Causes and consequences of non-random drop-outs for citizen science projects: lessons from the North American amphibian monitoring program

David M. Marsh[1,3] **and Bradley J. Cosentino**[2,4]

[1]Department of Biology, Washington and Lee University, Lexington, Virginia 22450 USA
[2]Department of Biology, Hobart and William Smith Colleges, Geneva, New York 14456 USA

Abstract: Citizen science holds great promise for collecting useful environmental data over large spatial scales. However, statistical issues that arise in the analysis of citizen science data may be relatively unfamiliar to scientists accustomed to data collected with traditional research methods. In particular, citizen science projects are often designed with standard randomization procedures, but volunteers may drop-out of a project in a highly non-random manner. For example, if volunteers are less likely to continue monitoring sites that are highly urbanized or polluted, these sites will be under-represented in analyses, and observed patterns could be biased accordingly. We tested for non-random drop-outs in the context of the North American Amphibian Monitoring Program (NAAMP), a road-based, citizen-science survey of calling frogs and toads. We found that discontinuation of survey routes by NAAMP volunteers was associated with high traffic volume, high noise levels, and low forest cover along these routes. The absolute increase in probability of dropping out of the program that was associated with these factors was often low (e.g., 2–10%), but much larger increases in drop-out probabilities (e.g., 40–70%) were predicted when traffic or noise were particularly high or when multiple factors were considered simultaneously. In addition, analysis of amphibian count data suggested that relatively low counts of amphibian and low species richness were also associated with increased probability that survey routes would be discontinued. Together, these non-random drop-outs led to the decreased representation of highly urbanized sites in our data set, and may have altered the estimated relationships between explanatory variables (e.g., traffic, forest cover) and amphibian species richness. Our results, therefore, suggest that citizen science projects need to be designed after careful consideration of the factors that promote retention of volunteers and the effects that non-random drop-outs may have on the data they generate. Stratification that takes non-random drop-outs into account may be necessary to ensure adequate representation of some kinds of survey sites in citizen science projects.

Key words: frogs, anuran, call survey, bias, volunteer, data analysis, data quality, NAAMP, statistics, toad

Citizen science projects increasingly generate high volume data sets over broad geographic areas (Dickinson et al. 2010). Large-scale citizen science projects have focused on issues as wide-ranging as bird distributions (e.g., Link and Sauer 1998) and plant responses to climate change (Chung et al. 2011). Many citizen science projects are also highly relevant to issues of water quality and aquatic habitat conservation. For example, FreshWater Watch (https://freshwater watch.thewaterhub.org/) is a global network of citizen scientists who collect water quality data such as pH, nitrate and phosphate levels, and turbidity at sites of their choosing. MiniSASS (http://www.minisass.org/en/) provides a general protocol for monitoring stream health through aquatic arthropod communities and is being applied across southern Africa. At a smaller scale, Florida Lakewatch (http://lakewatch.ifas.ufl.edu/how.shtml) coordinates water chemistry monitoring throughout the state, and Virginia Save Our Streams (http://www.vasos.org/) monitors the species diver-

E-mail addresses: [3]marshd@wlu.edu; [4]cosentino@hws.edu

DOI: 10.1086/701672. Received 30 June 2017; Accepted 22 August 2018; Published online 26 March 2019.
Freshwater Science. 2019. 38(2):292–302. © 2019 by The Society for Freshwater Science.

sity of aquatic insects as a measure of stream quality and biodiversity. Similar programs exist throughout North America and Europe and are expanding in other parts of the world.

From a scientific perspective, the potential for generating large data sets is one of the most exciting aspects of citizen science. However, the benefits of 'big data' are often tempered by issues of data quality (Hochachka et al. 2012, Kosmala et al. 2016). A number of studies have evaluated the quality of data generated by citizen science projects (Engel and Voshell 2002, Genet and Sargent 2003, Delaney et al. 2008, Crall et al. 2011, Kremen et al. 2011). Most of these studies focus on the data collection process (e.g., volunteers' ability to sample and identify species of interest), which volunteers may either perform well (e.g., Delaney et al. 2008, Crall et al. 2011) or poorly (McClintock et al. 2010, Miller et al. 2012). However, data quality may also be constrained by the project design. For example, citizen science projects are often faced with a trade-off between optimizing the study design and facilitating volunteer participation (Dickinson et al. 2010). In particular, randomization of study sites is not always feasible for citizen science projects. Volunteers may be more amenable to monitoring a stream they know well than one that is randomly assigned to them. Similarly, volunteers may be more interested in water quality near their own home than at a randomly generated site. For this reason, some citizen science projects forego randomization and allow volunteers to choose their own study locations. In these cases, it is necessary to estimate and correct for bias in sampling locations before making general inferences about species distributions, habitat conditions, or water quality across a landscape (Bird et al. 2014, Isaac et al. 2014).

To allow for more direct inference, some citizen science projects do attempt to randomize study locations as in traditional research designs. For example, in the UK's National Amphibian and Reptile Recording Scheme (NARRS; http://www.narrs.org.uk/), volunteers are assigned a random 1-km^2 square plot from within a 5- × 5-km grid centered on their residence. Within the selected square, volunteers monitor the pond closest to the SW corner to reduce selection bias. Similarly, the North American Amphibian Monitoring Program (NAAMP) starts with randomly-generated roadside survey routes. Coordinators in each state assign volunteers a route near their home, and volunteers identify precise sampling locations (i.e., water bodies) along this route. Thus, NAAMP survey routes can be thought of as a random selection of roadside habitats within the region.

However, even when survey locations are initially assigned at random, volunteer monitors may drop out of the project at any time. If these drop-outs are non-random (i.e., they are associated with monitoring site characteristics), substantial discrepancies may arise between the sample data and the broader area of interest. These drop-outs could, therefore, have several distinct effects on citizen science data. First, non-random drop-outs could lead to the under-representation of some site types in the final data set. For example, if volunteers tend to shy away from very urbanized or very polluted sites, these sites may end up under-represented. Second, non-random drop-outs could influence the observed relationships among variables within a data set. If the factors that affect retention of volunteers are the same factors that are being investigated in a study (e.g., urbanization, land use), predictor variables could be confounded with observation effort. Thus, there is a need to understand the factors that influence retention of volunteers as well as the potential effects of drop-outs on patterns within the data. The expansion of citizen science programs has exacerbated this need.

