CITIZEN SCIENCE:
THEORY AND PRACTICE

Evaluation of the Spatial Biases and Sample Size of a Statewide Citizen Science Project

RESEARCH PAPER

]u[ubiquity press

ROLAND KAYS
MONICA LASKY
ARIELLE W. PARSONS
BRENT PEASE
KRISHNA PACIFICI

*Author affiliations can be found in the back matter of this article

ABSTRACT

Although quality control for accuracy is increasingly common in citizen science projects, there is still a risk that spatial biases of opportunistic data could affect results, especially if sample size is low. Here we evaluate how well the sampling locations of North Carolina's Candid Critters citizen science camera trapping project represented available land cover types in the state and whether the sample size (4,295 sites) was sufficient to estimate ecological parameters (i.e., species occupancy) with low bias and error. Although most sampling was opportunistic, we used a "Plan, Encourage, Supplement" approach to improve our spatial coverage. We assessed potential biases by comparing seven dimensions of habitat (i.e., land cover, elevation, road density, etc.) sampled by camera traps with those available in the state, using a minimum sample threshold approach, and found that the variation of habitat across the state was sufficiently sampled. At the ecoregion level we sampled 99.2% (±0.01) of the variation of potential habitat "adequately" and 96.4% (±0.03) "very adequately." Supplemental sampling by staff helped meet sampling adequacy for 6.8% of ecoregion-habitat classes, especially in less populated parts of the state. Compared with results from the full data set, the relative bias and error with subsets of the data dropped below 10% relatively quickly with increasing sample size for estimates of occupancy, suggesting that results estimated with the full sample are robust, although the precision of particular ecological relationships were more variable. These analyses show that opportunistic sampling can be representative of large areas if sample size is high enough and that a priori sampling goals can help improve coverage by encouraging volunteers to sample in certain places or through supplemental data collection by staff.

CORRESPONDING AUTHOR:
Roland Kays
North Carolina Museum of Natural Sciences, US
rwkays@ncsu.edu

KEYWORDS:
sample design; camera trap; spatial bias; mammals; large scale

TO CITE THIS ARTICLE:
Kays, R, Lasky, M, Parsons, AW, Pease, B and Pacifici, K. 2021. Evaluation of the Spatial Biases and Sample Size of a Statewide Citizen Science Project. *Citizen Science: Theory and Practice,* 6(1): 34, pp. 1–12. DOI: *https://doi.org/10.5334/cstp.344*

INTRODUCTION

Data quality is central to citizen science projects producing trustworthy scientific outcomes. As the citizen science approach became more popular, some scientists voiced concerns about the potential for volunteers to produce data sets without large amounts of error (Dickinson, Zuckerberg, and Bonter 2010) that would be able to detect changes in the population status of wild species (Danielsen et al. 2014). In response to those challenges, many best practices in citizen science now feature mechanisms to assess the quality of data. Most of this development has focused on the accuracy and trustworthiness of their observations (e.g., Crall et al. 2011; Hunter et al. 2013; Kosmala et al. 2016). However, even if citizen observations are accurate, there is an additional risk for sampling bias because most are opportunistic samples (Bird et al. 2014).

Because citizen science projects rely on volunteers to collect data, it is difficult to follow a systematic sampling design. For example, in a study of two Canadian aquatic monitoring programs, Millar et al. (2019) found clustered and biased sampling around lakeshore houses, which they referred to as the "cottage effect." This spatial bias can affect biological inference; for example, Weiser et al. (2020) found strong bias when building statistical models that included non-probabilistically selected sites to survey butterflies (also known as preferential sampling; Diggle, Menezes, and Su 2010), which are typical of citizen science, although they noted that this might be less of a problem with larger data sets. There is some evidence that larger sample sizes might help offset these biases, as spatial models with different sampling schemes were similar for large data sets on toads in the United Kingdom (Petrovan, Vale, and Sillero 2020). Similarly, Callaghan et al. (2020) found that citizen science data can be as good as professional records for continental diversity mapping once a minimum sample size is met. There has also been work on accounting for biases in data collection during analysis. For example, Johnston et al. (2020) showed how spatial bias can be corrected for in spatial models by weighting based on the probability of an area being sampled, and Isaac et al. (2014) showed how filtering and data corrections could improve power to detect trends. However, both of these studies highlight the limitations of post-hoc corrections, emphasizing the importance of reducing bias at the point of data collection.

Here we address the questions of spatial bias and sample-size requirements in a citizen science data set collected as part of the statewide North Carolina's Candid Critters (NCCC) camera trapping wildlife survey. Recognizing these potential problems at the start of the project, we created an a priori study design. During the study, we monitored our progress and encouraged volunteers to run cameras in sites that would help meet those sampling goals. For the most poorly sampled areas, we supplemented the volunteer data with professional data collection, employing a hybrid citizen/professional scientist sampling program. We refer to this as the "Plan, Encourage, Supplement" strategy, which is similar to the strategies introduced by Callaghan et al. (2019).

The objective of the NCCC project was to document the distribution and ecological relationships of mammals in the state through statistical models that relate the occurrence of species in the camera traps to environmental data such as land cover, habitat type, or degree of human disturbance. In this paper, our goal is to assess potential biases in citizen science data by comparing the dimensions of habitat sampled by our cameras with those available in the state, using a minimum sample threshold approach (Callaghan et al. 2020). We also evaluate the sample-size robustness by comparing the relative bias and precision of wildlife metrics (i.e., species occupancy) calculated with subsets of the full data set. This subsetting approach provides insight into the extent to which finer-scale comparisons can be made between regions, seasons, or less common species that might naturally have fewer detections with a given survey effort. We expect these results will be broadly useful for professionals setting up citizen science projects who plan to use opportunistic data to study ecological relationships and want to ensure they end up with enough samples to statistically test hypotheses, and the samples are distributed to be representative of the larger study area.

METHODS

The North Carolina's Candid Critters project worked with citizen scientists and the North Carolina Wildlife Resources Commission personnel to survey wildlife with camera traps across the state over 3 years (Lasky et al. 2021a). Camera traps have advantages over some traditional wildlife data sources because they collect data across all seasons and times of day on a variety of warm-blooded species > 100 g. The photos allow species identifications to be verified (all NCCC pictures were reviewed by experts), and the duration a camera is deployed gives a clear record of sampling effort. The citizen science approach also allowed us to sample private land (52% of sites), for which it is typically difficult for researchers to obtain permission, even though it can represent a large part of a study area (here 86% in southern forests, Butler and Wear 2013).

