts of these biases on estimated population trends. Therefore, if the objective of a study is to make inferences about bird population trends, we recommend estimating detectability and its variation across time and space.

RESULTS

Estimating sampling biases

Sampling biases and their interactions in the Brisbane eBird dataset were explored at multiple scales across both temporal and spatial dimensions using the modified Hoover Index (H') as described above. For brevity, only the results for yearly (temporal) and per-habitat (spatial) biases are presented here (Fig. 1); other scales are presented in Figures S4–S11 in the Supporting Online Information. At the yearly scale, the rapid uptake of eBird in the city is evident in the dramatic increase in the number of checklists over the study period, leading to a nonuniform distribution throughout time and a large estimate of sampling bias (Fig. 1a). In the spatial dimension, sampling rates varied across habitats, with a moderate estimate of sampling bias. Non-Remnant, Estuary and Wetland

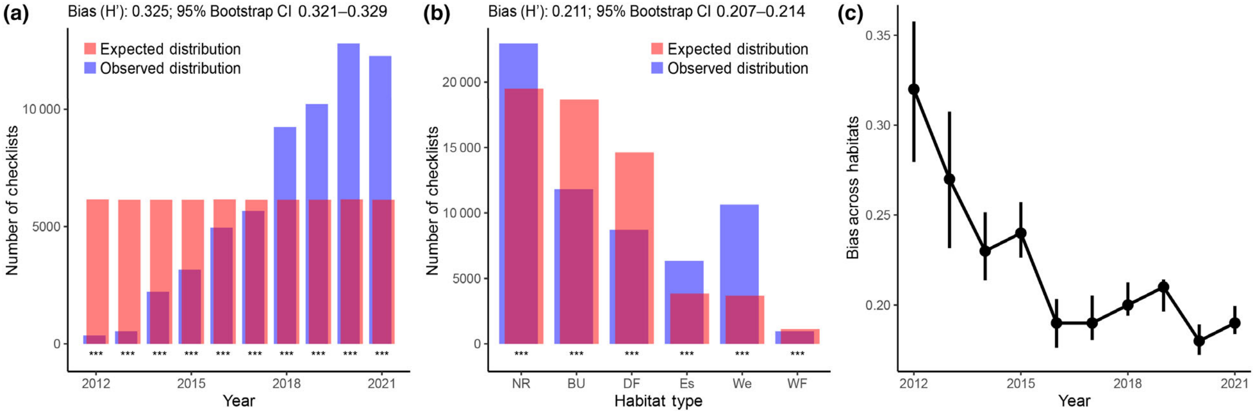


Figure 1. Key sampling biases in the Brisbane eBird Dataset. Biases and their interactions were explored at multiple scales across temporal and spatial dimensions using the modified Hoover Index. Figures are presented here for two biases (see Supporting Online Information for other scales): (a) yearly – a temporal sampling bias, (b) per-habitat – a spatial sampling bias and (c) the interaction between the two. Translucent red indicates the null distribution under representative sampling, blue the observed distribution and dark pink the overlap between red and blue bars. Note that for both the temporal and spatial biases presented here (a and b), all categories (years or habitats) have a significantly different number of checklists from that expected under the null hypothesis (***$P < 0.001$). Error bars in the spatial–temporal interaction (c) denote the 95% bootstrap confidence interval for the spatial sampling bias in that year. Key to habitats: NR, Non-Remnant; BU, Built Up; DF, Dry Forest; Es, Estuary; We, Wetland; WF, Wet Forest.

© 2024 The Author(s). *Ibis* published by John Wiley & Sons Ltd on behalf of British Ornithologists' Union.

1474919x, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/ibi.13343 by University Of St Andrews University, Wiley Online Library on [16/07/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

habitats were significantly oversampled relative to their prevalence, whereas Built Up, Dry Forest and Wet Forest habitats were significantly undersampled (Fig. 1b). The degree of sampling bias across habitats has changed over time, with a decline between 2012 and 2016 and little change thereafter (Fig. 1c), suggestive of an interaction between the two biases and that spatial sampling bias has weakened over time.

Assessing the impacts of sampling biases in simulations

Three models of varying complexity were fit to six different scenarios of sampling bias (Fig. 2, Supporting Online Information Table S2). The interactive model was able to account for all temporal biases, spatial biases and spatiotemporal bias interactions that we simulated and was also able to accurately estimate occupancy on a per-habitat level. The additive model was also able to account for all the biases and interactions that we simulated, but only when occupancy was averaged across the two habitats, with results within individual habitats being inaccurate. The basic model consistently performed the worst out of the three, being only able to estimate occupancy at an average level, and only able to do so accurately in the absence of a spatial bias or interaction between biases. The lack of a habitat term in the basic model prevented any distinction between habitats and made it susceptible to the impact of spatial biases, leading to often incorrect inferences and lower accuracy than the other two models.

When occupancy was averaged across both habitats, all models were unsurprisingly robust to a temporal bias alone (Scenario 1). When a decline in occupancy was added (Scenario 2), only the interactive model accurately reflected the truth on a per-habitat level, although the basic and additive models did correctly identify an average decline. All models were also robust to a temporal bias and a spatial bias in conjunction, provided there was no interaction between the two, and no change in occupancy (Scenario 3). However, the basic model underestimated the average occupancy rate due to the undersampling of the habitat in which the species was common (yellow). When a spatiotemporal interaction was added (Scenario 4), the basic model incorrectly estimated an increase in average occupancy, whereas the other two models continued to accurately estimate the truth. When a temporal bias, spatial bias and decline in occupancy were simulated (Scenario 5), all models correctly identified the decline, although only the interactive model was accurate on a per-habitat level. Finally, when an interaction was added between the temporal and spatial biases alongside a decline in occupancy (Scenario 6), only the interactive model was accurate on a per-habitat level; the additive model only identified the average decline and the basic model incorrectly identified an average increase, due to a progressive oversampling of the habitat in which the species was common and not declining (yellow).