In 2013, we began working with NAAMP data in a collaborative project to determine the relationships between pond-breeding amphibian distributions and land use across the Eastern and Central United States (Cosentino et al. 2014, Marsh et al. 2017). We were surprised to discover a weak but non-zero relationship in these data between the number of times a site had been surveyed and the amount of forest cover at that site. The NAAMP survey routes had been developed through randomization, so it was not clear to us why this relationship existed. In this paper, we re-analyze NAAMP data from 2 previously compiled survey route data sets and 1 newly compiled data set to test for non-random drop-outs by volunteers. Specifically, we ask what land cover variables (i.e., forest cover, agricultural cover, developed cover, wetland area, road density, traffic volume, and noise level) influence volunteer retention in NAAMP beyond 1 or 2 y. We also ask whether initial frog detections affect volunteer retention. We then analyze the effects that non-random drop-outs have on the distributions of these explanatory variables and on their apparent relationships with observed amphibian species richness. Finally, we outline strategies for either reducing the frequency of non-random drop-outs in citizen science projects or building non-random drop-outs into the project design itself.

METHODS

NAAMP surveys

NAAMP is a citizen-science monitoring initiative organized by the U.S. Geological Survey (Weir and Mossman 2005) that ran from 1997 to 2015 across most states in the Eastern and Central United States. NAAMP was based on night-time surveys for calling anurans (frogs and toads), which were monitored from roadside survey locations. NAAMP volunteers were assigned randomly-generated driving routes that were as close as possible to their residence (Weir and Mossman 2005). Observers initially traversed routes during the daytime to select 10 sampling locations (stops). Stops were located at least 0.8 km apart at sites where bodies of water were visible within 200 m of the road. The route surveys were carried out after dark in time windows that spanned the breeding season of anurans in the region. During a survey, observers would get out of their cars at each stop and record the number of anuran calls heard over a 5-minute survey period. In addition, observers would record the number of cars passing by on the road and whether or not background noise may have interfered with their observations. Surveys were usually carried out 3× /y, though occasionally they were carried out more or less often at the discretion of the volunteer.

NAAMP volunteers were trained by state coordinators in data collection and species identification. Audio files of local species were supplied to volunteers, along with information on which species were likely to be heard. Beginning in 2006, new volunteers were required to complete an online 'Frog Quiz' that required them to successfully identify species from recorded calls.

Data compilation

Our analyses relating survey discontinuation to route characteristics are based on 3 data sets. The 1st and 2nd data sets were compiled in 2013 and 2014, respectively, in conjunction with previous analyses of NAAMP data (Cosentino et al. 2014, Marsh et al. 2017). These two data sets are structured differently—in the 1st (hereafter 'nested data') multiple survey stops are sampled from each survey route, whereas in the 2nd ('non-nested data'), only one stop is sampled from each route. Analyses of these two data sets can be viewed as replications that ensure conclusions are not specific to a single approach for compiling NAAMP data. The 3rd data set was compiled specifically for the analysis of the effects of frog detections during volunteers' initial surveys, as initial detections had not been recorded in our previous analyses. Methods for compiling these data sets are summarized below, and the data themselves are available as Supplemental Materials.

In 2013, we created the nested data set by compiling anuran data from multiple stops within each of 406 NAAMP survey routes from 13 states: Florida (FL), Massachusetts (MA), Minnesota (MN), Missouri (MO), North Carolina (NC), New Hampshire (NH), New York (NY), Pennsylvania (PA), South Carolina (SC), Texas (TX), Virginia (VA), Vermont (VT), and West Virginia (WV). These surveys were conducted between 1997 and 2012. Routes were classified by region as 'North' or 'South' as these regions tend to have different land uses and different anuran species composition. We also characterized landscape structure within 1-km buffers around each survey stop. To avoid spatial overlap in landscape buffers among stops, we compiled data only for stops 1, 4, 7, and 10 within each route, which ensured that most stops were ≥2 km apart. For each stop, we calculated total number of surveys, proportion of surveys in which interfering noise was recorded ('noise level'), presence or absence of each species across all surveys, total number of species detected across all surveys (i.e., species richness), and mean number of cars passing by during the 5-m anuran counts ('traffic volume'). In practice, noise level and traffic volume were highly correlated. In some analyses it was possible to distinguish the effects of these variables, whereas in others we were not able to disentangle their effects. Surveys encompassed different time periods for different stops, because observers started and ended their participation in NAAMP in different years, or in some cases observers changed without survey interruption. The median number of surveys per stop in our data set was 12 (i.e., 4 y; range = 1–45 surveys).

We used qGIS (version 1.8; QGIS Geographic Information System, Open Source Geospatial Foundation Project) or ArcGIS (version 10.2, Environmental Systems Research Institute, Redlands, California) software to characterize landscape conditions present at each NAAMP stop. Landscape analysis was based on spatial data we imported from the National Land Cover Database (NLCD; Fry et al. 2011), the National Wetlands Inventory (NWI; US Department of the Interior 2013), and the TIGER road database (US Census Bureau 2013). We used these layers to calculate the following variables for 1-km buffers around each NAAMP stop: proportion of land that was forested, agricultural, and developed; total wetland area; and total linear road length ('road density'). To make these data comparable to those from the non-nested data set (see below), all variables for the 4 stops within each route were averaged to yield 1 summary value for each route. Landcover variables were inversely correlated (e.g., more forest cover necessarily means less agricultural cover), so we avoided constructing multivariate models that included pairs of highly correlated landcover variables.

In 2014, we created the non-nested data set by selecting 1 random NAAMP stop within each of 567 survey routes in the same 13 states as above. These surveys spanned the years 1997 to 2013. In the non-nested data, routes were classified into three regions (North, South, and Midwest) as the addition of routes from MO and MN give us sufficient data to split off Midwestern routes. Landscape metrics for these stops were compiled as in the previous data set. However, since we used only 1 stop for each route, we did not average landscape metrics across stops. Overlap between the stops in the nested and non-nested data was ~30%, so analyses of these data sets were largely independent.

The nested and non-nested data sets were used to assess if route discontinuation was related to landcover variables. However, we also hypothesized that route discontinuation might be affected by the number of frogs detected during the 1st few years of surveys by each volunteer. In the prior data compilations we had not distinguished these early surveys from later surveys. We, therefore, compiled a 3rd data set by extracting data from all NAAMP routes in FL, MO, VA, VT, and WV spanning the years 2002 to 2013 (203 routes total). For each route we determined the total number of recorded anuran detections per survey for each of the first 2 y of surveys.