However, because most (65%) of our cameras were set by volunteers, we had limited ability to determine where exactly they were set ahead of time. Nonetheless, as part of our Plan, Encourage, Supplement strategy, we had a priori sampling goals, and recruited and encouraged volunteers to set cameras in ways that would meet those goals. We supplemented by having staff cameras, as well. When planning the project, we used counties as an operational unit, aiming to get samples from all 100 counties. We initially quantified the land-cover types across the state as open, forested, or developed, and quantified the proportion of each primary land-cover type across counties (Lasky et al. 2021a). We then used this as a guideline for creating our study design on public lands, and for monitoring where cameras were run on private lands during the study, aiming for the simple goal of having samples in proportion to available land-cover types in each county. However, when evaluating if our data represents the state, we focused on whether we had an adequate sample of a given habitat type (to discover animal-habitat relationships) rather than an exact proportional match to what is available, since analyses were done at larger scales than county (i.e., state or ecoregion).

As the study progressed, we noticed areas that were being under sampled, especially rural areas, and designed campaigns to obtain more sampling in those locations by recruiting volunteers locally or by encouraging existing volunteers to travel outside their home county to set cameras on public land (e.g., visit a state park). This encouragement was done as part of our ongoing engagement activities with volunteers; it included newsletters, social media, and webinars (Lasky et al. 2021a), and is similar to the dynamic incentives suggested by Callaghan et al. (2019). If sampling was still lower than project goals, we supplemented the data by deploying cameras ourselves or by working with professional partners to run cameras in specific target areas.

SPATIAL BIAS

A recent empirical assessment of camera-trapping study design recommends a minimum of 40 to 60 camera points per habitat category (Kays et al. 2020). Based on this, we considered > 40 camera points to be an adequate sampling of a given habitat type, and > 60 to be a very adequate sample. Although this sample-size criteria might still not be sufficient for very rare species, it was derived from a wide variety of species and sites and serves as a good benchmark for our objectives.

Our approach (***Box 1***) to evaluate the representativeness of our sampling involved categorizing a variety of habitat types in the state, and evaluating whether each dimension was adequately sampled by our cameras. We did this first at the state level, but also at the finer-scale level of ecoregions. There are three primary ecoregions in North Carolina (mountains, piedmont, coastal), and using this division allowed us to also evaluate if ecological relationships could be different in these different regions. To describe the habitats of the state, we generated 10,000 randomly distributed points across the state, and after removing those in open water, used 9,586 random points in terrestrial habitats (***Figure 1***). At each of these points, we quantified the habitat with seven different covariates describing the natural and human infrastructures that are likely to affect animal distribution (***Table 1***). We then broke the variation in these each of these ecological covariates

Box 1 Overview of our approach for determining spatial bias in NCCC camera sites

Objective: Determine how representative the locations sampled by NCCC cameras were compared with all available habitat in the state.

Criteria: If a habitat type was sampled by > 40 camera sites it was judged as adequate, while those with > 60 were very adequate.

Habitat Availability: Habitat variation in the state was quantified using 9,586 random points. At each point, we measured seven different aspects of habitat (***Table 1***). For each habitat variable we split the statewide variation into 10 categories (bins in ***Figure 2***). We then used the above criteria to evaluate if each bin of each habitat variable had been sufficiently sampled.

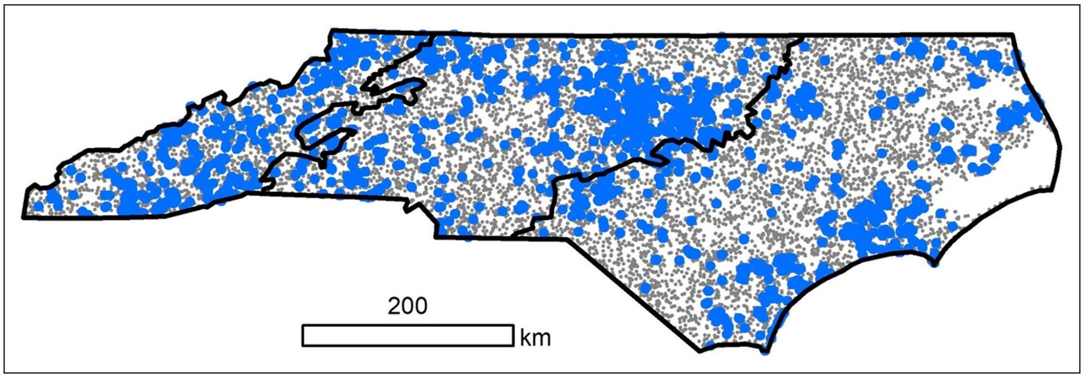

Figure 1 Map of the 9,586 random points (*grey*) across North Carolina, USA used to quantify variation in available habitat types, and the 4,295 sites sampled with camera traps (*blue*). Black lines indicate the three primary ecoregions (from west to east): mountains, piedmont, coastal.

into bins, each representing 10% of the total statewide variation as measured by the random points (***Figure 2***). In the case of highly skewed distributions, we had fewer than 10 bin levels of available habitat because one category would have a very high proportion of the total variation (e.g., 60% of random points had 0 roads within 250 m). We used these categories to describe the dimensions of variation of habitat in the state, and in each ecoregion.

For each of the seven habitat covariates (i.e., elevation, road density, etc.; ***Table 1***), we then quantified the habitat at the 4,295 sites sampled by our camera traps and placed them into the same ten bins as the available habitat (***Figure 2***) to evaluate if we had an adequate (> 40) or very adequate (>60) sampling in each dimension. Because ecological relationships might be different across the state (i.e., animals in the mountains might be more or less likely to occupy a site based on local forest cover than animals on the coast), we also considered how many camera traps fit into these categories separately in the three major ecoregions of the state (coastal, piedmont, and mountains).

SUFFICIENT SAMPLE SIZE

Occupancy modeling was the primary statistical model used to estimate ecological relationships and to map

HABITAT	DESCRIPTION	UNIT	SOURCE
Cover type	Open, forest, developed, or other		(Homer et al. 2015)
Elevation	Elevation	m	USGS
Large core forests	Area within a 5 km radius consisting of continuous forest fragment (forest parcels >2 ha)	%	(Homer et al. 2015)
Developed	Area within a 5 km radius consisting of developed land use	%	(Homer et al. 2015)
Houses	Housing density within 1 km radius of the site	houses/km^2	(Hammer et al. 2004)
Roads	Road density within a 250 m radius of site	km/km^2	NCDOT
Tree Cover	Tree cover at 30 m pixel resolution	%	(Hansen et al. 2013)

Table 1 Habitat variables considered for the analysis of how representative the data were of the state.