Assessing the impacts of sampling biases in real-world data

By applying the modelling methodology from the simulation study to eBird data, we found that the three models produced different trend estimates. We illustrate this with two example species, the Dusky Moorhen *Gallinula tenebrosa* and Crested Pigeon *Ocyphaps lophotes* (Fig. 3). Under the basic model (only time as a covariate), both species showed a significant negative trend over the study period. Similarly, under the additive model (time and habitat as covariates), all habitats continued to show a significant negative trend. However, when the interaction term was added, trends varied across habitats: for Dusky Moorhen, two habitats showed a significant negative trend, two a nonsignificant trend and two a significant positive trend; for Crested Pigeon, four habitats showed a significant negative trend, one a nonsignificant trend and one a significant positive trend. This suggests that these two species have different trends in some habitats, a pattern only identified by the interactive model. By detecting habitat-specific trends, we show that spatial sampling bias across habitats will sometimes lead to incorrect conclusions about the overall trend, when using a basic or additive model.

We then generalized these results to all 157 species modelled, summarizing per-habitat reporting rate trend estimates (Fig. 4). Generally speaking, inter-model agreement (i.e. consensus on the significance and sign of the trend; Fig. 4) was highest between basic and additive models and lower when these two models were compared with the interactive model (the interactive model was the only model that correctly identified occupancy trends in all of our simulated scenarios).

© 2024 The Author(s). *Ibis* published by John Wiley & Sons Ltd on behalf of British Ornithologists' Union.

1474919x, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/ibi.13343 by University Of St Andrews University, Wiley Online Library on [16/07/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