Predictors of NAAMP route discontinuation

In a prior study (Cosentino et al. 2014), we found a positive relationship between survey effort and observed species richness adjusted for net primary productivity ($r_p = 0.21$) that was most pronounced for routes with <9 surveys (typically <3 y of data, Fig. 1). For sites that had ≥9 surveys, the correlation between survey number and observed richness was <0.10. Routes with <9 surveys appeared to underestimate species richness, so we sought to identify why some volunteers stop their routes after only 1 or 2 y. Two thresholds were used to classify a route as continued or discontinued: 1) whether routes were surveyed for >1 y, or 2) whether routes were surveyed for >2 y. These thresholds were used to classify routes for both the nested and the non-nested data sets. Our rationale for classifying routes as continued based on whether >1 y of surveys were completed is that these routes may be most indicative of the factors that cause volunteers to end their participation. Our rationale for classifying based on whether >2 y of surveys were completed was that our prior analyses suggested that >2 y of data were required to produce reliable estimates of species richness on a route. These 2 classification schemes yielded similar results, and we present both below for comparison.

Univariate relationships with route discontinuation

Once routes were classified as discontinued or continued, we used logistic regression models on each data set to determine which factors predicted route discontinuation. We examined: 1) geographic region, 2) proportion of the landscape within 1 km of survey stops that was forested, agricultural, wetland, or developed, 3) road density within 1 km, 4) mean traffic volume during surveys, and 5) proportion of total surveys with noise interference ('noise level').

We first evaluated each of these factors individually to assess whether they were related to the probability of route discontinuation. We had noted a positive relationship between forest cover and survey continuation during our previous analysis of NAAMP data, so results with respect to forest cover for the nested data should not be viewed as an a priori hypothesis test. However, our analysis of the role of forest cover in the non-nested data should represent a valid test of this relationship.

Multivariate relationships with route discontinuation

Once we analyzed each variable individually, factors that were consistently related to survey discontinuation were entered into a single multivariate logistic regression model for each data set/discontinuation threshold to evaluate their combined influence on survey continuation. To keep these models simple, we only included variables that were significantly related to survey continuation in all 4 univariate analyses (2 discontinuation thresholds and 2 data sets). This approach is perhaps overly conservative with respect to variable selection, but our primary goal was to examine the magnitude of effects of individual variables, not necessarily to identify the best overall model. We used calculations of logit probabilities to estimate the effect size of each variable, and used McFadden's pseudo-R^2 to evaluate model fit (McFadden 1974). Values of pseudo-$R^2 > {\sim}0.20$ represent a good model fit (McFadden 1974).

We also hypothesized that survey discontinuation might be influenced by the relative number of frogs detected on any

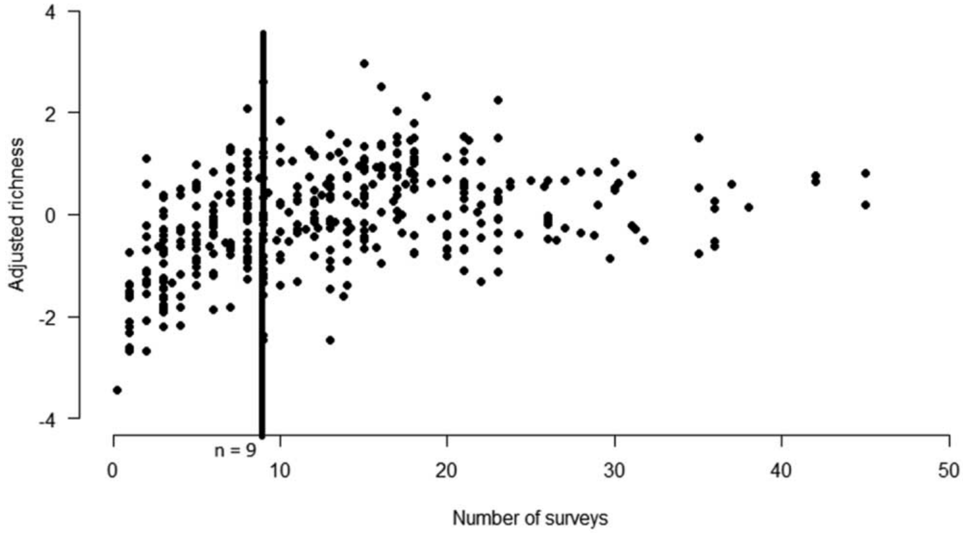

Figure 1. Scatterplot for the relationship between observed anuran species richness and number of times each route was surveyed. Richness values (y-axis) show residual richness after net primary productivity (NPP) was accounted for, which represents the number of species observed relative to the number expected based on NPP. The vertical line shows 3 y of surveys, after which the relationship between species detected and survey number was less pronounced.

particular route. Specifically, when volunteers detect fewer frog calls in their initial surveys, they might be less interested in continuing their routes in subsequent years. To evaluate this possibility, we created a 3rd data set by extracting data from the 1st year of surveys for a subset of states: FL, MO, VA, VT, and WV. These states were chosen because they spanned the different regions and included both continued and discontinued routes. For each route we calculated the total number of frog detections recorded per survey, irrespective of species. We then asked whether the number of detections per survey was related to whether a survey route was continued beyond the 1st year. We performed this analysis with and without covariates for route characteristics (e.g., traffic volume, noise) that could have separately influenced survey continuation.

Direct and indirect effects on route discontinuation

We used a structural equation model (SEM) (Grace 2006, Kline 2015) to further examine the effects of urbanization, landcover, traffic, noise, and frog richness on survey continuation. SEMs allow the use of latent variables to represent theoretical constructs that are estimated by measured variables (Grace et al. 2010, Kline 2015). For example, we represented the latent variable 'urbanization' by 2 measured variables: proportion of developed land and total road length within 1 km. Other latent variables in our model included forest cover, traffic, noise, frog species richness, and route discontinuation. In our study, each of these latent variables was represented by single measured variables, but in principle additional measured variables might be related to these latent variables.

