Journal of Applied Ecology

BRITISH ECOLOGICAL SOCIETY

Journal of Applied Ecology 2017, **54,** 628–637 doi: 10.1111/1365-2664.12682

Evaluating citizen vs. professional data for modelling distributions of a rare squirrel

Courtney A. Tye[1]†, Robert A. McCleery[1]*, Robert J. Fletcher Jr[1], Daniel U. Greene[1] and Ryan S. Butryn[2]

[1]Department of Wildlife Ecology and Conservation, University of Florida, Gainesville, FL 32611, USA; and [2]Fish and Wildlife Research Institute, Florida Fish and Wildlife Conservation Commission, Gainesville, FL 32601, USA

Summary

1. To realize the potential of citizens to contribute to conservation efforts through the acquisition of data for broad-scale species distribution models, scientists need to understand and minimize the influences of commonly observed sample selection bias on model performance. Yet evaluating these data with independent, planned surveys is rare, even though such evaluation is necessary for understanding and applying data to conservation decisions.

2. We used the state-listed fox squirrel *Sciurus niger* in Florida, USA, to interpret the performance of models created with opportunistic observations from citizens and professionals by validating models with independent, planned surveys.

3. Data from both citizens and professionals showed sample selection bias with more observations within 50 m of a road. While these groups showed similar sample selection bias in reference to roads, there were clear differences in the spatial coverage of the groups, with citizens observing fox squirrels more frequently in developed areas.

4. Based on predictions at planned field surveys sites, models developed from citizens generally performed similarly to those developed with data collected by professionals. Accounting for potential sample selection bias in models, either through the use of covariates or via aggregating data into home range size grids, provided only slight increases in model performance.

5. *Synthesis and applications*. Despite sample selection biases, over a broad spatial scale opportunistic citizen data provided reliable predictions and estimates of habitat relationships needed to advance conservation efforts. Our results suggest that the use of professionals may not be needed in volunteer programmes used to determine the distribution of species of conservation interest across broad spatial scales.

Key-words: citizen science, data aggregation, fox squirrel, habitat relationships, opportunistic data, predictive performance, road bias, sample selection bias, species distribution models, validation

Introduction

Development of effective conservation strategies for the planet's growing biodiversity crisis (Ceballos *et al.* 2015) necessitates an understanding of species' broad-scale distributions and habitat associations (Kremen *et al.* 2008; Hochachka *et al.* 2012). Unfortunately, data from planned broad-scale surveys that can be used to generate this information are often lacking (Isaac *et al.* 2014). To address this shortcoming, ecologists have increasingly turned to the public for the collection of species occurrence and habitat data (Silvertown 2009).

The increasing involvement of citizens from the non-scientific community (hereafter, citizens) in research (citizen science) has transformed how large-scale environmental monitoring and research programmes are conducted (Conrad & Hilchey 2011). Citizens now gather data for an array of projects that would otherwise be difficult, if not impossible, to obtain due to time and resource limitations (Dickinson, Zuckerberg & Bonter 2010). In addition

*Correspondence author. E-mail: ramccleery@ufl.edu

†Courtney Tye is deceased. This manuscript was derived from Courtney Tye's research for her M.S. degree. She drafted the first version of the manuscript. The final version was approved by her husband Barry Tye.

to providing data, citizen science may increase positive public engagement with scientific research and natural resource issues (Devictor, Whittaker & Beltrame 2010; Dickinson, Zuckerberg & Bonter 2010).

It is now a common practice to use data collected by citizens to model the distribution, abundance and species richness of plants and animals (Pearce & Boyce 2006; Silvertown 2009; Dickinson *et al.* 2012). This trend has likely arisen from the proven ability of citizens to collect large amounts of observations across broad spatial scales (Devictor, Whittaker & Beltrame 2010; Hochachka *et al.* 2012) and the relative ease with which species distribution and habitat models can be generated (Phillips, Anderson & Schapire 2006).