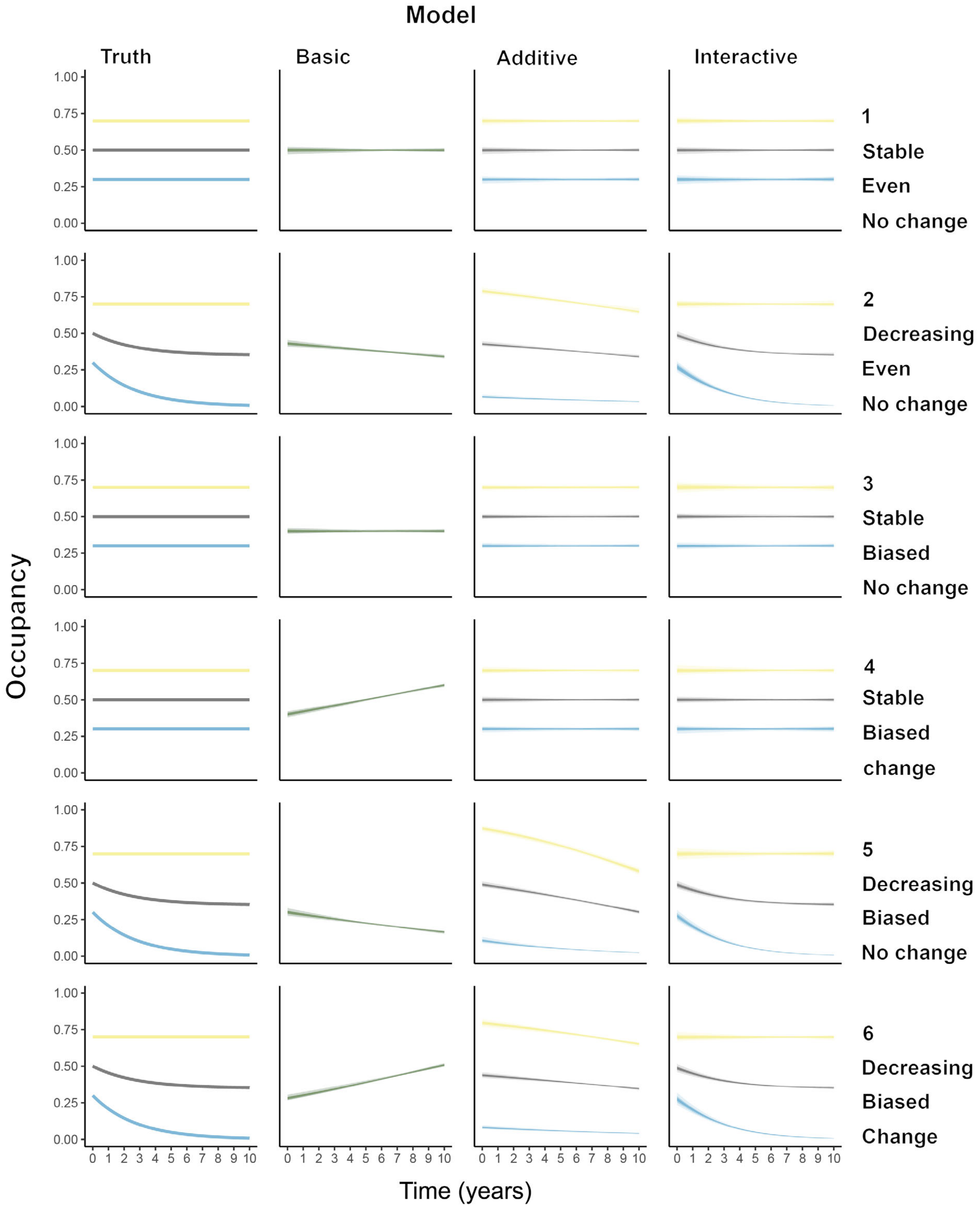


Figure 2. Results of six simulated scenarios. The results of three Generalized Linear Models are presented here, alongside the simulated truth; models generated estimates of occupancy for two habitats (yellow and blue); the average of the two habitats is shown in grey. Median model results are shown in opaque colours, whereas shaded colours indicate the 10th–90th centiles of all 100 simulations. In all models, occupancy starts at 0.7 for the yellow habitat and 0.3 for the blue.

© 2024 The Author(s). *Ibis* published by John Wiley & Sons Ltd on behalf of British Ornithologists' Union.

1474919x, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/ibi.13343 by University Of St Andrews University, Wiley Online Library on [16/07/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

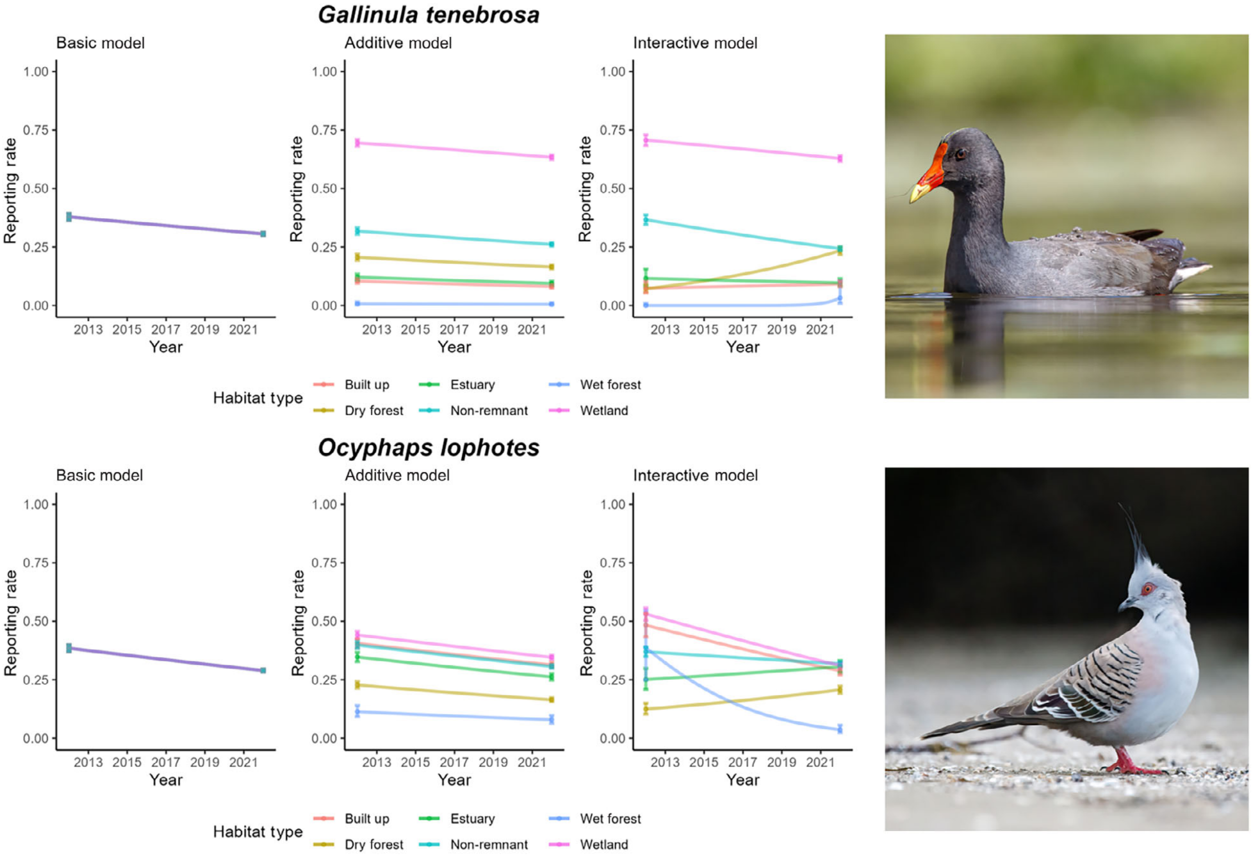
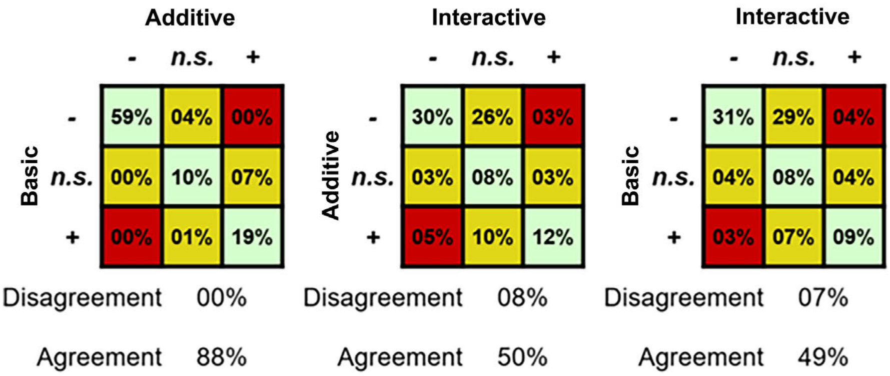


Figure 3. Generalized Linear Model predictions for two common Brisbane species, Dusky Moorhen *Gallinula tenebrosa* and Crested Pigeon *Ocyphaps lophotes*. Models generated estimates of reporting rate through time for six habitats; the estimate and confidence intervals for reporting rate at the start and end of the study period are shown with dots.