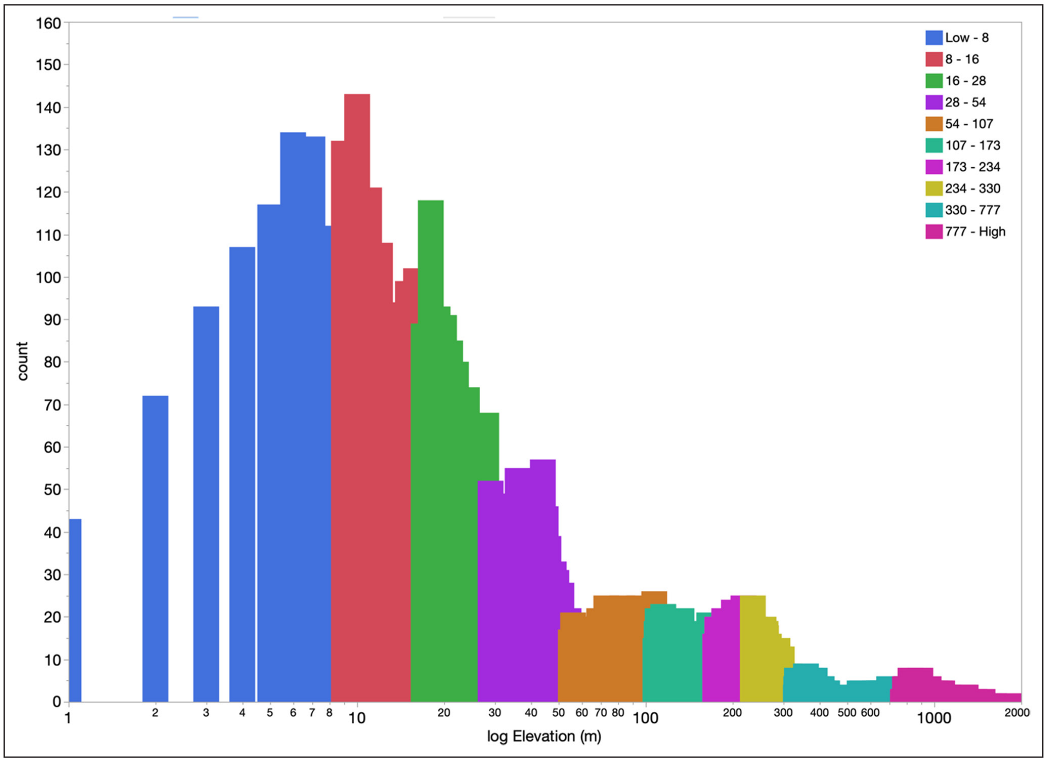

Figure 2 Distribution of elevational habitats available in North Carolina measured at 9,586 random points across the states. Colors show the 10 categories that each represent ~10% of the variation in available habitat. We used these categories to see if our camera samples were representative of the dimensions of a given habitat type.

animal distribution, so we evaluated how sensitive our estimates were to changes in sample size for a common (white-tailed deer, *Odocoileus virginianus*) and a rarer (coyote, *Canis latrans*) species (*Box 2*). If the data set was not robust, we would expect occupancy estimates to have high uncertainty and to vary greatly with changes in sample size, compared with the full sample. That is, we would expect a smaller random sample of the 4,295 sites (e.g., 10% of the total data set) to result in a different estimate of species-specific occupancy probability. If the sample was robust, however, we would expect estimates of bias and precision to level off well before the full-sample estimate.

Box 2 Overview of our approach for determining if the sample size was robust

Objective: Determine whether sufficient data had been collected to estimate the occupancy of common (white-tailed deer, *Odocoileus virginianus*) and rarer (coyote, *Canis latrans*) species.

Criteria: We used a measure of relative bias and a measure of precision relative mean square error (RMSE) to quantify the accuracy of occupancy estimates with, estimates < 0.1 RMSE or bias judged as accurate.

Approach: We used the full data set to estimate "true" results and then used subsamples to evaluate the effect of sample size. Analyses were conducted at the ecoregion level.

We used multiscale single-species, single-season occupancy models (Mordecai et al. 2011; Schmidt et al. 2013; MacKenzie et al. 2017) to estimate species-specific occupancy probabilities across the state as a function of habitat (e.g., proportion of forest cover in 5 km^2 area) and of site-level characteristics (e.g., whether camera was deployed within a residential yard) (*Table 2*). Occupancy models are a common way to analyze camera trap data while accounting for imperfect detection and variation in survey effort. The multiscale approach models occupancy probability at two scales—unit (i.e., grid cell) and subunit (i.e., immediate area surrounding camera) levels—all while accounting for imperfect detection of the observations. That is,

$$Z_i \sim Bernoulli\left(\psi_i\right)$$
$$\psi_i = \mathrm{X}_i^T \boldsymbol{\beta}$$
$$a_{ij} | Z_i \sim Bernoulli\left(Z_i \times \theta_{ij}\right)$$
$$\theta_{ij} = \mathrm{X}_{ij}^T \boldsymbol{\gamma}$$
$$y_{ij} | a_{ij} \sim Binomial\left(n_{ij}, \mu_{ij}\right)$$
$$\mu_{ij} = a_{ij} \times p_{ij}$$
$$logit\left(p_{ij}\right) = \mathrm{X}_{ij}^T \boldsymbol{\alpha}$$

where Z_i is the true but unknown occurrence status in grid cell *i*; $\boldsymbol{\psi}_i$ is the probability of occupancy in grid cell *i*, which is a function of covariates $\boldsymbol{X}$; and estimated regression coefficients $\boldsymbol{\beta}$ represent the effects of habitat effects (Habitat category in *Table 2*). a_{ij} represents the small-scale occurrence (or suitability) of the immediate camera-level conditions, *j*, and is conditional on the occurrence status in

CATEGORY	COVARIATE	DESCRIPTION
Habitat	Forest cover	% forested in 5 km^2 buffer
Habitat	Housing density	Average housing density (houses/km) in 5 km^2 square buffer
Habitat	Contagion index ("Clumpiness")	The propensity for a 5 km^2 square raster pixel of a given land-cover class to be neighboring a different land-cover class
Habitat	PRD	Patch richness density. Number of land-cover types per 100 ha in 5 km^2 square buffer
Site variation	Yard	Categorical predictor of whether the camera was placed within a residential yard
Site variation	Richness	Number of species detected at the camera site
Nuisance	Precipitation	Precipitation rate averaged over camera deployment period (Mesinger et al. 2006)
Nuisance	Temperature	Temperature averaged over camera deployment period (Dee et al. 2011)
Nuisance	EVI	Enhanced vegetation index; a measure of greenness at the camera site.
Nuisance	Julian	Julian day of the year
Nuisance	Detection distance	Furthest distance away that the camera was triggered by a human
Nuisance	Bait	Categorical of whether bait was used at the camera site
Survey effort	Trap nights	Length (days) of camera trap deployment. Used to control for variation in effort (i.e., catch per unit effort)

Table 2 Covariates used in occupancy models.

grid cell *i* (i.e., a camera can be occupied only if the grid cell is occupied); $\boldsymbol{\theta}_{ij}$ is the small-scale occupancy probability and is again a function of environmental conditions immediately surrounding the camera (Site variation category in ***Table 2***); and $\boldsymbol{\gamma}$ represents the estimated regression coefficients of those covariates. y_{ij} are the species-specific camera observations and are conditional on the site being suitable (i.e., we can only detect the species given that the site is suitable for that species); with n_{ij} the number of days a given camera ran and p_{ij} is the site-level detection probability, or the probability of detecting the species given that the species is present, and is a function of site-level detection covariates (Nuisance category in ***Table 2***) with α representing the regression coefficients of those effects.