While opportunistic, citizen-collected, presence-only data have proven valuable (Elith *et al.* 2006), there are a number of potential issues that have been identified when using this type of data to model species distributions. Citizen-generated data often can lack records of species absences (Isaac & Pocock 2015). Although absences can sometimes be inferred from species lists (i.e. birds and butterflies), detection of species with few ecologically and physiologically similar conspecifics is rarely recorded in lists. In situations where absences cannot be inferred, only measures of relative suitability probability can typically be modelled (Hastie & Fithian 2013). Opportunistic, citizen-collected, presence-only data may also suffer from quality issues and sample selection biases, where some sites are more likely to be surveyed than others (e.g. road bias; Phillips *et al.* 2009). Sample selection bias can arise due to spatial or geographic constraints on citizens' sampling, where citizens disproportionally collect from areas that are convenient and highly accessible (Dennis, Sparks & Hardy 1999; Phillips & Dudik 2008; van Strien, van Swaay & Termaat 2013). Citizens appear more likely to sample areas closer to where they live and to under sample remote regions (Dennis, Sparks & Hardy 1999; Isaac & Pocock 2015). Opportunistic data are also likely to have road bias with more detections close to roads and trails (Crall *et al.* 2010). Large numbers of background samples or 'pseudo-absences' can reduce the influence of sample selection bias in presence-only data (Barbet-Massin *et al.* 2012), but these biases have a much stronger effect on presence-only models than for models that use confirmed absences (Phillips *et al.* 2009). From a quality perspective, there is concern that citizens – unlike professionals – do not have the ability or expertise to find or identify (detect) species of interest (Fitzpatrick *et al.* 2009; Silvertown 2009). The quality of citizen-generated data may also suffer if citizens are asked to provide data that do not align with their interests and skills (Lukyanenko, Parsons & Wiersma 2014).

One approach to improve data quality and reduce sample selection bias may be using volunteers that are trained professionals to collect data. Professionals should be more invested in projects' outcomes and due to their training, familiarity with data, and use of protected and restricted lands, they may be more likely to survey remote areas and away from roads. They may also be better able to identify and detect rare animals. Accordingly, limiting data collection to professional volunteers has the potential to reduce sampling bias near roads and areas of high population density and may reduce issues of species detection and misidentification. Alternatively, from a modelling perspective, incorporating road biases as a covariate into models (Warton, Renner & Ramp 2013) and aggregating records on coarser scales may reduce some sample biases (Elith *et al.* 2011; Fourcade *et al.* 2014). However, aggregating records is to risk throwing out a portion of the data collected or not using all the information from place-based records (Isaac & Pocock 2015). The loss of data is troubling because the large samples produced by citizens may increase the predictive performance of some models (e.g. Hernandez *et al.* 2006), allowing researchers to overcome sample biases.

There are outstanding questions about the use of opportunistic species observations in species distribution models, and there is a clear need to find better ways to turn the large quantities of citizen-generated data into useful information (Hochachka *et al.* 2012; Isaac *et al.* 2014; Isaac & Pocock 2015). To realize the potential of citizens to generate the data needed to produce broad-scale species distribution maps (Phillips & Dudik 2008; Devictor, Whittaker & Beltrame 2010; Hochachka *et al.* 2012), there is a need to understand and minimize the potential for sample selection bias and data quality issues on model performance. This is only possible through a rigorous model assessment and validation. Nonetheless, distribution models are rarely evaluated with independent data and the use of planned field surveys (i.e. prospective sampling; Fielding & Bell 1997) to determine model validity is even less common (Elith *et al.* 2006). More specifically, the assessment of citizen science-generated models with independent data is almost non-existent (but see Kadoya *et al.* 2009) and particularly troublesome given their considerable potential for sample selection bias.

To quantify the influences of potential sample selection bias and evaluate the performance of models created with opportunistic observations from citizens, we used the fox squirrel *Sciurus niger* in the state of Florida as an illustrative case study. Information on the spatial ecology (distribution and habitat utilization) of Florida's fox squirrels is critical to their management and conservation (Florida Fish and Wildlife Conservation Commission [FWC] 2013). However, statewide locational information on the fox squirrel has been unavailable because systematically sampling throughout Florida (170 304 km^2) has not been a viable option.

Our goal was to understand and account for sample selection biases in opportunistic citizen collated data to generate a reliable distribution map of the fox squirrel in Florida. Our specific objectives were to (i) compare predictive models produced from opportunistic observations from professionals and citizens; (ii) compare the potential sample selection bias of citizen vs. professionally collected data; (iii) understand the influence of roads on data acquisition and model performance; and (iv) determine whether data from professionals and citizens altered the interpretation of environmental relationships of fox squirrel across the state of Florida.

Materials and methods

STUDY SPECIES

The fox squirrel is a mid-large sized (800–1200 g) tree squirrel that occurs naturally throughout most of the south-eastern and mid-western USA (Steele & Koprowski 2001). In the south-eastern USA, fox squirrels are thought to be experiencing declines in abundance and distribution due to habitat loss from land conversion and fire suppression (Weigl *et al.* 1989; Loeb & Moncrief 1993). Florida has four fox squirrel subspecies (Sherman's [*S. n. shermani*], Big Cypress [*S. n. avicennia*], Bachman's [*S. n. bachmani*] and the south-eastern [*S. n. niger*]). Both Sherman's and Big Cypress fox squirrels, whose combined distribution covers most of peninsular Florida, are rare, difficult to detect (Eisenberg *et al.* 2011) and listed as Species of Conservation Concern by the state of Florida (Florida Fish and Wildlife Conservation Commission 2012). All four of these subspecies appear to have sparse and patchy distributions across a broad array of habitats (Loeb & Moncrief 1993). However, once detected they are easily recognizable and distinguishable from other species of squirrel due to their unique coloration (Tye *et al.* 2015).