Basic \ Additive	-	n.s.	+
-	59%	04%	00%
n.s.	00%	10%	07%
+	00%	01%	19%

Disagreement 00%

Agreement 88%

Additive \ Interactive	-	n.s.	+
-	30%	26%	03%
n.s.	03%	08%	03%
+	05%	10%	12%

Disagreement 08%

Agreement 50%

Basic \ Interactive	-	n.s.	+
-	31%	29%	04%
n.s.	04%	08%	04%
+	03%	07%	09%

Disagreement 07%

Agreement 49%

Figure 4. Confusion matrices for the pairs of Generalized Linear Models applied to 157 common species in the Brisbane eBird dataset, aggregated across all six habitats. The trend estimate per-habitat for each species was calculated for each model, then allocated to a square in the confusion matrix according to the model result. Values in each square represent the percentage of all 942 (157 × 6) species–habitat trends that correspond to that pair of model outcomes. Numbers on the leading/main diagonal of each matrix (coloured in pale green) indicate species for which the two models agreed (same sign and significance), whereas values off this diagonal indicate either model ambiguity (one significant, one non-significant change; coloured in yellow) or disagreement (one positive, one negative change; coloured in red).

© 2024 The Author(s). *Ibis* published by John Wiley & Sons Ltd on behalf of British Ornithologists' Union.

1474919x, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/ibi.13343 by University Of St Andrews University, Wiley Online Library on [16/07/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Importantly, there was only 49% agreement (the leading/main diagonal of the matrices; pale green) between the interactive and basic models, and 7% disagreement (significant changes of estimates in opposite directions; red), indicating that sampling biases can lead to the wrong conclusions if not addressed during analysis.

DISCUSSION

Sampling biases in citizen science data present a key quality concern that analysts must consider when using such datasets. Here, we found varying types and degrees of sampling bias across different temporal and spatial scales in a subset of the eBird dataset. Importantly, we found evidence that some of these temporal and spatial sampling biases interacted with one another. When assessing habitat-specific population trends using simulations, these interactions between sampling biases were accounted for by increased model complexity, by explicitly including an interaction term between temporal and spatial variables. Similarly, when comparing trends using real data, we found frequently contradictory results between model types, demonstrating that simpler models that do not account for sampling biases in the data will often produce biased conclusions about average population trends. Together, our analyses highlight the importance of estimating sampling biases and/or using modelling approaches that accommodate spatially and temporally varying sampling biases when analysing citizen science data or other observation-based data (Binley & Bennett 2023). We demonstrate some potential issues that may arise if sampling biases are not appropriately dealt with during analysis.

We found strong temporal sampling bias at the yearly scale, with a rapid increase in the number of checklists over the study period (Bias (H') = 0.32; Fig. 1). This pattern is common across most large-scale biodiversity databases (Isaac & Pocock 2015), and reflects the increasing popularity and accessibility of citizen science among members of the public and of increased acceptance and promotion of citizen science by professional scientists (Follett & Strezov 2015). We also observed spatial sampling bias across habitat types (Bias (H') = 0.21; Fig. 1). Spatial sampling biases are well-documented in the literature (Meyer *et al.* 2016, Tang *et al.* 2021), including when considered across habitat types (Petersen *et al.* 2021, Shirey *et al.* 2021). Importantly, however, several of the temporal and spatial biases we estimated differed from those described in the literature for other databases. For example, protected areas in Brisbane are significantly undersampled in the eBird dataset (Bias (H') = 0.07; Supporting Online Information Fig. S3), in contrast to many other regions and datasets where protected areas are comparatively oversampled (e.g. Botts *et al.* 2011, Tulloch *et al.* 2013, Tang *et al.* 2021). We also observed limited seasonal sampling bias (Bias (H') = 0.03; Supporting Online Information Fig. S2), in contrast to other regions and datasets for which there is large variation in sampling effort throughout the year (e.g. La Sorte & Somveille 2019, Di Cecco *et al.* 2021). These variations in published sampling bias illustrate the risk in attempting to generalize patterns of bias across datasets or across broad spatial or temporal extents.

We investigated the effects of spatial and temporal sampling biases on model inference in two ways: a simulation study, with controlled sampling biases and known trends, and a real-world modelling exercise using eBird data, with uncontrolled sampling biases and unknown trends. The key finding of the simulation study was that increased sampling bias led to poorer trend estimates using simple models, particularly under scenarios with interactions between biases (Fig. 2, Supporting Online Information Table S2). When we applied the same models to eBird data, results diverged considerably (Fig. 4). There was only 49% agreement (same trend across models) – and 7% disagreement (different trend) – between the per-habitat trends produced by the interactive and basic models. Based on the results from our simulations, we expect that the more complex interactive models are more likely to be correct than the simpler models but cannot be sure of this as we have no measure of 'truth' to compare against. Indeed, sampling biases are a key reason that citizen science datasets can produce different (often less accurate) results to structured datasets, particularly when estimating population trends (Kamp *et al.* 2016, Boersch-Supan *et al.* 2019, Neate-Clegg *et al.* 2020).