To evaluate whether our estimates of occupancy were robust, we initially fit occupancy models to the complete NCCC data set and used these values to reflect our best estimate of species-specific occupancy. We then subset the data set for the three ecoregions of the state, four seasons, and region-season. Seasons were the quarters of the year, where winter was January through March, spring was April through June, etc. For the region-season combination, we further divided the data set among ecoregions and seasons (e.g., mountain ecoregion during winter season). We then systematically subsampled (i.e., reduced the number of cameras available to inform occupancy) the NCCC data set to understand how our estimates of occupancy would change as we reduced the number of cameras available to inform our models. At each percentage, we randomly selected cameras 20 different times—resulting in 20 different groups of cameras—and at each random sample we fit the same model that was used with the full data set. We recorded the estimated occupancy probabilities from the subsampled data, and after the 20 random samples, we summarized how those estimates compared with the full data set using relative bias (RBIAS) and relative root mean square error (RRMSE):

$$RRMSE = \frac{\sqrt{\left(\frac{1}{r}\right)\sum_{l=1}^{r}(\hat{\theta}_l - \theta_l)^2}}{\bar{\theta}}$$

$$RBIAS = \frac{\left(\frac{1}{r}\right)\sum_{l=1}^{r}\left(\hat{\theta}_l - \theta_l\right)}{\bar{\theta}}$$

where *r* is the number of replicates (in this case, 20), $\hat{\theta}_l$ is the estimated parameter (mean) at replicate *l*, θ_l was the true parameter value calculated using the full data set, and $\bar{\theta}$ is the mean of the true parameter values across all replicates. RRMSE is a measure of how variable the estimates of occupancy were across the replicates, and RBIAS is a measure of how different the estimates were from the full data set estimate. The ultimate goal is to have RRMSE and RBIAS near zero, which would reflect little variation across each replicate, and an estimated occupancy nearly identical to the full data set and therefore robust to data reduction.

RESULTS

SPATIAL COVERAGE

At the statewide level, we obtained very adequate (>60 sites) sampling of all habitat dimensions (58 categories across seven habitat dimensions; Appendix 1). The lowest sampling was 161 cameras for the 28–54 m elevation category. At the ecoregion level, we had adequate samples for 93% of the habitat-ecoregion categories (136/146 samples; ***Table 3***; Appendix 1). However, most of these were

TREE COVER %	STAFF CAMS	VOLUNTEER CAMS	ALL CAMS	RANDOM %	ADEQUATE SAMPLE (>40)	VERY ADEQUATE SAMPLE (>60)
0	**17**	51	68	13%	13%	13%
0–40	**10**	**24**	**34**	4%	0	0
40–81	**22**	78	100	9%	9%	9%
81–94	50	96	146	12%	12%	12%
94–98	**33**	101	134	12%	12%	12%
98–100	**18**	**26**	44	3%	3%	0
100	99	348	447	45%	45%	45%
				Total	**95.6%**	**92.2%**

Table 3 Example of results for the representation of the camera trap sampling for one ecoregion (mountains) for one habitat type (tree cover). Columns show number of cameras set in each habitat type by staff, volunteers, and total. Habitat types with <40 camera samples (*bold*) were judged to be insufficiently sampled. In this example, additional sampling by staff ensured adequate sampling for the 98–100% category but not for the 0–40% category. The proportional availabilities of a habitat categories for that ecoregion are given by the % of random points that fell into that category, which are then summed if they are sampled adequately (>40 pts) or very adequately (>60) to quantify the total % of a given habitat type adequately sampled in a given ecoregion. Additional habitats/ecoregion results are in Appendix 1.

HABITAT COVARIATE	COASTAL	MOUNTAINS	PIEDMONT	AVERAGE
Tree cover	100	95.6	100	98.5
Elevation	99.7	99.9	99.7	99.8
Large forests	98.9	93.7	99.8	97.5
Developed	100	100	99.6	99.9
Houses	100	100	99.2	99.7
Land use	100	96.3	100	98.8
Roads	100	100	100	100
Average	99.8	97.9	99.8	99.2

Table 4 Percentage of the area in three ecoregions of the state adequately sampled (>40 sites) by camera traps in the North Carolina's Candid Critters(NCCC) project across seven habitat dimensions. See Table 3 for an example of how this was estimated for one habitat/ecoregion and Appendix 1 for all results.

in categories that were extremely rare in a given ecoregion (e.g., low elevation areas in the mountains ecoregion). When considering the proportion of area each habitat made up of each ecoregion, we had adequate sampling for habitat types covering 97.5–100% of the state's area, and for three ecoregions, 97.9–99.8% of the area (***Table 4***). At the more stringent, very adequate level we had lower coverage of the mountains ecoregion (90%; Appendix 1), but similar levels for the other two ecoregions. For these seven habitat characteristics and three ecoregions, on average, NCCC data obtained an adequate sampling for dimensions of habitats covering 99.2% of the area in the state (***Table 4***) and a very adequate sampling for 96.4% of the state (Appendix 1).

The additional sampling done by staff allowed us to improve sampling for 35 habitat-ecoregion categories that would have been under sampled with volunteers alone, providing sampling for 9 categories that had been under sampled, and bringing another 26 up to the level of very adequate sampling. Many of these categories consisted of significant proportions of the total habitat type, such that this supplemental sampling allowed us to get a representative sample for important aspects of the state's variation (Appendix 1). This was most important for the mountains ecoregion where this additional sampling raised the percentage of adequately sampled available habitat from 91.9% to 97.9%, and increased our proportion of area with very adequate sampling from 86% of habitat to 90%.

SAMPLE SIZE ROBUSTNESS

Occupancy estimates were robust to changes in sample size, with error and bias falling below our goal of 0.1 quickly (i.e., 10%; ***Figure 3***; Supplemental Figure 1). Not surprisingly, error and bias fell below our threshold for the more common deer, with smaller sample size than the less-common coyote. We also plotted the sensitivity of ecological relationships described by the model to changes in sample size in terms of error and bias (Supplemental Figure 2). The error associated with estimating these relationships approached, but did not meet, our 10% goal in most cases. Interestingly, for a given number of camera sites, error was lower for the more restrictive season-region models than the season-only or region-only (***Figure 4***), suggesting a fair amount of spatio-temporal variability in these relationships. Bias, on the other hand, decreased rapidly after few hundred sites in most cases (Supplemental Figure 2).