WEBSITE AND DATA COLLECTION

We developed a web-based tool (webpage; https://public.myfwc.com/hsc/foxsquirrel/GetLatLong.aspx) to allow natural resource professionals and the general public (citizens) to submit sightings of fox squirrels in Florida. We relied solely on volunteers to collect data for this project because their intrinsic motivations have been liked to increased effort and participation and effort in collective scientific research programmes (Nov, Arazy & Anderson 2011). We promoted the site to professionals by sending agency wide (e.g. Florida Fish and Wildlife Conservation Commission, Florida State Parks) and targeted emails (e.g. Big Cypress National Preserve, St. Marks National Wildlife Refuge), posting flyers in offices of natural resources agencies and promoting the website at local natural resources conferences and workshops. We promoted the site to the public at extension events (e.g. Florida Black Bear & Wildlife Conservation Festival), meetings of conservation organization (e.g. Audubon clubs) and through local newspapers (e.g. Tampa, Sarasota, Gainesville and Orlando) and newsletters (e.g. Florida Cattlemen Association, Florida Master Naturalist). Finally, we used our social and professional networks to promote the website via Facebook.

The web-based tool recorded georeferenced locations (latitude and longitude in decimal degrees) using a Google map application to record sightings in the data base. Along with the georeferenced fox squirrel locations, we asked each participant to enter their name, date of the sighting, organization (if applicable) and email address. Additionally, we asked participants whether they were a member of the general public or a natural resource professional (biologist, natural resource extension, forester, land manager, etc.). A comments box allowed participants to enter further sighting information such as surrounding land use, behaviour and individual description. We also encouraged participants to provide pictures of fox squirrels to confirm the validity of their sightings. We activated the website from 20 August 2011 to 1 April 2012 to coincide with peak fox squirrel activities during pine (*Pinus* spp.) and oak (*Quercus* spp.) masting events (Weigl *et al.* 1989). To eliminate erroneous squirrel locations from the survey, we reviewed each location and removed locations that appeared to be from user error (i.e. locations in the middle of a water body). Questionable points were verified or removed after we gathered more information from the participant that submitted the data.

ENVIRONMENTAL VARIABLES

We fit distribution models to a suite of environmental variables that we expected would influence the fox squirrels' distribution. We assessed the strength of the relationship between our variables using the correlation coefficient (r), and defined variables >0·7 as being highly correlated. For most variables, we transformed the grain size to 30 × 30 m pixels and generated a sum or average for each pixel based on a 25-ha neighbourhood, the average home range of fox squirrels in Florida (Kantola & Humphrey 1990). We removed duplicate squirrel locations within the same 30 × 30 m pixel area. We generated pixel averages using focal statistics in Spatial Analyst ArcMap 10 (Environmental Systems Research Institute, Redlands, CA, USA). To determine the average (*treeA*) and standard deviation (*treeSTD*) of tree canopy cover in area surrounding an observation, we used the 2011 National Land Cover Database (Homer *et al.* 2015). We estimated the majority land cover type (*LC_maj*) surrounding each squirrel location by classifying the Florida Natural Areas Inventory (FNAI 2012) land covers into 19 relevant categories (see Appendix S1 in Supporting Information). To generate estimates of edge, we considered the intersection of land cover types as edges and summed the amount of edge (*sumedge*) available to a fox squirrel (25 ha neighbourhood). We estimated the amount of forest available (*forest_patch*) to a fox squirrel in the surrounding environment by summing all the area of forest land covers within 25 ha of the squirrel location. However, *forest_patch* was highly correlated with *treeA* ($r = 0{\cdot}80$), so we did not include *forest_patch* in the analysis. We quantified the elevation at each location using the 2005 US. Geological Survey National Elevation Dataset (NED) for the state of Florida. Finally, to account for potential road sample bias, we incorporated a nuisance variable that estimated the distance from each squirrel location to the nearest roadway (*roadbias*) using the US. Census Bureau's Florida 2013 Topologically Integrated Geographic Encoding and Referencing (TIGER) data layer.