The results of our analyses therefore underline the need for continued development and increased application of robust methods of accounting for, or alleviating the effect of, sampling biases in citizen science datasets. There has been significant research effort in this field in recent years, with increasingly sophisticated models being developed

© 2024 The Author(s). *Ibis* published by John Wiley & Sons Ltd on behalf of British Ornithologists' Union.

1474919x, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/ibi.13343 by University Of St Andrews University, Wiley Online Library on [16/07/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

to address the biases found in citizen science datasets (e.g. Bird *et al.* 2014, Fink *et al.* 2023). In this study, we explored the effect of sampling biases on Generalized Linear Models due to their simplicity and ease of manipulation, as the increased complexity of other model families (e.g. occupancy models, *N*-mixture models, or Machine Learning methods) makes it challenging to systematically test the effects of biases on their outputs. Our findings suggest that these models may be appropriate for citizen science data provided that they account for any sampling biases that are present and interactions among those biases, and that the data and ecological system meet the assumptions of the model. Further work in this area is needed, because many current methods do not account for strong biases or for interactions between multiple biases.

Sampling bias in citizen science data is an area that requires more research. As addressed above, we have assumed in this paper that our chosen spatiotemporal variables fully explain the sampling bias observed. This assumption is almost certainly violated in our work and given the amount of spatiotemporal variation in the observation process we expect it to be violated in almost every analysis of citizen science data. However, the analytical framework we have presented here enables analysts to identify the largest sources of sampling bias and to account for these in analyses. This approach can therefore be used to mitigate the effect of a complex and multi-faceted observation process by identifying the covariates associated with the largest sampling bias.

Further research is also needed to better understand the properties and behaviour of the Hoover Index (standard and modified), which we apply here for the first time in this field. Other similar measures (e.g. χ^2 test, Gini Index) could be explored as potential methods to estimate sampling biases. Extending the index to work with continuous predictor variables would also be worthwhile. Finally, there is scope for further simulation work to extend this study by exploring other biases (e.g. multiple biases within the same dimension), more complex interactions, or the performance of other, more sophisticated, model families.

In conclusion, sampling biases in citizen science datasets present a key challenge that must be addressed if these datasets are to be used to generate reliable information about biodiversity. In this paper, we present an easily applicable method for estimating sampling biases that can be used to assess biases in any detection–nondetection dataset. As a result of this work, we recommend that users of citizen science data use the methods presented here to consider and estimate sampling biases and their interactions across biologically relevant dimensions in their datasets as part of their analytical pipelines. This will ensure a comprehensive understanding of the structures of sampling biases that may influence results and the risks associated with them, in turn allowing for well-informed model specification and application of other methods to mitigate sampling biases if necessary. Once relevant sampling biases have been estimated, analysts may make informed decisions about model specification, depending on their given research questions and desired outcomes. Due consideration of sampling biases, if combined with appropriate model specification and careful interpretation of results, is the best way to increase the robustness of any analysis conducted using citizen science data.

We thank Dr Simone Blomberg, who provided statistical support at key junctures, and Dr John Dwyer who co-ordinated the academic component of this work. Many people read drafts and provided feedback at various points, and we thank them all. To meet institutional and research funder open access requirements, any accepted manuscript arising shall be open access under a Creative Commons Attribution (CC BY) reuse licence with zero embargo.

AUTHOR CONTRIBUTIONS

Louis J. Backstrom: Conceptualization; methodology; formal analysis; writing – original draft; project administration; visualization. **Corey T. Callaghan:** Methodology; writing – review and editing; supervision. **Hannah Worthington:** Writing – review and editing; supervision. **Richard A. Fuller:** Writing – review and editing; conceptualization; supervision; project administration. **Alison Johnston:** Writing – review and editing; supervision; project administration.

ETHICAL NOTE

None.

FUNDING

None.

© 2024 The Author(s). *Ibis* published by John Wiley & Sons Ltd on behalf of British Ornithologists' Union.

1474919x, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/ibi.13343 by University Of St Andrews University, Wiley Online Library on [16/07/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

CONFLICT OF INTEREST STATEMENT

The authors have no conflict of interest to disclose.

Data Availability statement

All datasets used have been cited in the text. A GitHub repository with full code and outputs is available at https://github.com/Louis-Backstrom/LJB-Honours-Public. Key R packages used include *auk* (Strimas-Mackey *et al.* 2021), *tidyverse* (Wickham *et al.* 2019) and *sf* (Pebesma 2018).

REFERENCES

August, T., Fox, R., Roy, D.B. & Pocock, M.J.O. 2020. Data-derived metrics describing the behaviour of field-based citizen scientists provide insights for project design and modelling bias. *Sci. Rep.* **10**: 11009.

Bhaskaran, K. & Smeeth, L. 2014. What is the difference between missing completely at random and missing at random? *Int. J. Epidemiol.* **43**: 1336–1339.

Binley, A.D. & Bennett, J.R. 2023. The data double standard. *Methods Ecol. Evol.* **14**: 1389–1397.