DISCUSSION

The opportunistic nature of sampling within most citizen science projects puts them at risk of providing biased results, especially if sample size is low. We used a Plan, Encourage, Supplement strategy to try to obtain a balanced sample by having a priori sampling goals, by encouraging volunteers to help meet those goals, by setting additional cameras ourselves in underrepresented areas, and in general, by obtaining a large sample size. Here we present a unique approach to evaluating the representativeness of our data, comparing it with a random sample of seven environmental covariates (i.e., dimensions of habitat) at both the statewide and ecoregion levels. We show that we obtained an adequate sample across the spectrum of variation in these habitat dimensions at the state level, and for most of the variation at the ecoregion level. Dimensions that were under sampled were rare, such that we adequately sampled 99.2% of the area of the state at the ecoregion level, and very adequately sampled 96.4%. Furthermore, sample sizes were robust enough to estimate occupancy for important wildlife species, although some of the ecological relationships we discovered between covariates and occupancy were sensitive to reductions in

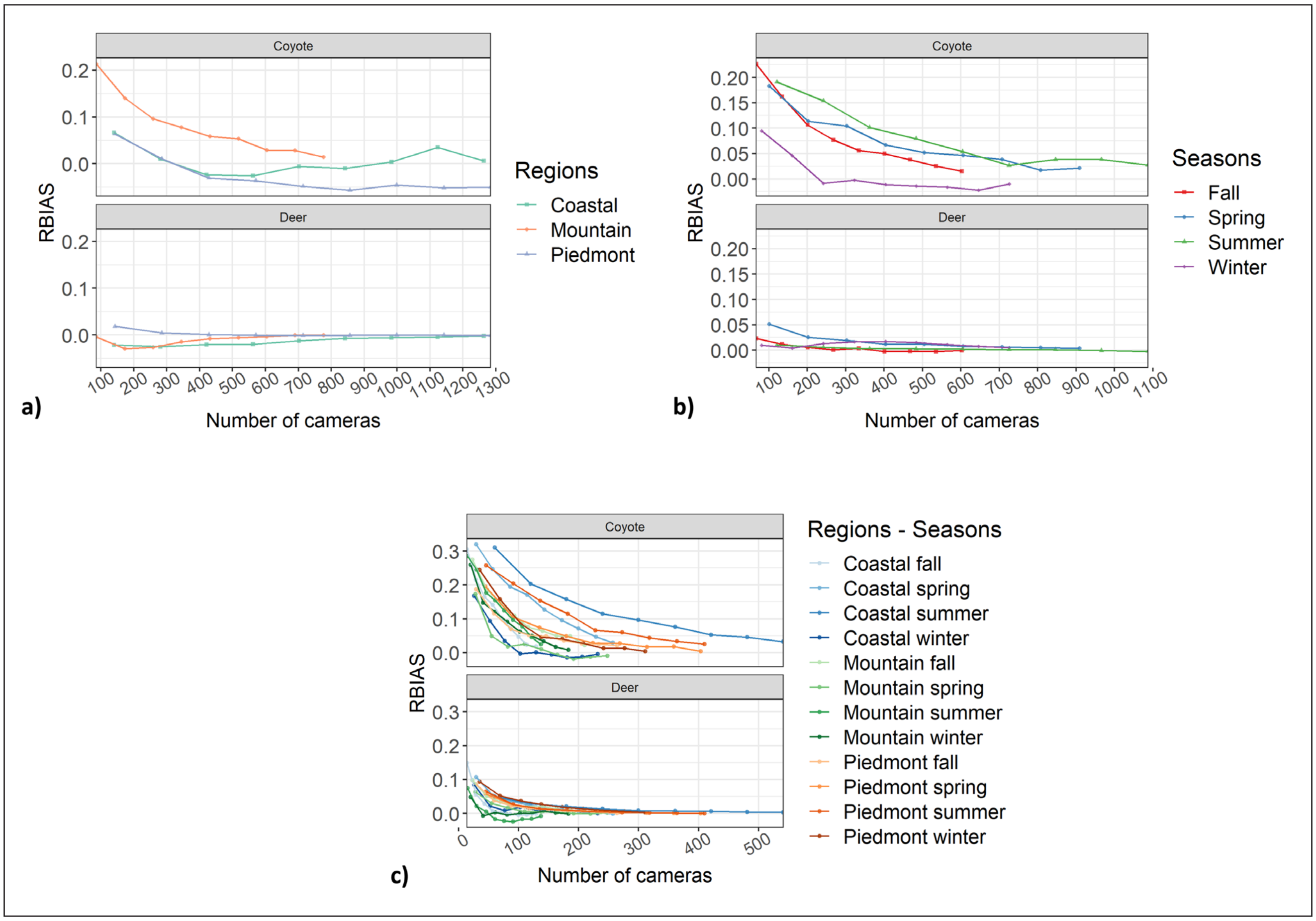

Figure 3 Graphs showing relative bias (a measure how different the estimates were from the full data set estimate) of occupancy estimates for white-tailed deer and coyote. These were calculated with subsets of the full NCCC dataset for **a)** three ecoregions, **b)** four seasons, and **c)** ecoregion-seasons. Our estimates reached our goal for bias (<0.1) at very small sample sizes for the common deer and after sampling 250–300 sites for the less common coyote across all spatio-temporal divisions. The lack of change in these estimates with increasing sample size also indicates a stable, robust result. Results were similar for estimates of error (Supplemental Figure 1).

sample size. Additionally, our subsetting exercise suggested these ecological relationships are variable over space and time (i.e., nonstationary).

This analysis shows that citizen scientists can obtain representative data at large scales and demonstrates how a hybrid study design can improve sample coverage. By having a priori sampling goals, we were able to encourage volunteers to sample in ways that gave us a more balanced sample, which was sufficient in the more populated piedmont ecoregion. This approach also would have obtained good coverage in the mountains and coastal ecoregions (92% and 98% coverage, respectively), but by supplementing this with data collected by staff, we were able to boost coverage to > 98% for both. This approach is also cost-effective over large areas (Lasky et al. 2021a), and provides the opportunity to sample on private land, which makes up 86% in the region (Butler and Wear 2013), but is difficult to sample without involving citizens. We think this Plan, Encourage, Supplement strategy could be useful to others designing citizen science projects, whether they are trying to get a representative sample of a region, as we did, or are targeting specific rare species or habitats.