DISTRIBUTION MODELLING

To address the objectives of the study, we developed models from four data sets of observations: (i) citizen and professional observations combined; (ii) citizen observations only; (iii) professionals observations only; and (iv) a subset of the citizen observations equivalent in size to the professional data set. For each of these data sets, we ran models with four different configurations to account for potential sample selection biases in the data. We created models that included a covariate to account for road bias (Warton, Renner & Ramp 2013), a subsampling grid that randomly selected one detection within each 25-ha grid cell (Fourcade *et al.* 2014), both a covariate for roads and a subsampling grid and no added spatial adjustments. Using these 16 different data/configuration combinations, we modelled fox squirrel distribution as a function of environmental variables across Florida utilizing the maximum entropy program Maxent (version 3.3.3k, http://www.cs.princeton.edu/˜schapire/maxent/), called through the dismo package in R. We fit models using only linear and quadratic relationships in an effort to not overfit our data (Merow, Smith & Silander 2013) and to add in biological interpretation. Additionally, we altered the tuning parameter to adjust for model complexity (β = 1–20). We set Maxent to randomly generate 10 000 background points to allow for the distribution analysis (Barbet-Massin *et al.* 2012). We used these same background points for each model assessment to allow formal comparison between models (Merow, Smith & Silander 2013). To compare variation in predicted environmental relationships, we used a nonparametric bootstrap (n = 500 samples) to generate partial predictions for each environmental variable and its associated uncertainty. To determine the relative importance of variables contributing to the Maxent models, we used a jackknife procedure based on five sets of simulations (Phillips *et al.* 2009). To ensure that our conservative modelling approach did not limit the overall predictive ability of our models, we also ran all of our models using the hinge and threshold features in Maxent that allows for more complex relationships to be modelled, but conclusions did not change.

INDEPENDENT FIELD VALIDATION

We conducted field surveys for fox squirrels using passive camera traps to collect independent data for validation of Maxent models. Typically, species distribution models are validated using variants of cross-validation; however, such assessments are not based on truly independent data (Fielding & Bell 1997; McCarthy *et al.* 2012) and can provide misleading inferences (and unwarranted optimism; Araújo *et al.* 2005) on the predictive ability of models. We conducted our surveys within the core range of Sherman's fox squirrels in central and northern Florida on public and private lands (Fig. 1). The vegetative communities at our sites were highly variable, and included open grasslands, pine-dominated forests, hardwood-dominated forests, mixed pine–hardwoods, bottomland forest and clear cuts. The canopy trees varied, but the dominant pine trees included longleaf *P. palustris*, slash *P. elliottii* and loblolly *P. taeda* pines, and the dominant oaks were turkey (*Quercus laevis*, live *Q. virginiana*, laurel *Q. laurifolia* and water *Q. nigra* oaks). Vegetation management practices included cattle grazing, timber production, mowing, burning and no active management.

For prospective sampling, we used a stratified random design to selected 14 points in each of three strata (high > 0·60, medium 0·30–0·59 and low probability of occurrence ≤0·29) based on Maxent models (described above). Additionally, we selected forty 7·65 km^2 landscapes across the study region. We selected landscapes using a stratified random design to capture major vegetative communities in the region. We placed 10 grids in upland pine habitats, 10 in mesic pine and hardwood communities and placed the remaining 20 without regard to a vegetative community. Within each landscape, we randomly placed five 5·3-ha survey grids. Each grid consisted of 9 sampling points in a 3 × 3 grid arrangement with 115 m spacing. For this analysis, to reduce spatial autocorrelation we randomly selected a subset of 252 points (84 in each strata) giving a total of 294 points to be used for validation. We surveyed each sampling point for at least eight nights from 1 January to 1 June in 2013 and 2014, with a passive digital camera (Bushnell Trophy Cam model 119436c, Bushnell Outdoor Products, Overland Park, KS). We placed baits of pecans *Carya illinoinensis* and cracked corn *Zea mays* 2·5 m from the camera.

MODEL PERFORMANCE

We first evaluated potential road bias and the spatial occurrence location from both citizen and professionals. We compared environmental covariates at locations for data collected by citizens and professionals to determine whether these data sources were capturing different environmental gradients using MANOVA, and determined key environmental covariates that may explain differences with a subsequent linear discriminant analysis (McCarthy *et al.* 2012). We also determined whether data collected by citizens and professionals as a function of distance from roads differed from availability based on background point locations.

We assessed the ability of all 16 models to predict occurrence at our independent field validation locations using two threshold independent measures, Area under the curve (AUC) and the correlation coefficient (r). The AUC of the receiver-operating characteristic (ROC) plot is a measure of overall predictive accuracy ranging from 0 to 1·0, where 1 is accurate and a 0·5 represents random chance (Fielding & Bell 1997). We also used two threshold-dependent measures – the true skill statistic (TSS) and the kappa statistic (Fielding & Bell 1997; Liu, White & Newell 2011). For TSS and kappa, we set a threshold cut-off based on maximizing the sum of the specificity and sensitivity (Liu, White & Newell 2013).

Results

Our web survey totalled 4222 vetted locations of fox squirrels in 66 of 67 counties in Florida (no reports were received from Broward County) from 2673 different people (Fig. 1). Of the locations recorded, 73% were from citizens and the remaining 27% were submitted by natural resource professionals. We removed 67 points from further analysis because they occurred on the edge of two layers.