SEMs are particularly useful for understanding systems in which there is a network of direct and indirect effects. In our system, route discontinuation could be directly related to any of the latent variables in the model, but some pathways may be indirect. For example, urbanization could directly influence route discontinuation because urban areas have a greater pool of potential volunteers to survey routes (so that volunteers could be replaced without discontinuing the route). Alternatively, urbanization could indirectly affect the probability of route discontinuation by affecting the amount of habitat in the landscape, the degree of traffic and noise during surveys, or the number of frog species on a route. We used the *lavaan* package in R (R Core Development Team 2016) to fit the SEM with diagonally weighted least squares estimation and robust standard errors (Rosseel 2012). Traffic volume and road length were divided by a constant to put them on a similar scale to other variables (Rosseel 2012). We used standardized regression coefficients to compare the strength and significance of pathways in the model. We assessed model fit with a χ^2 test and root mean square error of approximation (RMSEA). Good model fit is indicated by $p > 0.05$ and RMSEA <0.05 (Kline 2015). We fit the SEM to the non-nested data set rather than the nested data set because the non-nested data set included ~30% more routes.

Effects of route discontinuation on observed patterns in NAAMP data

To assess if route discontinuation affected patterns in NAAMP data, we conducted 2 sets of analyses. First, we examined the statistical distribution of important explanatory variables when we included all survey routes vs when we restricted data to routes with ≥3 y of surveys (i.e., routes that were included in Cosentino et al. 2014 and Marsh et al. 2017). In particular, we wanted to know whether highly urbanized sites were being lost because volunteers assigned these sites were less likely to keep surveying them. For this analysis, we compared the frequency of high end values (i.e., traffic volume >5.0 cars/survey, noise level >0.5, forest cover <10%) for explanatory variables in the full sample of routes to frequencies of these values in the sample of routes surveyed for at least 3 y. Second, we examined how removal of routes with only 1 or 2 y of data changed the observed correlations between landscape variables and observed species richness. That is, we asked to what extent route discontinuation appeared to alter the basic patterns observed in the data. All analyses were performed in R (version 3.3: R Project for Statistical Computing, Vienna, Austria) and raw data are available as Supplemental Materials.

RESULTS

Univariate relationships with survey discontinuation

Nested data Region, forest cover, wetland area, traffic volume, and noise level each was associated with the probability that a route would be discontinued after a single year (Table 1). The discontinuation probability for sites in the northern region was 8%, compared with 22% in the southern region. High forest cover was negatively associated with the probability of survey discontinuation, whereas high wetland area, traffic volume, and noise level were positively associated with survey discontinuation (Fig. S1). These associations were generally weak, and comparison of predicted discontinuation rates at 25th percentile values for each individual variable vs 75th percentile values yielded increases of only 2 to 5% in absolute probability of route discontinuation (Table 1). However, models predicted very high discontinuation probabilities near the maximum observed values for traffic volume (60%) and noise level (49%).

Results based on survey discontinuation after 2 y were similar to the results after 1 y (Fig. S2). Again, discontinuation was more likely in the southern region (48 vs 25% in the northern region), and was more likely at sites with low forest cover, high wetland area, high traffic volume, and high noise (Table 1). In this data set, developed land cover was also associated with a higher probability of route discontinuation. For these models, increases in variable values from the 25th

Table 1. Results from univariate logistic regressions of probability of discontinuing surveys on site characteristics. Results are shown for samples analyzed for 2 data sets with 2 different cut-offs for classifying survey routes as 'continued' (either surveyed >1 or >2 y). For each model, we show the regression coefficient and its standard error (SE) for each association with route discontinuation. When parameters differ significantly from 0 (shown in bold), we also give the predicted probability of a volunteer discontinuing their route at the 25th, 75th, and maximum probability (max) percentile values for each variable. * indicates parameter for region in the nested data represents the relative effect of being in the southern vs the northern USA; † indicates 2 parameters for region the non-nested data represent the relative effects of being in the midwestern United States and the southern vs the northern USA.

						Probability at different percentiles		
Data set	Threshold	Variable	beta	SE	p	25th	75th	Max
Nested	**>1 y**	**Region***	**1.22**	**0.32**	**0.0001**	**NA**	**NA**	**NA**
Nested	**>1 y**	**Forest cover**	**−2.66**	**0.68**	**9×10^{-5}**	**0.26**	**0.24**	**0.29**
Nested	>1 y	Developed cover	0.44	1.22	0.72			
Nested	>1 y	Agricultural cover	0.60	0.58	0.30			
Nested	**>1 y**	**Wetland area**	**6.2×10^{-6}**	**2.4×10^{-7}**	**0.01**	**0.12**	**0.16**	**0.44**
Nested	>1 y	Road density	-1.1×10^{-5}	3.7×10^{-5}	0.76			
Nested	**>1 y**	**Traffic volume**	**0.065**	**0.022**	**0.003**	**0.11**	**0.13**	**0.60**
Nested	**>1 y**	**Noise Level**	**2.06**	**0.60**	**0.0005**	**0.11**	**0.16**	**0.49**
Nested	**>2 y**	**Region***	**1.01**	**0.22**	**4×10^{-6}**	**NA**	**NA**	**NA**
Nested	**>2 y**	**Forest cover**	**−1.64**	**0.45**	**0.0003**	**0.41**	**0.27**	**0.47**
Nested	**>2 y**	**Developed cover**	**2.29**	**0.94**	**0.02**	**0.33**	**0.37**	**0.75**
Nested	>2 y	Agricultural cover	0.037	0.44	0.93			
Nested	**>2 y**	**Wetland area**	**5.0×10^{-7}**	**2.0×10^{-7}**	**0.01**	**0.33**	**0.39**	**0.69**
Nested	>2 y	Road density	3.4×10^{-5}	2.6×10^{-5}	0.19			
Nested	**>2 y**	**Traffic volume**	**0.056**	**0.021**	**0.006**	**0.30**	**0.34**	**0.69**
Nested	**>2 y**	**Noise Level**	**1.59**	**0.52**	**0.002**	**0.32**	**0.40**	**0.69**
Non-nested	**>1 y**	**Region†**	**−0.02/1.63**	**0.45/0.35**	**0.97/2×10^{-6}**	**NA**	**NA**	**NA**
Non-nested	**>1 y**	**Forest cover**	**−1.26**	**0.53**	**0.02**	**0.14**	**0.04**	**0.16**
Non-nested	>1 y	Developed cover	−1.07	1.89	0.57			
Non-nested	>1 y	Agricultural cover	−0.21	0.56	0.56			
Non-nested	>1 y	Wetland cover	0.94	0.66	0.66			
Non-nested	>1 y	Road density	-3.3×10^{-6}	2.5×10^{-5}	0.89			
Non-nested	**>1 y**	**Traffic volume**	**0.07**	**0.02**	**0.001**	**0.08**	**0.10**	**0.66**
Non-nested	**>1 y**	**Noise Level**	**1.72**	**0.44**	**0.0001**	**0.08**	**0.11**	**0.33**
Non-nested	**>2 y**	**Region†**	**0.36/1.68**	**0.31/0.27**	**0.25/3×10^{-10}**	**NA**	**NA**	**NA**
Non-nested	**>2 y**	**Forest cover**	**−0.99**	**0.40**	**0.01**	**0.25**	**0.17**	**0.27**
Non-nested	>2 y	Developed cover	1.24	1.10	0.26			
Non-nested	>2 y	Agricultural cover	−0.30	0.43	0.49			
Non-nested	>2 y	Wetland cover	0.15	0.57	0.79			
Non-nested	>2 y	Road density	2.5×10^{-5}	1.7×10^{-5}	0.15			
Non-nested	**>2 y**	**Traffic volume**	**0.08**	**0.20**	**0.0001**	**0.16**	**0.19**	**0.85**
Non-nested	**>2 y**	**Noise Level**	**1.89**	**0.38**	**8×10^{-7}**	**0.16**	**0.21**	**0.56**