As citizen science databases grow, they are more likely to be used in biodiversity modeling research projects. However, some question whether big unstructured, biodiversity data sets typical of citizen science actually mean more knowledge (Bayraktarov et al. 2019). This is less of a concern for sensor-based citizen science projects, including camera traps, because the primary data come from the sensor, and is thus more structured. In the case of camera traps, this means that effort is automatically calculated (time in the field), data quality can be checked (by examining photographs), and the likely absence of a species can be inferred (from lack of pictures). However, even with accurate measures from the sensors, there is still a risk of spatial bias based on where the sensors are actually deployed by citizens, especially if site choice is opportunistic (Weiser et al. 2020). Our approach for detecting spatial bias in relation to seven dimensions of

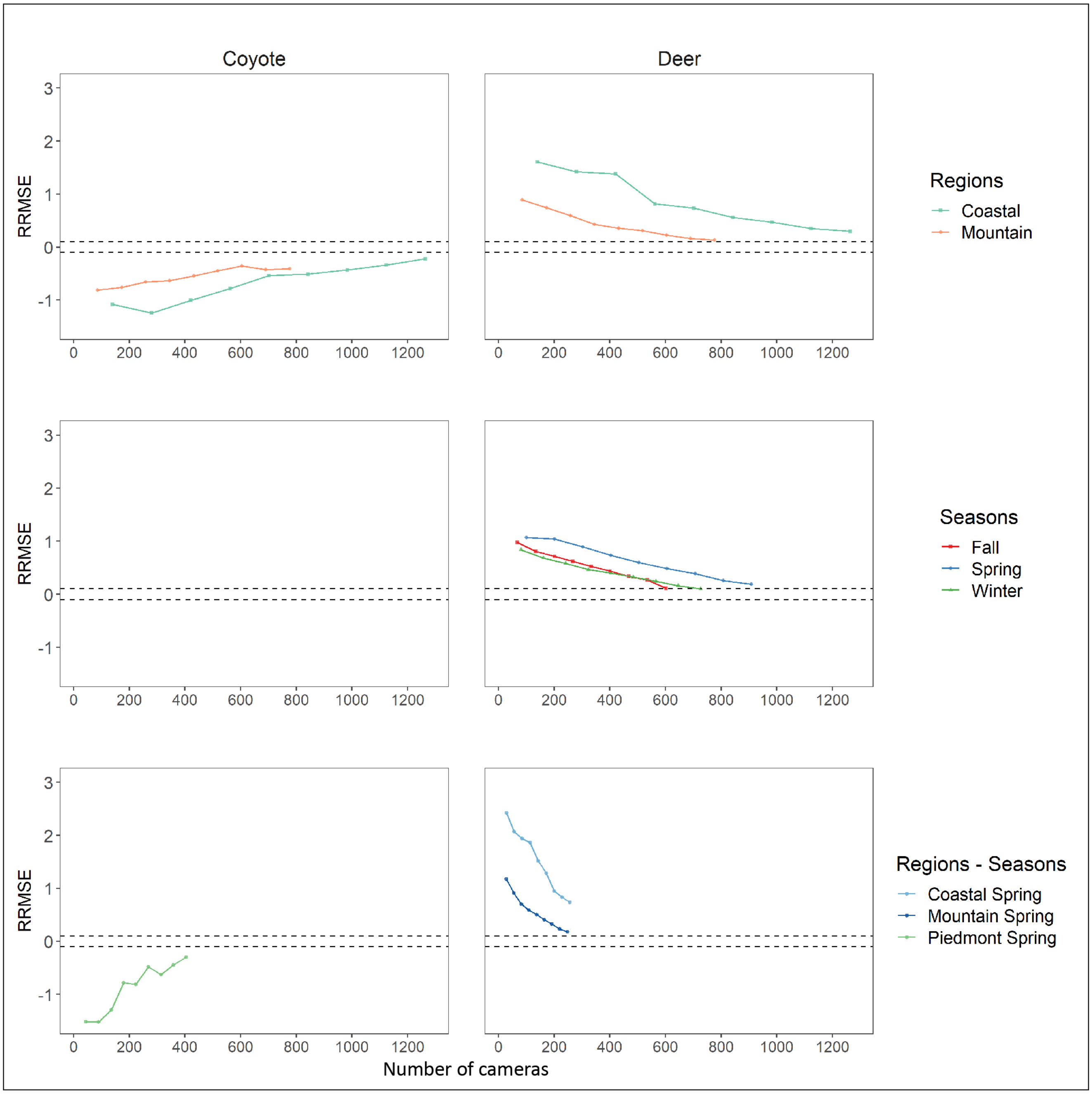

Figure 4 Graphs showing the changes in the relative error (relative root mean square error [RRMSE], a measure of how variable the estimates of occupancy were across the replicates) with larger sample size for estimated ecological relationships for agriculture land cover in occupancy models for coyotes and deer. Models were run across regions (*top*), seasons (*middle*), and regions-seasons (*bottom*). Only significant model effects are shown. Error estimates approached our 10% goal more rapidly with more restricted models (i.e., region-season) suggesting spatio-temporal variability in these relationships added variation to the larger-scale models. Full results for changes in error and bias of all covariates are available in Supplemental Figure 2.

habitat is unique, and the results provide more confidence in the utility of this data set specifically, and the potential for the Plan, Encourage, Supplement approach in general.

Our decision to categorize the habitat variation seen across the state into ten bins, and then check if each of those categories was adequately sampled, was somewhat arbitrary. Spreading the variation into more categories would certainly result in fewer being adequately sampled by our scheme. However, we felt that having each category represent 10% of the variation observed in the state was a biologically relevant measure. We also recognize that there are other more fine-scale ways to classify habitat types, for example, recognizing categorical difference in forest type based on tree-species composition. However, categorical fine-scale habitat classifications are not often used in large-scale ecological models, because most types are absent from a given area, requiring exhaustive sampling to get coverage. Such an analysis would undoubtably uncover that some rare habitat types were under sampled. In addition, there are some habitats that are not realistic to sample with camera traps set by volunteers or professionals (e.g., cliffs, pocosins, swamps).

We used occupancy models as our gauge of how sensitive ecological results from this data set would be to changes in sample size, because these are one of the most common metrics derived from this type of data. To allow for spatial and temporal variation in these relationships, we conducted separate analyses for three ecoregions, four seasons, and also twelve region-season combinations. Bias and error dropped below our goal of 0.1 with relatively small sample sizes for occupancy estimates. However, the precision of our estimation of ecological relationships was less robust. Interestingly, there was quite a lot of variability in which factors were significant depending on the spatiotemporal grouping of the data, suggesting substantial local differences in the ecology of these species (nonstationarity). For example, agriculture was a significant predictor of deer distribution in the costal and mountain regions but not the piedmont, and in the fall, winter, and spring, but not in the summer (Supplemental Figure 2). Likewise, the more restricted region-season model reached lower error with fewer samples than the larger-scale models, further suggesting nonstationary ecological relationships. These results show how larger data sets can enable more detailed research questions (e.g., seasonal or geographic differences in ecology of a species) and in our case, still estimate occupancies accurately, although with less confidence in specific ecological relationships. Although more data could help untangle these ecological nuances more precisely, this density of data could also be useful for more sophisticated spatiotemporal modeling approaches designed to accommodate this type of dynamic (Meehan, Michel, and Rue 2019).