There were clear differences in the spatial coverage of fox squirrel observations from professionals and citizens (P < 0·0001; Fig. 2a). Compared to professionals, citizens were more likely to sample fox squirrels in urban areas and less likely to sample them in remote prairies and forests (Figs 1 and 2a). Similarly, professionals and citizens differed spatially on their observation with respect to roads (P < 0·0001; Fig. 2b). Both groups concentrate their sampling <50 m from a road with under sampled areas >200 m from roads. Unlike professional observations, citizen-collected data also appeared to spike again 100 m from a road (Fig. 2b).

Despite the differences in the coverage of data sets, there were only negligible differences (AUC < 0·015,

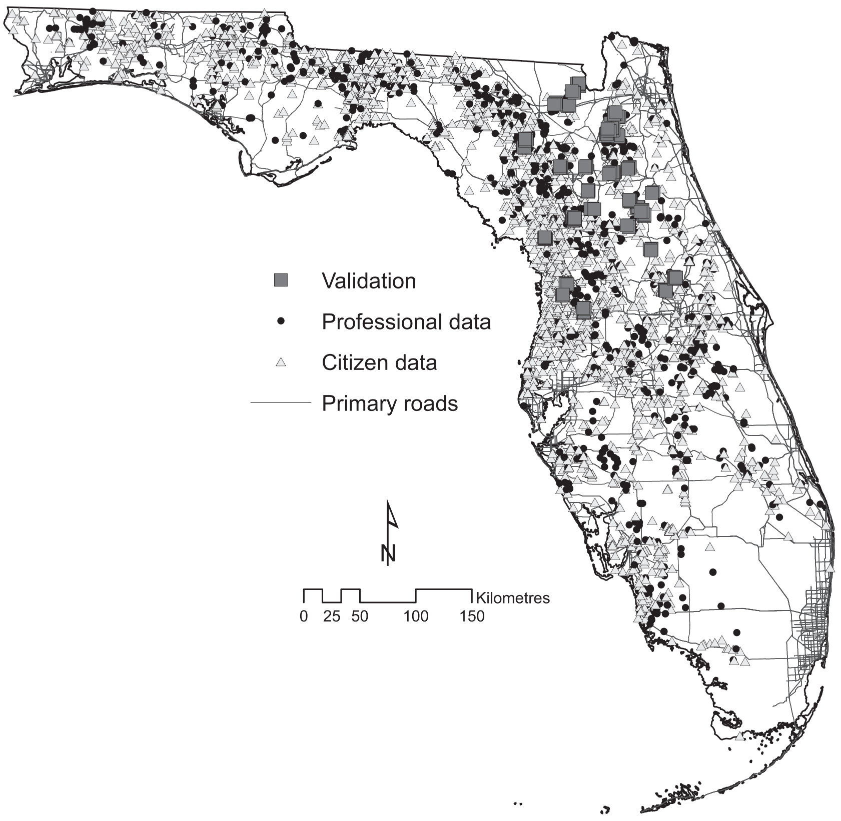

Fig. 1. Study area for distribution modelling of fox squirrels, including presence locations from professional data, citizen data and validation locations.

$r < 0{\cdot}013$, $\mathrm{TSS} < 0{\cdot}08$, $\kappa < 0{\cdot}08$) in model performance among the four different sources of data (citizen and professional, citizen only, professionals only, and a subset of citizens) with the three configurations aimed at reducing sample bias (road bias, grid, grid and road bias, and no adjustments, Table 1.). Using the AUC metric for validation, professionally collected data consistently, but only negligibly, outperformed citizen-collected data. Alternatively, on average the data collected by citizens showed only negligibly better performance than professionally collected data, based on *r*, TSS and Kappa statistics (Table 1). Additionally, we found using the hinge and threshold features in Maxent did not improve our model's predictive ability.

The sample size variation within this study had little influence on overall model performance. The differences between citizen-collected data ($n = 3101$) and a subsample of this data equivalent to the number of professional observations ($n = 1121$) were negligible (Table 1). Additionally, models using the smaller samples of observations from professionals were comparable to models using the larger samples from citizen-generated observations ($n = 3101$). On average, both data aggregation and adding a covariate to correct for road bias increased three of the four validation metrics, but only marginally (Table 1).