Bird, T.J., Bates, A.E., Lefcheck, J.S., Hill, N.A., Thomson, R.J., Edgar, G.J., Stuart-Smith, R.D., Wotherspoon, S., Krkosek, M., Stuart-Smith, J.F., Pecl, G.T., Barrett, N. & Frusher, S. 2014. Statistical solutions for error and bias in global citizen science datasets. *Biol. Conserv.* **173**: 144–154.

Boakes, E.H., Gliozzo, G., Seymour, V., Harvey, M., Smith, C., Roy, D.B. & Haklay, M. 2016. Patterns of contribution to citizen science biodiversity projects increase understanding of volunteers' recording behaviour. *Sci. Rep.* **6**: 33051.

Boersch-Supan, P.H., Trask, A.E. & Baillie, S.R. 2019. Robustness of simple avian population trend models for semi-structured citizen science data is species-dependent. *Biol. Conserv.* **240**: 286.

Botts, E.A., Erasmus, B.F.N. & Alexander, G.J. 2011. Geographic sampling bias in the South African frog Atlas project: implications for conservation planning. *Biodivers. Conserv.* **20**: 119–139.

Bowler, D.E., Callaghan, C.T., Bhandari, N., Henle, K., Benjamin Barth, M., Koppitz, C., Klenke, R., Winter, M., Jansen, F., Bruelheide, H. & Bonn, A. 2022. Temporal trends in the spatial bias of species occurrence records. *Ecography* **2022**: e06219.

Boyd, R.J., Powney, G.D., Burns, F., Danet, A., Duchenne, F., Grainger, M.J., Jarvis, S.G., Martin, G., Nilsen, E.B., Porcher, E., Stewart, G.B., Wilson, O.J. & Pescott, O.L. 2022. ROBITT: a tool for assessing the risk-of-bias in studies of temporal trends in ecology. *Methods Ecol. Evol.* **13**: 1497–1507.

Boyd, R.J., Powney, G.D. & Pescott, O.L. 2023. We need to talk about nonprobability samples. *Trends Ecol. Evol.* **38**: 521–531.

Boyd, R.J., Stewart, G.B. & Pescott, O.L. 2024. Descriptive inference using large, unrepresentative nonprobability samples: an introduction for ecologists. *Ecology* **105**: e4214.

Brown, E.D. & Williams, B.K. 2019. The potential for citizen science to produce reliable and useful information in ecology. *Conserv. Biol.* **33**: 561–569.

Buckland, S.T. & Johnston, A. 2017. Monitoring the biodiversity of regions: key principles and possible pitfalls. *Biol. Conserv.* **214**: 23–34.

Burgess, H.K., DeBey, L.B., Froehlich, H.E., Schmidt, N., Theobald, E.J., Ettinger, A.K., HilleRisLambers, J., Tewksbury, J. & Parrish, J.K. 2017. The science of citizen science: exploring barriers to use as a primary research tool. *Biol. Conserv.* **208**: 113–120.

Callaghan, C.T., Rowley, J.J.L., Cornwell, W.K., Poore, A.G.B. & Major, R.E. 2019. Improving big citizen science data: moving beyond haphazard sampling. *PLoS Biol.* **17**: e3000357.

Chandler, M., See, L., Copas, K., Bonde, A.M.Z., López, B.C., Danielsen, F., Legind, J.K., Masinde, S., Miller-Rushing, A.J., Newman, G., Rosemartin, A. & Turak, E. 2017. Contribution of citizen science towards international biodiversity monitoring. *Biol. Conserv.* **213**: 280–294.

Commonwealth of Australia. 2021. Collaborative Australian Protected Areas Database (CAPAD) 2020 – Terrestrial, Version 11. *Department of Climate Change, Energy, the Environment and Water.* Available at: https://www.environment.gov.au/fed/catalog/search/resource/details.page?uuid=%7B4448CACD-9DA8-43D1-A48F-48149FD5FCFD%7D (accessed 1 June 2023).

Cornell Lab of Ornithology. 2021. eBird Basic Dataset. Version: EBD_relDec-2021. *Cornell Lab of Ornithology.* Available at: https://ebird.org/data/download (accessed 1 June 2023).

Courter, J.R., Johnson, R.J., Stuyck, C.M., Lang, B.A. & Kaiser, E.W. 2013. Weekend bias in citizen science data reporting: implications for phenology studies. *Int. J. Biometeorol.* **57**: 715–720.

Di Cecco, G.J., Barve, V., Belitz, M.W., Stucky, B.J., Guralnick, R.P. & Hurlbert, A.H. 2021. Observing the observers: how participants contribute data to iNaturalist and implications for biodiversity science. *Bioscience* **71**: 1179–1188.

Dickinson, J.L., Zuckerberg, B. & Bonter, D.N. 2010. Citizen science as an ecological research tool: challenges and benefits. *Annu. Rev. Ecol. Evol. Syst.* **41**: 149–172.

Fink, D., Johnston, A., Strimas-Mackey, M., Auer, T., Hochachka, W.M., Ligocki, S., Jaromczyk, L.O., Robinson, O., Wood, C., Kelling, S. & Rodewald, A.D. 2023. A double machine learning trend model for citizen science data. *Methods Ecol. Evol.* **14**: 2435–2448.

Follett, R. & Strezov, V. 2015. An analysis of citizen science cased research: usage and publication patterns. *PLoS One* **10**: e0143687.