In summary, we showed that sensor-based citizen science projects can obtain a robust sample size that is representative of habitat variation across large scales. We implemented the dynamic incentives strategy of Callaghan et al. (2019) as a three-part Plan, Encourage, Supplement strategy that can be useful for other citizen science projects, and we think that that our approach for assessing the representativeness of opportunistic samples could be applied to other data sets before using them in modeling exercises.

DATA ACCESSIBILTY STATEMENT

The camera trap data used in these analyses is available in (Lasky et al. 2021b).

SUPPLEMENTARY FILES

The supplementary files for this article can be found as follows:

- **Supplemental File 1: Supplemental Figure 1.** Graphs showing relative error of occupancy estimates for white-tailed deer and coyote calculated with subsets of the full NCCC data set. DOI: *https://doi.org/10.5334/cstp.344.s1*
- **Supplemental File 2: Supplemental Figure 2.** Graphs showing changes in relative error and relative bias in estimating ecological relationships with occupancy models using different sample sizes for coyotes and white-tailed deer. DOI: *https://doi.org/10.5334/cstp.344.s2*
- **Supplemental File 3: Appendix 1.** Worksheets showing sample size results for all ecoregion and habitat type combinations including total and just those set by volunteers or staff. DOI: *https://doi.org/10.5334/cstp.344.s3*

ACKNOWLEDGEMENTS

We would like to thank all 580 citizen science volunteers and numerous project interns for their contribution of time, effort, and data to the NC Candid Critters Project. We would also like to thank NC Public Libraries and NC Cardinal, in particular J. Pippin, for helping us distribute gear to volunteers. Thanks to Jonathan Shaw for a review of the methodology. This project would not have been possible without the assistance of NC Wildlife Resources Commission employees and resources to help NC Candid Critters set camera traps during Fall Fawn Frenzy. We sincerely thank NC State Parks, the US National Park Service, the US Fish and Wildlife Service, the US Forest Service, and numerous other public land agencies for providing us access to their lands. Finally, we thank the NC Museum of Natural Sciences and NC State University for their continual support.

FUNDING INFORMATION

The NCCC project was funded by a grant from the North Carolina Wildlife Resources Commission.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

Roland Kays led the NCCC project and the conceptual development and writing of the paper. Monica Lasky managed volunteers for the NCCC project and oversaw data management. Brent Pease, Arielle Parsons, and Krishna Pacifici conducted the analyses. All authors contributed critical manuscript review and revisions.

AUTHOR AFFILIATIONS

Roland Kays *orcid.org/0000-0002-2947-6665*
North Carolina Museum of Natural Sciences, US

Monica Lasky *orcid.org/0000-0002-9567-4643*
North Carolina Museum of Natural Sciences, US

Arielle W. Parsons *orcid.org/0000-0003-1076-2896*
North Carolina Museum of Natural Sciences, US

Brent Pease *orcid.org/0000-0003-1528-6075*
North Carolina State University, US

Krishna Pacifici *orcid.org/0000-0002-7518-7186*
North Carolina State University, US

REFERENCES

Bayraktarov, E, Ehmke, G, O'Connor, J, Burns, EL, Nguyen, HA, McRae, L, Possingham, HP and **Lindenmayer, DB.** 2019. Do big unstructured biodiversity data mean more knowledge? *Frontiers in Ecology and Evolution*, 6: 239. DOI: *https://doi.org/10.3389/fevo.2018.00239*

Bird, TJ, Bates, AE, Lefcheck, JS, Hill, NA, Thomson, RJ, Edgar, GJ, Stuart-Smith, RD, Wotherspoon, S, Krkosek, M, Stuart-Smith, JF, Pecl, GT, Barrett, N and **Frusher, S.** 2014. Statistical solutions for error and bias in global citizen science datasets. *Biological Conservation*. DOI: *https://doi.org/10.1016/j.biocon.2013.07.037*

Butler, BJ and **Wear, DN.** 2013. Forest ownership dynamics of southern forests. In: Wear, DN, Greis, JG, (eds.) 2013. *The Southern Forest Futures Project: technical report. Gen. Tech. Rep. SRS-GTR-178.* Asheville, NC: USDA-Forest Service, Southern Research Station, 178: 103–121.

Callaghan, CT, Roberts, JD, Poore, AGB, Alford, RA, Cogger, H and **Rowley, JJL.** 2020. Citizen science data accurately predicts expert-derived species richness at a continental scale when sampling thresholds are met. *Biodiversity and Conservation*, 29(4): 1323–1337. DOI: *https://doi.org/10.1007/s10531-020-01937-3*

Callaghan, CT, Rowley, JJL, Cornwell, WK, Poore, AGB and **Major, RE.** 2019. Improving big citizen science data: Moving beyond haphazard sampling. *PLOS Biology*, 17(6): e3000357. DOI: *https://doi.org/10.1371/journal.pbio.3000357*

Crall, AW, Newman, GJ, Stohlgren, TJ, Holfelder, KA, Graham, J and **Waller, DM.** 2011. Assessing citizen science data quality: An invasive species case study. *Conservation Letters*, 4(6): 433–442. DOI: *https://doi.org/10.1111/j.1755-263X.2011.00196.x*

Danielsen, F, Pirhofer-Walzl, K, Adrian, TP, Kapijimpanga, DR, Burgess, ND, Jensen, PM, Bonney, R, Funder, M, Landa, A, Levermann, N and **Madsen, J.** 2014. Linking Public Participation in Scientific Research to the Indicators and Needs of International Environmental Agreements. *Conservation Letters*, 7(1): 12–24. DOI: *https://doi.org/10.1111/conl.12024*

Dee, DP, Uppala, SM, Simmons, AJ, Berrisford, P, Poli, P, Kobayashi, S, Andrae, U, Balmaseda, MA, Balsamo, G, Bauer, P, Bechtold, P, Beljaars, ACM, van de Berg, L, Bidlot, J, Bormann, N, Delsol, C, Dragani, R, Fuentes, M, Geer, AJ, Haimberger, L, Healy, SB, Hersbach, H, Hólm, E V, Isaksen, L, Kållberg, P, Köhler, M, Matricardi, M, Mcnally, AP, Monge-Sanz, BM, Morcrette, JJ, Park, BK, Peubey, C, de Rosnay, P, Tavolato, C, Thépaut, JN and **Vitart, F.** 2011. The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*. DOI: *https://doi.org/10.1002/qj.828*

Dickinson, JL, Zuckerberg, B and **Bonter, DN.** 2010. Citizen Science as an Ecological Research Tool: Challenges and Benefits. *Annual Review of Ecology, Evolution, and Systematics*, 41: 149–172. DOI: *https://doi.org/10.1146/annurev-ecolsys-102209-144636*