Overall, the different data sources generated similar environmental relationships (Fig. 4) and the importance of environmental predictors was similar between models. The distribution of fox squirrels across Florida from similar citizen and professional models (subsampled, not adjusted for road bias) had similar predictive accuracy and fox squirrel distribution responded similarly to environmental variables (Figs 4 and 5). Based on jackknifing, the rank order of variable importance for the best professional model was as follows: land cover > SD of tree cover > elevation > tree cover > edge. For citizen data, the rank order of variable importance was as follows: land cover > tree cover > SD of tree cover > elevation > edge. Models generated using both professional and citizen data sources showed fox squirrels responded positively to pinelands and negatively to coastal uplands, high-intensity urban areas, extractive uses, prairies and water, and hardwood wetlands and mangroves (Fig. 3). Models generated with citizen data suggested fox squirrels responded positively to low-intensity urban areas and parks, cemeteries and golf courses, while models with data from professionals did not (Figs 3 and 5). Fox squirrel occurrence also increased with variability in tree cover and decreased with overall increasing canopy cover from trees (Fig. 4). Fox squirrel occurrence decreased at the lowest elevations (<20 m) and highest elevations (>60 m) and responded negatively in increasing amounts of edge.

Discussion

Citizen science is increasingly used in ecology and conservation, yet there are ongoing concerns regarding the value of such data (Kery, Gardner & Monnerat 2010;

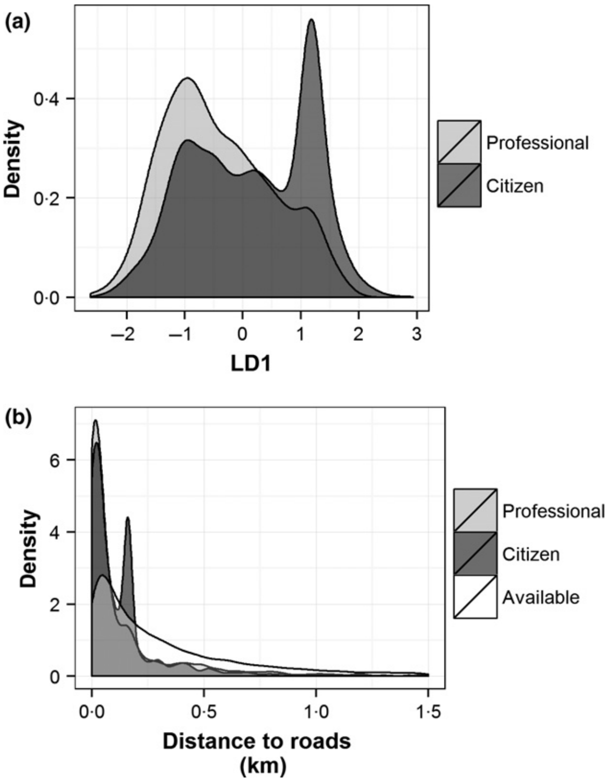

Fig. 2. (a) Differences in environmental space (not including roads) of professional and citizen data based on predicted scores of linear discriminant analysis and (b) road bias. For linear discriminant analysis, differences were driven by land cover types, not tree cover, elevation or edge, where negative loadings on linear discriminant function 1 (LD1) were associated with dry prairie and barren land covers, while positive loadings were associated with coastal uplands, mangroves and exotics. In (b), available refers to background sample used for model building.

Isaac *et al.* 2014). Our results provide new insight to this issue by illustrating that while citizen science data show sample selection bias, such bias does not result in lower predictive ability of models relative to data collected by professionals.

Our web-based survey yielded a large amount of data on fox squirrel occurrence from throughout Florida (Fig. 1) in a relatively short time (194 days). Citizens reported the majority of these data and submitted the only observations for three counties. Generating this amount of data would have taken an extraordinary effort in the field and the web-based survey was a vast improvement over the previous approach of using mail surveys to document fox squirrel occurrences in Florida (Williams & Humphrey 1979; Eisenberg *et al.* 2011).

We tried to enhance the quality of our data by calling on natural resource professionals to volunteer, but data from professionals and citizens alike had considerable sample selection biases. Both groups had disproportionally more observations of squirrels close to roads, and there is no evidence that fox squirrels select for these areas (Steele & Koprowski 2001). Road biases have been shown to be a consistent problem with opportunistically collected data (Kadmon, Farber & Danin 2004; Grand *et al.* 2007; Albert *et al.* 2010). The proportion of fox squirrel observations closer to roads was more pronounced for citizens, where there were two peaks of citizen observations adjacent to (<50 m) and just removed from roads (≈100 m; Fig. 2b). This second peak may be from a citizen's observations at places they frequent. Our data clearly showed a difference in the land cover extent of citizen data (Fig. 2a) which produced models predicting greater occurrence of fox squirrels in low-intensity urban areas, golf courses and parks more than models derived from professional data (Figs 3 and 5).