Gallant, J., Wilson, N., Dowling, T., Read, A. & Inskeep, C. 2011. SRTM-derived 1 Second Digital Elevation Models, Version 1.0. *Geoscience Australia*. Available at: http://pid.geoscience.gov.au/dataset/ga/72759 (accessed 1 June 2023).

Hoover, E.M. 1941. Interstate redistribution of population, 1850–1940. *J. Econ. Hist.* **1**: 199–205.

Horns, J.J., Adler, F.R. & Şekercioğlu, Ç.H. 2018. Using opportunistic citizen science data to estimate avian population trends. *Biol. Conserv.* **221**: 151–159.

© 2024 The Author(s). *Ibis* published by John Wiley & Sons Ltd on behalf of British Ornithologists' Union.

1474919x, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/ibi.13343 by University Of St Andrews University, Wiley Online Library on [16/07/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Hughes, A.C., Orr, M.C., Ma, K., Costello, M.J., Waller, J., Provoost, P., Yang, Q., Zhu, C. & Qiao, H. 2021. Sampling biases shape our view of the natural world. *Ecography* **44**: 1259–1269.

Isaac, N.J.B. & Pocock, M.J.O. 2015. Bias and information in biological records. *Biol. J. Linn. Soc.* **115**: 522–531.

Isaac, N.J.B., van Strien, A.J., August, T.A., Zeeuw, M.P. & Roy, D.B. 2014. Statistics for citizen science: extracting signals of change from noisy ecological data. *Methods Ecol. Evol.* **5**: 1052–1060.

Johnston, A., Moran, N., Musgrove, A., Fink, D. & Baillie, S.R. 2020. Estimating species distributions from spatially biased citizen science data. *Ecol. Model.* **422**: 927.

Johnston, A., Matechou, E. & Dennis, E.B. 2023. Outstanding challenges and future directions for biodiversity monitoring using citizen science data. *Methods Ecol. Evol.* **14**: 103–116.

Kamp, J., Oppel, S., Heldbjerg, H., Nyegaard, T. & Donald, P.F. 2016. Unstructured citizen science data fail to detect long-term population declines of common birds in Denmark. *Divers. Distrib.* **22**: 1024–1035.

Kelling, S., Johnston, A., Hochachka, W.M., Iliff, M., Fink, D., Gerbracht, J., Lagoze, C., La Sorte, F.A., Moore, T., Wiggins, A., Wong, W.K., Wood, C. & Yu, J. 2015. Can observation skills of citizen scientists Be estimated using species accumulation curves? *PLoS One* **10**: e0139600.

Kelling, S., Johnston, A., Bonn, A., Fink, D., Ruiz-Gutierrez, V., Bonney, R., Fernandez, M., Hochachka, W.M., Julliard, R., Kraemer, R. & Guralnick, R. 2019. Using semistructured surveys to improve citizen science data for monitoring biodiversity. *Bioscience* **69**: 170–179.

La Sorte, F.A. & Somveille, M. 2019. Survey completeness of a global citizen-science database of bird occurrence. *Ecography* **43**: 34–43.

Meyer, C., Weigelt, P. & Kreft, H. 2016. Multidimensional biases, gaps and uncertainties in global plant occurrence information. *Ecol. Lett.* **19**: 992–1006.

Neate-Clegg, M.H.C., Horns, J.J., Adler, F.R., Kemahlı Aytekin, M.Ç. & Şekercioğlu, Ç.H. 2020. Monitoring the world's bird populations with community science data. *Biol. Conserv.* **248**: 653.

Newson, S.E., Massimino, D., Johnston, A., Baillie, S.R. & Pearce-Higgins, J.W. 2013. Should we account for detectability in population trends? *Bird Study* **60**: 384–390.

Pebesma, E.J. 2018. Simple features for R: standardized support for spatial vector data. *R J.* **10**: 439–446.

Pescott, O.L., Humphrey, T.A., Stroh, P.A. & Walker, K.J. 2019. Temporal changes in distributions and the species atlas: how can British and Irish plant data shoulder the inferential burden? *Br. Irish Bot.* **1**: 250–282.

Petersen, T.K., Speed, J.D.M., Grøtan, V. & Austrheim, G. 2021. Species data for understanding biodiversity dynamics: the what, where and when of species occurrence data collection. *Ecol. Solut. Evid.* **2**: e12048.

Queensland Department of Resources. 2022. Built up areas – Queensland, Version 6.13. *Queensland Spatial Catalogue – QSpatial*. Available at: https://qldspatial.information.qld.gov.au/catalogue/custom/viewMetadataDetails.page?uuid=%7B063A413F-7910-4E6B-8389-24E06AF4508C%7D (accessed 1 June 2023).

Queensland Herbarium. 2019. Wetland areas – Queensland, Version 5. *Queensland Spatial Catalogue – QSpatial*. Available at: https://qldspatial.information.qld.gov.au/catalogue/custom/viewMetadataDetails.page?uuid=%7B135EB151-D406-4094-9E9F-40ABC5AA0C7B%7D (accessed 1 June 2023).

Queensland Herbarium. 2022. Biodiversity status of 2019 remnant regional ecosystems – Queensland, Version 12.2. *Queensland Spatial Catalogue – QSpatial*. Available at: https://qldspatial.information.qld.gov.au/catalogue/custom/viewMetadataDetails.page?uuid=%7B8FDF54D2-654C-4822-8295-1D8E8E772373%7D (accessed 1 June 2023).