to the 75th percentile were associated with absolute increases in discontinuation probabilities of 4 to 14%.

Non-nested data When routes were categorized based on whether or not they were surveyed for >1 y, the discontinuation probability in the South (22%) was substantially higher than in the Midwest (5%) or North (6%). Forest cover was negatively associated with route discontinuation, whereas traffic volume and noise were positively associated with the probability that a route would be discontinued (Fig. S3; Table 1). However, neither wetland area nor developed cover was a significant predictor of route continuation in the 2014 sample. Increasing variable values from their 25th to 75th percentiles resulted in a 2 to 10% change in the probability of route discontinuation.

Categorizing routes based on completion of >2 y of surveys produced similar results (Fig. S4). Discontinuation rates were higher in the South and at sites with low forest cover, high traffic volume, and high noise (Table 1). Increasing variable values from their 25th to their 75th percentiles resulted in a 3 to 8% change in the probability of route discontinuation. However, the maximum values for traffic volume and noise levels were associated with high discontinuation probabilities (up to 85%).

Multivariate analysis of route discontinuation

In the univariate analyses for the nested and non-nested data sets, region, forest cover, traffic volume, and noise level were all consistently related to the probability of survey discontinuation. Thus, we included these 4 explanatory variables in the multivariate logistic regression models presented below, with the proviso that correlations exist among the predictor variables. Traffic volume and noise level were moderately correlated with each other ($r = 0.57$), whereas the correlations between forest cover and other predictor variables were lower ($r = -0.06$ for traffic volume and $r = -0.11$ for noise level).

For the nested data, the multivariate model predicted higher discontinuation rates after 1 year for routes in the South, for routes with low forest cover, and for routes with high traffic volume, and high noise levels (Table S1). The multivariate model predicted discontinuation probabilities as low as 3% for sites in the North (where discontinuation probability was lower) with high forest cover (75th percentile), and low (25th percentile) traffic volume and road noise, and as high as 22% for sites in the South (where discontinuation probability was higher) with low (25th percentile) forest cover and high (75th percentile) traffic and noise levels. McFadden's pseudo-R^2 for this model was 0.27, indicating a good fit to the data. The multivariate model for discontinuing surveys before 3 y was very similar (Table S2). Sites in the North with high (75th percentile) forest cover and low (25th percentile) traffic and noise were predicted to have a discontinuation probability of 16%, whereas sites in the South with the low forest cover (25th percentile) and high (75th percentile) traffic and noise were predicted to have a 48% discontinuation probability. McFadden's pseudo-R^2 for this model was 0.19.

For the non-nested data set, routes in the South had an increased probability of discontinuation after 1 y compared to routes in the Midwest or North. Routes with less forest cover also had lower probabilities of route continuation. Noise level was significantly related to discontinuation probability, but traffic volume was not a significant predictor of discontinuation when noise level was included in the model (Table S3). Predicted discontinuation probability was only 2% for routes in the North with high (75th percentile) forest cover, and low (25th percentile) traffic and noise, but increased to 25% for sites in the South with low (25th percentile) forest cover, and high (75th percentile) traffic and noise. The model for discontinuation before 3 y was very similar, once again with significant effects of region, forest cover, and noise level, but not traffic volume (Table S4). Predicted discontinuation probability from this model was 5% for routes in the North with high (75th percentile) forest cover and low (25th percentile) traffic and noise and 46% for Southern routes with low (25th percentile) forest cover and high (75th percentile) traffic and noise. Traffic volume and noise were moderately correlated within the data set ($r = 0.61$), and both variables were weakly negatively correlated with forest cover ($r = -0.12$ for traffic volume and $r = -0.17$ for noise). McFadden's pseudo-R^2 for models using both threshold criteria with the non-nested data was 0.24.

Effects of detection frequency

We hypothesized that the number of frog and toad detections in the 1st year of surveys would influence whether a volunteer would continue with their surveys. For a model that included only a categorical variable for state and initial detection frequency, fewer detections was associated with an increased probability of route discontinuation ($b = -0.22$, $p = 0.01$). A change in the number of detections in the 25th percentile value (4.17) to the 75th percentile value (9.17) resulted in an increase in the predicted probability of discontinuing a survey from 18% to 29%. For a model that included state as a categorical variable and forest cover and noise level as covariates, detection frequency was again a significant predictor of route discontinuation ($b = -0.23$, $p = 0.01$). In this model, predicted survey discontinuation went from 10% for sites with high (75th percentile) initial detection frequency, and median forest cover and noise to 32% for sites with few initial detections (25th percentile) and median forest cover and noise. Anecdotally, there were 7 routes in our data set that produced ≤1 average detections per survey in the 1st year. Of these 7 routes, 6 of them (84%) were not resurveyed in the following year (compared to 22% overall).

Structural equation model

Results of the SEM (Fig. 2) largely agreed with the multivariate logistic regression for each data set. The model identified direct effects of forest cover and noise level on the probability of survey discontinuation, but not traffic volume. Additionally, the model identified species richness as an additional influence on survey continuation, such that sites with more frog species were more likely to be resurveyed in subsequent years.