Surprisingly, when comparing model predictions to independent planned surveys, efforts to reduce bias produced only slight improvements in overall model performance (Table 1). Despite the considerable road biases in observations, explicitly modelling this variation with a covariate did not improve overall performance as in previous studies (Warton, Renner & Ramp 2013). A considerable portion of Florida's available land occurs <250 m of a road, so all data sets may have captured at least some observations across environmental gradients (see also McCarthy *et al.* 2012), even though data sets varied in capturing the frequency of these gradients (Fig. 2a). Also contradicting previous work, we found the larger samples sizes did not measurably improve model performance (Hernandez *et al.* 2006). This result may be because our observations were concentrated in certain areas of the state, rendering the increased sample from citizens of little value. In fact, aggregating sampling into 25-ha grids reduced the number of citizen-generated samples by approximately 30%.

One potential benefit of using professionals along with citizens to collect data was that they sampled in slightly different areas and more remote areas (Figs 1 and 2a). Nonetheless, models created using professionally collected data did not perform better than models using citizen data. Citizens did not appear to have identification issues that might have influenced data quality. We received over 400 pictures of fox squirrels from citizens with no misidentification. While a portion of our validation points were specifically stratified according to Maxent predictions, some of the validation points were stratified across forest types thought to be habitat for fox squirrels. Consequently, the different land cover extent of professional observations more closely matched land cover at our independent validation points. Our validation points generally occurred in areas of reduced human activity. It is possible that if our validation points were collected in areas with more urbanization, the citizen data could have outperformed the professional data (Fig. 1). Alternatively, citizen observations in urban areas may have suffered from sample selection bias because more people frequent these areas. Further assessments using independent sampling across the entire land cover and geographic gradient considered by citizens would be useful.

Table 1. Evaluation of fox squirrel Maxent models based on citizen science data, professional data or both data sets combined (all), when predicting to independent, prospective sampling locations. Models considered potential road bias using a distance to road covariate, as well as general sample bias by subsampling data based on a 25-ha grid. For each bias combination, models with the highest area under the curve (AUC) statistic, correlation coefficient (*r*), true skill Statistic (TSS), and kappa statistic are in bold. Across all model combinations, models with highest AUC, r, TSS and kappa are in italics and bold

	Sample selection bias considered?						
Data used	Road	Sample (subsampling)	β*	AUC	*r*	TSS	κ
All	No	No	12	0·735	0·299	0·445	0·216
Professional	No	No	15	**0·740**	**0·301**	0·412	0·244
Citizen	No	No	15	0·733	0·299	0·459	0·226
Citizen (*N* = Prof)	No	No	15	0·730	0·296	**0·466**	**0·246**
All	Yes	No	12	0·737	0·305	0·459	0·226
Professional	Yes	No	15	**0·738**	0·296	0·422	0·199
Citizen	Yes	No	20	0·735	**0·308**	**0·475**	***0·254***
Citizen (*N* = Prof)	Yes	No	15	0·732	**0·308**	0·466	0·246
All	No	Yes	12	0·739	0·302	0·464	0·199
Professional	No	Yes	12	***0·744***	**0·307**	0·440	0·176
Citizen	No	Yes	20	0·735	0·300	**0·464**	0·218
Citizen (*N* = Prof)	No	Yes	15	0·728	0·300	0·462	0·220
All	Yes	Yes	12	0·739	0·306	0·459	0·218
Professional	Yes	Yes	15	**0·740**	0·303	0·459	0·192
Citizen	Yes	Yes	20	0·735	0·308	0·466	**0·242**
Citizen (*N* = Prof)	Yes	Yes	15	0·732	***0·312***	***0·490***	0·237

*Best β-value, based on AUC (β considered = 1, 3, 6, 9, 12, 15, 20).

In contrast to previous work showing improved predictive ability from aggregating point-based occurrences into large grids cell (Fourcade *et al.* 2014), our results suggest that at best aggregating point-based occurrences into gridded cells produces only trivial improvements in model prediction. Unlike other studies that commonly used larger cells (≥1 km^2) for aggregation (Warton, Renner & Ramp 2013; Isaac *et al.* 2014), we selected smaller (25 ha) biologically relevant cells (average fox squirrel home range). Larger aggregations might have reduced more of the sample selection bias and increased the model's predictability, but this broader scale would have limited our ability to observe species' responses to the environment on the most biologically relevant scales.