R Core Team. 2022. *R: a language and environment for statistical computing*. Version 4.2.0. Vienna: R Foundation for Statistical Computing. Available at: https://www.R-project.org/ (accessed 1 June 2023).

Schmill, M.D., Gordon, L.M., Magliocca, N.R., Ellis, E.C. & Oates, T. 2014. GLOBE: analytics for assessing global representativeness. In *2014 Fifth International Conference on Computing for Geospatial Research and Application*: 25–32. IEEE.

Shirey, V., Belitz, M.W., Barve, V. & Guralnick, R. 2021. A complete inventory of North American butterfly occurrence data: narrowing data gaps, but increasing bias. *Ecography* **44**: 537–547.

Steen, V.A., Elphick, C.S. & Tingley, M.W. 2019. An evaluation of stringent filtering to improve species distribution models from citizen science data. *Divers. Distrib.* **25**: 1857–1869.

Steiner, R., Niemi, G., Nicoletti, F., Evans, M.J., Zlonis, E. & Etterson, M.A. 2022. Changes in survey effort can influence conclusions about migration phenology. *J. Raptor Res.* **56**: 171–179. https://doi.org/10.3356/jrr-21-22

Strimas-Mackey, M., Miller, E. & Hochachka, W. 2021. *auk: eBird data extraction and processing with AWK*. R package version 0.5.1. Available at: https://cornelllabofornithology.github.io/auk/ (accessed 1 June 2023).

Tang, B., Clark, J.S. & Gelfand, A.E. 2021. Modeling spatially biased citizen science effort through the eBird database. *Environ. Ecol. Stat.* **28**: 609–630.

Tulloch, A.I.T., Mustin, K., Possingham, H.P., Szabo, J.K. & Wilson, K.A. 2013. To boldly go where no volunteer has gone before: predicting volunteer activity to prioritize surveys at the landscape scale. *Divers. Distrib.* **19**: 465–480.

Ver Hoef, J.M., Johnson, D., Angliss, R. & Higham, M. 2021. Species density models from opportunistic citizen science data. *Methods Ecol. Evol.* **12**: 1911–1925.

Ward, D.F. 2014. Understanding sampling and taxonomic biases recorded by citizen scientists. *J. Insect Conserv.* **18**: 753–756.

Wickham, H., Averick, M., Bryan, J., Chang, W., Mcgowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K. & Yutani, H. 2019. Welcome to the Tidyverse. *J. Open Source Softw.* **4**: 1686.

Received 11 June 2023;
Revision 24 April 2024;
revision accepted 15 June 2024.
Associate Editor: Ben Smit.

© 2024 The Author(s). *Ibis* published by John Wiley & Sons Ltd on behalf of British Ornithologists' Union.

1474919x, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/ibi.13343 by University Of St Andrews University, Wiley Online Library on [16/07/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Figure S1. Map of broad habitat types and protected areas in Brisbane City. The location of Brisbane within Australia is indicated in the inset. Unlike many other urban areas, the Brisbane City Local Government Area contains extensive green space, particularly in the northwest of the city, and covers a range of broad habitat types.

Figure S2. Temporal sampling biases in the Brisbane eBird Dataset. Biases are presented here across four scales: yearly (top left), monthly (top right), daily (bottom left) and hourly (bottom right). Translucent red bars indicate the null distribution under representative sampling, blue the observed distribution, and dark pink the overlap between red and blue bars.

Figure S3. Spatial sampling biases in the Brisbane eBird Dataset. Biases are presented here across three scales: elevation (top left), protected area status (top right) and habitat type (bottom).

Figure S4. Yearly interactions in the Brisbane eBird Dataset. Interactions are presented here across three secondary temporal scales: monthly (top), daily (middle) and hourly (bottom).

Figure S5. Per-habitat type interactions in the Brisbane eBird Dataset. Interactions are presented here across two secondary spatial scales: elevation (top) and protected area status (bottom).

Figure S6. Yearly interactions in the Brisbane eBird Dataset. Interactions are presented here across three secondary spatial scales: elevation (top), habitat type (middle) and protected area status (bottom).

Figure S7. Monthly interactions in the Brisbane eBird Dataset. Interactions are presented here across two secondary temporal scales: daily (top) and hourly (bottom).

Figure S8. Daily interactions in the Brisbane eBird Dataset. Interactions are presented here across one secondary temporal scale: hourly.

Figure S9. Per-protected area status interactions in the Brisbane eBird Dataset. Interactions are presented here across two secondary spatial scales: elevation (top) and habitat type (bottom).

Figure S10. Monthly interactions in the Brisbane eBird Dataset. Interactions are presented here across three secondary spatial scales: elevation (top), habitat type (middle) and protected area status (bottom).

Figure S11. Daily interactions in the Brisbane eBird Dataset. Interactions are presented here across three secondary spatial scales: elevation (top), habitat type (middle) and protected area status (bottom).

Table S1. Key to habitat classifications for Regional Ecosystem types within Brisbane City.

Table S2. Summary of simulation results for each of the six scenarios of sampling bias tested. Two habitats (yellow/blue) were simulated, and three different models of varying complexity were fit to the simulated data (basic/additive/interactive).

© 2024 The Author(s). *Ibis* published by John Wiley & Sons Ltd on behalf of British Ornithologists' Union.

1474919x, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/ibi.13343 by University Of St Andrews University, Wiley Online Library on [16/07/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License