Effects on variable distributions

One potential effect of non-random survey discontinuation is loss of highly disturbed or urbanized sites from the

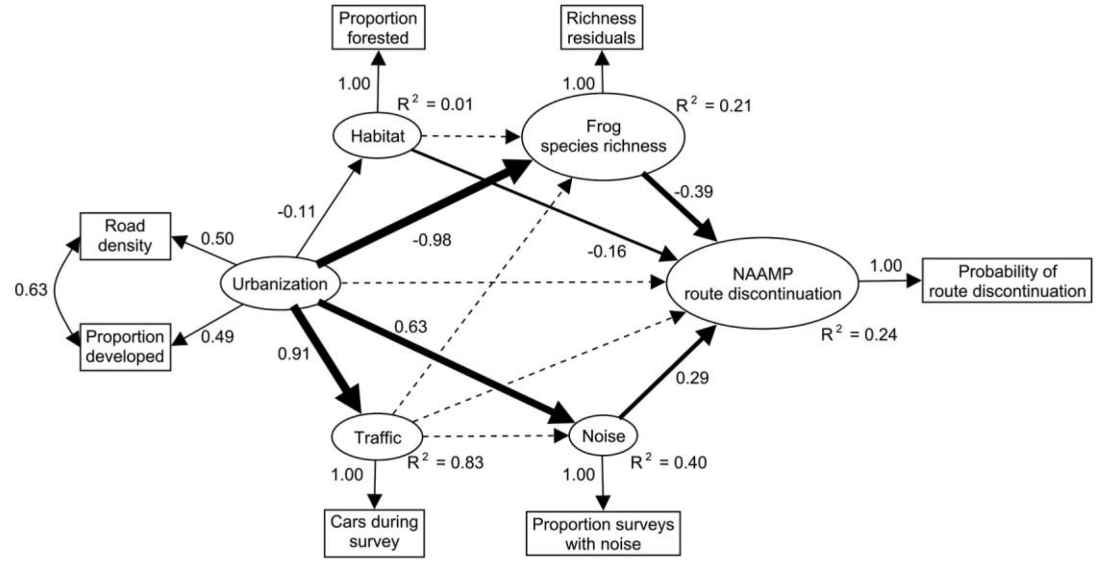

Figure 2. Results from a structural equation model (SEM) for route discontinuation within the North American Amphibian Monitoring Program. Measured variables are contained within boxes, whereas latent variables are shown within ellipses. Arrow thickness is proportional to SEM coefficient. Solid arrows indicate coefficients with 95% percent confidence intervals that differ from zero, whereas dashed arrows indicate coefficients with 95% confidence intervals that include zero. R^2-values show the proportion of variation explained for each variable. Model fit for the SEM was good ($\chi^2 = 6.57$, df = 6, $p = 0.36$; RMSEA = 0.01).

sample. Comparison of variable distributions for forest cover, traffic volume and noise all showed fewer extreme values when sites surveyed <3 y were excluded, though none of these differences were statistically significant. For forest cover, routes with <10% forest cover represented 20.1% of the full sample vs 19.1% of the sample of routes surveyed for ≥3 y (Fisher's exact test, $p = 0.75$). For traffic, routes with mean car count >5 represented 12.0% of the full sample but 8.7% of the sample of routes surveyed ≥3 y ($p = 0.10$). Finally, for noise level, sites with noise frequency of >0.5 represented 8.5% of the full sample, but 6.0% of the routes surveyed for ≥3 y ($p = 0.15$).

Effects of survey discontinuation on observed patterns in NAAMP data

Correlations between site characteristics and species richness were reduced in 5 out of 6 cases in which discontinued sites were excluded (Table 2). However, in all of these cases the 95% confidence intervals for correlation coefficients overlapped between the data sets (Table 2).

DISCUSSION

We found substantial evidence that discontinuation of NAAMP survey routes was non-random, and instead depended on survey site characteristics. In particular, sites with low forest cover, high vehicle traffic, and high noise levels were consistently associated with increased probabilities of survey discontinuation after 1 or 2 y. Further, our SEM analysis suggested that the correlated effects of traffic and noise on route discontinuation were probably associated with noise levels, whereas the effect of low forest cover appeared to be an independent of other variables. In addition, separate analyses

Table 2. Changes in correlations between landscape variables and species richness when discontinued routes are included (full) or excluded (≥3 y) from both the nested and the non-nested data sets. Species richness is estimated as the residual of the regression of raw richness on number of surveys in order to account for variation in survey effort. Results show that correlations between landscape variables and richness generally decreased when discontinued routes were excluded.

Data set	Variable	Correlation with rich ness (95% CI)
Nested, full	Forest cover	0.03 (−0.07 to 0.13)
Nested, ≥3 y	Forest cover	−0.08 (−0.20 to 0.05)
Nested, full	Traffic volume	−0.24 (−0.33 to −0.14)
Nested, ≥3 y	Traffic volume	−0.20 (−0.20 to −0.07)
Nested, full	Noise Level	−0.22 (−0.32 to −0.13)
Nested, ≥3 y	Noise Level	−0.15 (−0.27 to −0.03)
Non-nested, full	Forest cover	0.06 (−0.03 to 0.14)
Non-nested, ≥3 y	Forest cover	0.01 (−0.08 to 0.11)
Non-nested, full	Traffic volume	−0.21 (−0.29 to −0.13)
Non-nested, ≥3 y	Traffic volume	−0.09 (−0.18 to −0.01)
Non-nested, full	Noise Level	−0.24 (−0.31 to −0.16)
Non-nested, ≥3 y	Noise Level	−0.13 (−0.22 to −0.04)

found that route discontinuation was more likely when fewer total amphibians were detected and when observed species richness was lower. The absolute change in discontinuation probability was often low (e.g., 2–10%) over the mid-range values for site variables (e.g., 25^{th} to 75^{th} percentile values). However, at more extreme values for some variables, such as very high traffic or very few amphibians detected, the predicted probability that a survey site would be discontinued could be 50 to 85%.

The non-random drop-outs that we observed had several apparent effects on NAAMP data. First, these drop-outs led to an under-representation of highly urbanized sites in our sample. The effect sizes of these drop-outs were usually small, but NAAMP surveys began with few highly urbanized survey routes, and these drop-outs made it even more difficult to evaluate the effects of urbanization (Cosentino et al. 2014). A second effect of drop-outs was their influence on observed relationships between site characteristics and amphibian species richness in our sample. In most cases, correlations between site characteristics and amphibian richness decreased when we restricted analysis to sites with ≥3 y of data, as we did in our previous analyses of NAAMP data (Cosentino et al. 2014, Marsh et al. 2017). Reduced correlations among these variables are expected because sites with high levels of traffic and noise, and with low levels of amphibian richness, would have fewer surveys and, therefore, be more likely to be excluded. These reductions in correlation would occur regardless of whether the low recorded amphibian richness is real or a result of reduced detection caused by noise. In some cases, correlations did switch from being 'significant' (i.e., confidence limits not over-lapping zero) to being 'non-significant' (i.e., confidence limits overlapping with zero) when sites with few surveys were excluded, although the reductions in correlation coefficients always fell within the 95% confidence limits for these parameters.