Diggle, PJ, Menezes, R and **Su, T.** 2010. Geostatistical analysis under preferential sampling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*. DOI: *https://doi.org/10.1111/j.1467-9876.2009.00701.x*

Hammer, RB, Stewart, SI, Winkler, R, Radeloff, VC and **Voss, PR.** 2004. Characterizing spatial and temporal residential density patterns across the U.S. Midwest, 1940–1990. *Landscape and Urban Planning*, 69: 183–199. DOI: *https://doi.org/10.1016/j.landurbplan.2003.08.011*

Hansen, MC, Potapov, P V, Moore, R, Hancher, M, Turubanova, SA, Tyukavina, A, Thau, D, Stehman, SV, Goetz, SJ, Loveland, TR, Kommareddy, A, Egorov, A, Chini, L, Justice, CO and **Townshend, JRG.** 2013. High-resolution global maps of 21st-century forest cover change. *Science*, 342: 850–853. DOI: *https://doi.org/10.1126/science.1244693*

Homer, C, Dewitz, J, Yang, L, Jin, S, Danielson, P, Xian, G, Coulston, J, Herold, N, Wickham, J and **Megown, K.** 2015. Completion of the 2011 National Land Cover Database for the conterminous United States–representing a decade of land cover change information. *Photogrammetric Engineering & Remote Sensing*, 81(5): 345–354.

Hunter, J, Alabri, A and **Van Ingen, C.** 2013. Assessing the quality and trustworthiness of citizen science data. In:

Concurrency Computation Practice and Experience. 2013 John Wiley & Sons, Ltd. 454–466. DOI: *https://doi.org/10.1002/cpe.2923*

Isaac, NJB, van Strien, AJ, August, TA, de Zeeuw, MP and **Roy, DB.** 2014. Statistics for citizen science: Extracting signals of change from noisy ecological data. *Methods in Ecology and Evolution*. DOI: *https://doi.org/10.1111/2041-210X.12254*

Johnston, A, Moran, N, Musgrove, A, Fink, D and **Baillie, SR.** 2020. Estimating species distributions from spatially biased citizen science data. *Ecological Modelling*, 422: 108927. DOI: *https://doi.org/10.1016/j.ecolmodel.2019.108927*

Kays, R, Arbogast, BS, Baker-Whatton, M, Beirne, C, Boone, HM, Bowler, M, Burneo, SF, Cove, M V, Ding, P, Espinosa, S, Luis Sousa Gonçalves, A, Hansen, CP, Jansen, PA, Kolowski, JM, Knowles, TW, Guimarães Moreira Lima, M, Millspaugh, J, McShea, WJ, Pacifici, K, Parsons, AW, Pease, BS, Rovero, F, Santos, F, Schuttler, SG, Sheil, D, Si, X, Snider, M and **Spironello, WR.** 2020. An empirical evaluation of camera trap study design: how many, how long, and when? *Methods in Ecology and Evolution*, 11: 700–713. DOI: *https://doi.org/10.1111/2041-210X.13370*

Kosmala, M, Wiggins, A, Swanson, A and **Simmons, B.** 2016. Assessing data quality in citizen science. *Frontiers in Ecology and the Environment*, 14(10): 551–560. DOI: *https://doi.org/10.1002/fee.1436*

Lasky, M, Parsons, A, Schuttler, S, Mash, A, Larson, L, Norton, B, Pease, B, Boone, H, Gatens, L and **Kays, R.** 2021a. Candid Critters : Challenges and Solutions in a Large-Scale Citizen Science Camera Trap Project. *Citizen Science: Theory and Practice*, 6(1): 1–17. DOI: *https://doi.org/10.5334/cstp.343*

Lasky, M, Parsons, AW, Schuttler, SG, Hess, G, Sutherland, R, Kalies, L, Clark, S, Olfenbuttel, C, Matthews, J, Davis, G, McShea, WJ, Shaw, J, Dukes, C, Hill, J and **Kays, R.** 2021b. CAROLINA CRITTERS: a collection of camera trap data from wildlife surveys across North Carolina. *Ecology*. e03372. DOI: *https://doi.org/10.1002/ecy.3372*

MacKenzie, DI, Nichols, JD, Lachman, GB, Droege, S, Royle, JA and **Langtimm, CA.** 2002. Estimating site occupancy rates when detection probabilities are less than one. *Ecology*, 83: 2248–2255. DOI: *https://doi.org/10.1890/0012-9658(2002)083[2248:ESORWD]2.0.CO;2*

Meehan, TD, Michel, NL and **Rue, H.** 2019. Spatial modeling of Audubon Christmas Bird Counts reveals fine-scale patterns and drivers of relative abundance trends. *Ecosphere*. DOI: *https://doi.org/10.1002/ecs2.2707*

Mesinger, F, DiMego, G, Kalnay, E, Mitchell, K, Shafran, PC, Ebisuzaki, W, Jović, D, Woollen, J, Rogers, E, Berbery, EH, Ek, MB, Fan, Y, Grumbine, R, Higgins, W, Li, H, Lin, Y, Manikin, G, Parrish, D and **Shi, W.** 2006. North American regional reanalysis. *Bulletin of the American Meteorological Society*. DOI: *https://doi.org/10.1175/BAMS-87-3-343*

Millar, EE, Hazell, EC and **Melles, SJ.** 2019. The 'cottage effect' in citizen science? Spatial bias in aquatic monitoring programs. *International Journal of Geographical Information Science*, 33(8): 1612–1632. DOI: *https://doi.org/10.1080/13658816.2018.1423686*

Petrovan, SO, Vale, CG and **Sillero, N.** 2020. Using citizen science in road surveys for large-scale amphibian monitoring: are biased data representative for species distribution? *Biodiversity and Conservation*, 1–15. DOI: *https://doi.org/10.1007/s10531-020-01956-0*

Weiser, EL, Diffendorfer, JE, Lopez-Hoffman, L, Semmens, D and **Thogmartin, WE.** 2020. Challenges for leveraging citizen science to support statistically robust monitoring programs. *Biological Conservation*, 242. DOI: *https://doi.org/10.1016/j.biocon.2020.108411*

TO CITE THIS ARTICLE:
Kays, R, Lasky, M, Parsons, AW, Pease, B and Pacifici, K. 2021. Evaluation of the Spatial Biases and Sample Size of a Statewide Citizen Science Project. *Citizen Science: Theory and Practice,* 6(1): 34, pp. 1–12. DOI: *https://doi.org/10.5334/cstp.344*

Submitted: 03 June 2020 Accepted: 14 September 2021 Published: 16 December 2021

Citizen Science: Theory and Practice is a peer-reviewed open access journal published by Ubiquity Press.