We found a clear effect of noise and forest cover, but it seems likely that the influential factors would differ from one project to another. Few other studies have investigated the specific factors that influence drop-outs from citizen science projects (but see Tulloch and Szabo 2012). With roadside amphibian call surveys, volunteers may find heavily-trafficked and noisy survey routes less appealing. However, for other types of surveys (e.g., water quality), easily accessible (i.e., less remote) areas might be preferred for collecting data (Tulloch and Szabo 2012).

Our finding with respect to number of amphibian detections and species richness could indicate a more general problem. That is, for surveys of animals and plants, volunteer retention may consistently be a challenge when the target species are frequently undetected. If so, this would be a potentially important problem because accurately modeling species distributions requires substantial data on absence as well as presence. In addition, if low abundance sites are avoided by volunteers, citizen science studies may miss changes in populations at these sites. For citizen science studies of water quality, it is not obvious (to us, anyway) whether volunteers would tend to prefer sites that appear pristine and, therefore, safe to visit, or sites that appear highly polluted and, therefore, worthy of concern. Further studies of this issue would allow for a better understanding of the biases associated with volunteer water sampling and the actions needed to correct for them. In some cases, data may already exist to carry out these kinds of analyses.

When biases in drop-outs are apparent in citizen science projects, several actions can be taken to mitigate the issue. First, volunteer education can highlight the importance of monitoring undesirable sites. Most volunteers in citizen science projects value their contribution to the larger scientific enterprise, so convincing them of the importance of monitoring highly urbanized sites or sites where a target species is absent could increase volunteer retention. Providing volunteers with maps that identify under-sampled locations may also help fill in the gaps. Second, stratification in the randomization procedure can ensure adequate representation of sites that might otherwise be under-represented (Tulloch and Szabo 2012). If a citizen science project aims to investigate the effects of urbanization or pollution, it may be necessary to specifically assign more volunteers urbanized or polluted sites to account for drop-outs. NAAMP was not designed with the specific goal of understanding the effects of urbanization, but our own studies were limited by the small number of highly urbanized routes in our samples. Under-representation of high-end values for urbanization may be particularly important if species exhibit threshold responses to urbanization, such that responses occur only at high levels of habitat loss (With and Crist 1995, Radford et al. 2005). Third, where citizen science surveys are conducted in cooperation with local or state agencies and/or NGOs, these organizations could potentially step in to replace surveys lost to drop-outs. More generally, coordination between citizen science projects and agencies with full-time employees can help to increase data quality and survey reliability.

The presence of non-random drop-outs in citizen-science data provides an additional argument in support of explicitly modeling the data collection process along with the biological variables of interest (Royle 2004, Royle and Link 2005). When species detection by observers is modeled in conjunction with presence/absence, fewer situations may arise where survey effort is confounded with other variables of interest. Unfortunately, modeling probability of detection along with occupancy is not always possible because data may be sparse or collected using varied approaches. In our own analyses of NAAMP data, large sections of missing data and different survey years in different states made it difficult for us to fit species occupancy models that take detection into account (Cosentino et al. 2014). Furthermore, modeling the detection process as part of species richness estimation is frequently a challenging problem (Mao and Colwell 2005, Dorazio et al. 2006). For citizen science projects that rely on species richness (e.g., stream invertebrates) to assess habitat quality, many analyses will continue to make use of summary measures that can mask biases in data collection by volunteers.

The data we compiled to analyze NAAMP surveys were not originally designed to examine the issue of volunteer retention. As a result, our analyses for this paper are largely phenomenological, and rely on general patterns in the data (e.g., associations between land use variables and survey numbers) to make inferences about volunteer behavior. On some NAAMP routes that appeared to be continuously monitored over time, observers might actually have discontinued their participation and been replaced with new volunteers before the next set of surveys. In other cases, volunteers might have been reassigned to new NAAMP routes for reasons unrelated to the routes themselves. Thus, our analyses can provide only indirect evidence of non-random drop-outs from NAAMP. Future studies of volunteer behavior within citizen science projects would benefit from explicitly analyzing how volunteers respond to characteristics of their survey sites and the data they collect on a year-to-year basis. In addition, surveys of current and former volunteers could be used to enquire specifically about the factors that promote continued participation in the project.

There is a growing need to understand how to analyze the data citizen science projects produce as they become an increasingly common approach to collecting environmental data (Dickinson et al. 2010, Tulloch et al. 2013). Much as political polling relies on detailed studies of response rates among different demographic groups, accurate estimation from citizen science data requires an understanding of the factors that influence data collection by volunteers. Fortunately, common biases in citizen science data can potentially be corrected for once they are understood (Bird et al. 2014). We hope our study will motivate other researchers to examine volunteer behavior in the context of major citizen science initiatives, thereby improving the capacity of citizen science to address critical environmental issues.

ACKNOWLEDGEMENTS

Author contributions: DMM wrote most of the manuscript and contributed to the data analysis. BJC wrote part of the manuscript, produced figures, and contributed to the data analysis.

We thank Jim Regetz, Stephanie Hampton, and the staff at the National Center for Ecological Analysis and Synthesis for facilitating the 'Roads, Toads, and Nodes' project. Data were contributed from classes taught by J. J. Apodaca (Warren Wilson College), Karen Beard (Utah State University), Christine Bozarth (Northern Virginia Community College), Kristen Genet (Anoka-Ramsey Community College), Nancy Karraker (University of Rhode Island), Eran Kilpatrick (University of South Carolina, Salkehatchie), Tom Langen (Clarkson University), and James Vonesh (Virginia Commonwealth University). Kara Jones (University of Kentucky) and Emily Cole (Washington and Lee University) compiled NAAMP data into a useful format for analysis. Linda Weir (USGS) helped us understand NAAMP methods and data collection procedures. This research was funded by the U.S. National Science Foundation through TUES (Transforming Undergraduate Education in Science) program #1140475.

LITERATURE CITED

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

Chung, U., L. Mack, J. I., Yun, and S. H. Kim. 2011. Predicting the timing of cherry blossoms in Washington, DC and mid-Atlantic states in response to climate change. PLoS ONE 6:e27439.

Cosentino, B. J., D. M. Marsh, K. S. Jones, J. J. Apodaca, C. Bates, J. Beach, K. H. Beard, K. Becklin, J. M. Bell, C. Crockett, G. Fawson, J. Fjelsted, E. A. Forys, K. S. Genet, M. Grover, J. Holmes, K. Indeck, N. E. Karraker, E. Kilpatrick, T. A. Langen, S. G. Mugel, A. Molina, J. R. Vonesh, R. Weaver, and A. Willey. 2014. Citizen science reveals widespread negative effects of roads on amphibian distributions. Biological Conservation 180: 31–38.

Crall, A. W., G. J. Newman, T. J. Stohlgren, K. A. Holfelder, J. Graham, and D. M. Waller. 2011. Assessing citizen science data quality: an invasive species case study. Conservation Letters 4:433–442.

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

Dickinson, J. L., B. Zuckerberg, and D. N. Bonter. 2010. Citizen science as an ecological research tool: challenges and benefits. Annual Review of Ecology, Evolution, and Systematics 41: 149–172.

Dorazio, R. M., J. A. Royle, B. Söderström, and A. Glimskär. 2006. Estimating species richness and accumulation by modeling species occurrence and detectability. Ecology 87:842–854.

Engel, S. R., and J. R. Voshell. 2002. Volunteer biological monitoring: can it accurately assess the ecological condition of streams? American Entomologist 48:164–177.

Fry, J., G. Xian, S. Jin, J. Dewitz, C. Homer, L. Yang, C. Barnes, N. Herold, and J. Wickham. 2011. Completion of the 2006 National Land Cover Database for the Conterminous United States. Photogrammetric Engineering and Remote Sensing 77:858–864.

Genet, K. S., and L. G. Sargent. 2003. Evaluation of methods and data quality from a volunteer-based amphibian call survey. Wildlife Society Bulletin 31:703–714.

Grace, J. B. 2006. Structural Equation Modeling and Natural Systems. Cambridge University Press, Cambridge, UK.

Grace, J. B., T. M. Anderson, H. Olff, and S. M. Scheiner. 2010. On the specification of structural equation models for ecological systems. Ecological Monographs 80:67–87.

Hochachka, W. M., D. Fink, R. A. Hutchinson, D. Sheldon, W. K. Wong, and S. Kelling. 2012. Data-intensive science applied to broad-scale citizen science. Trends in Ecology and Evolution 27:130–137.

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

Kline, R. B. 2015. Principles and practice of structural equation modeling. 4th Edition. Guilford Press, New York.

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

Kremen, C., K. S. Ullman, and R. W. Thorp. 2011. Evaluating the quality of citizen-scientist data on pollinator communities. Conservation Biology 25: 607–617.

Link, W. A., and Sauer, J. R. 1998. Estimating population change from count data: application to the North American Breeding Bird Survey. Ecological Applications 8:258–268.

Mao, C. X., and R. K. Colwell. 2005. Estimation of species richness: mixture models, the role of rare species, and inferential challenges. Ecology 86: 1143–1153.

Marsh, D. M., B. J. Cosentino, K. S. Jones, J. J. Apodaca, K. H. Beard, J. M. Bell, C. Bozarth, D. Carper, J. F. Charbonnier, A. Dantas, E. A. Forys, M. Foster, J. General, K. S. Genet, M. Hanneken, K. R. Hess, S. Hill, F. Iqbal, N. E. Karraker, E. S. Kilpatrick, T. A. Langen, J. Langford, K. Lauer, A. J. McCarthy, J. Neale, S. Patel, A. Patton, C. Southwick, N. Stearrett, N. Steijn, M. Tasleem, J. M. Taylor, and J. R. Vonesh. 2017. Diversity and Distributions 23:158–170.

McClintock, B. T., L. L. Bailey, K. H. Pollock, and T. R. Simons. 2010. Experimental investigation of observation error in anuran call surveys. Journal of Wildlife Management 74:1882–1893.

McFadden, D. 1974. Conditional logit analysis of qualitative choice behavior. Pages 105–142 in P. Zarembka (editor). Frontiers in Econometrics. Academic Press, Cambridge, Massachusetts.

Miller, D. A. W., L. A. Weir, B. T. McClintock, E. H. Campbell Grant, L. L. Bailey, and T. R. Simons. 2012. Experimental investigation of false positive errors in auditory species occurrence surveys. Ecological Applications 22:1675–1688.

R Development Core Team. 2016. R version 3.3.1: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Radford, J. Q., A. F. Bennett, and G. J. Cheers. 2005. Landscape-level thresholds of habitat cover for woodland-dependent birds. Biological Conservation 124:317–337.

Rosseel, Y. 2012. lavaan: an R package for structural equation modeling. Journal of Statistical Software 48: 1–36.

Royle, J. A. 2004. Modeling abundance index data from anuran calling surveys. Conservation Biology 18:1378–1385.

Royle, J. A., and W. A. Link. 2005. A general class of multinomial mixture models for anuran calling survey data. Ecology 86:2505–2512.

Tulloch, A. I., H. P. Possingham, L. N. Joseph, J. Szabo, and T. G. Martin. 2013. Realizing the full potential of citizen science monitoring programs. Biological Conservation 165:128–138.

Tulloch, A. I. and J. K. Szabo. 2012. A behavioural ecology approach to understand volunteer surveying for citizen science datasets. Emu: Austral Ornithology 112:313–325.

US Census Bureau. 2013. 2013 TIGER/Line Shapefiles. (Available from: http://www2.census.gov/geo/pdfs/maps-data/data/tiger/tgrshp2013/TGRSHP2013_TechDoc.pdf)

US Department of the Interior, Fish and Wildlife Service. 2013. National Wetlands Inventory website. U.S. Fish and Wildlife Service, Washington, DC. (Available from: http://www.fws.gov/wetlands/)

Weir, L. A., and M. J. Mossman. 2005. North American Amphibian Monitoring Program. Pages 307–313 *in* M. J. Lannoo (editor). Amphibian Declines: the Conservation Status of United States Species. University of California Press, Berkeley, California.

With, K. A., and T. O. Crist. 1995. Critical thresholds in species' responses to landscape structure. Ecology 76:2446–2459.