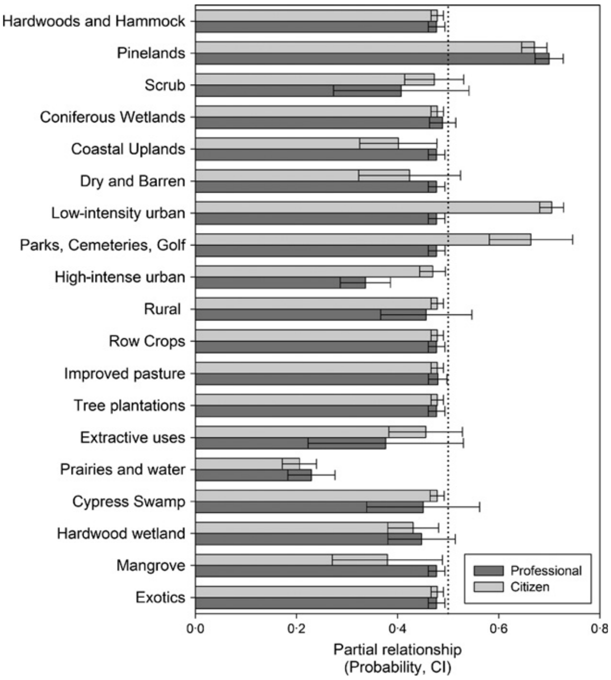

Fig. 3. Partial relationships from best models for professional and citizen data for land cover. Models based on subsampling to reduce bias, and uncertainty (95% CI) taken from a nonparametric bootstrap (*n* = 500 samples).

The data produced from our survey did not include the absences commonly used for species distribution models. While it is possible to infer absences from species lists and to explicitly model the probability of detection along with the factors that may influence it (Kery, Gardner & Monnerat 2010; Hochachka *et al.* 2012; van Strien, van Swaay & Termaat 2013), this approach is impractical for many species of conservation interest (e.g. carnivores, reptiles, marine mammals). Imperilled species such as Florida's fox squirrel are often found in low densities, have large ranges, and lack similarly sized conspecifics making it difficult and unrealistic to infer detection from species lists.

As predictors of relative fox squirrel activity, we found our best models provided useful ecological information about the factors that shape the fox squirrels' distribution. From our validation data, we found that when model probabilities were <0·4 we found few or no squirrels on validation plots (≈0·05) and when modelled probabilities were >0·4 the probability of squirrel presence on validation plots was ≈0·3. Fox squirrels occurred widely throughout Florida and were recorded in a variety of land cover types, but the highest levels of activity occurred in open pinelands (Fig. 3). Fox squirrels are believed to have evolved in savanna habitats and on the borders of the forest and prairie ecosystems (Steele & Koprowski 2001).

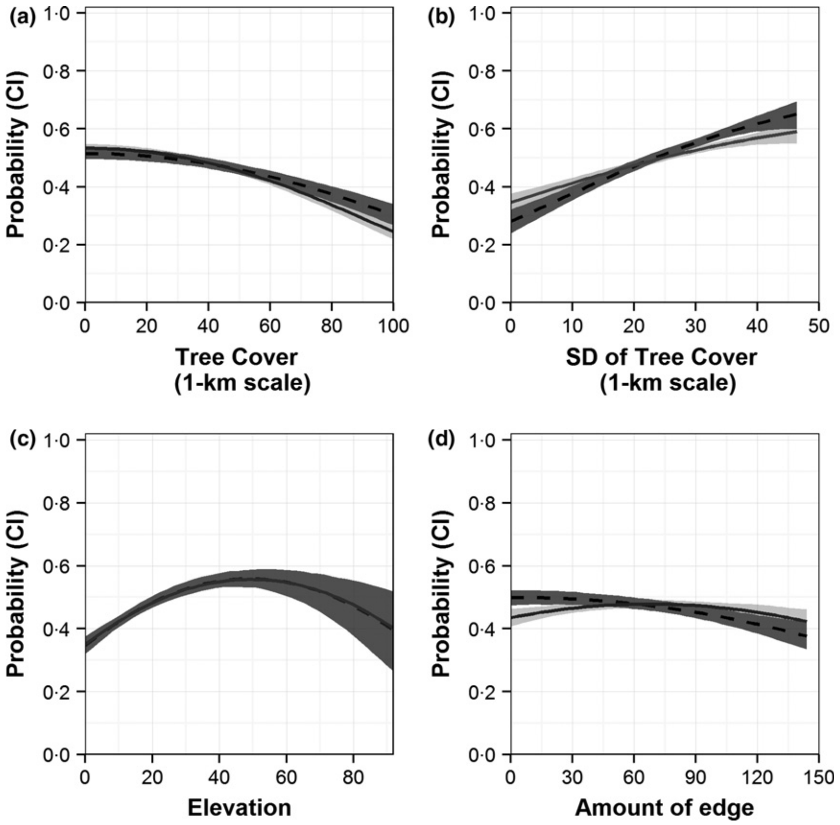

Fig. 4. Partial relationships (a) for tree cover, (b) SD of tree cover, (c) Elevation, and (d) edge from best models for professional data (solid line/light grey CI) and citizen data (dashed/dark grey CI). Models based on subsampling to reduce bias, and uncertainty (95% CI) taken from a non-parametric bootstrap (n = 500 samples).

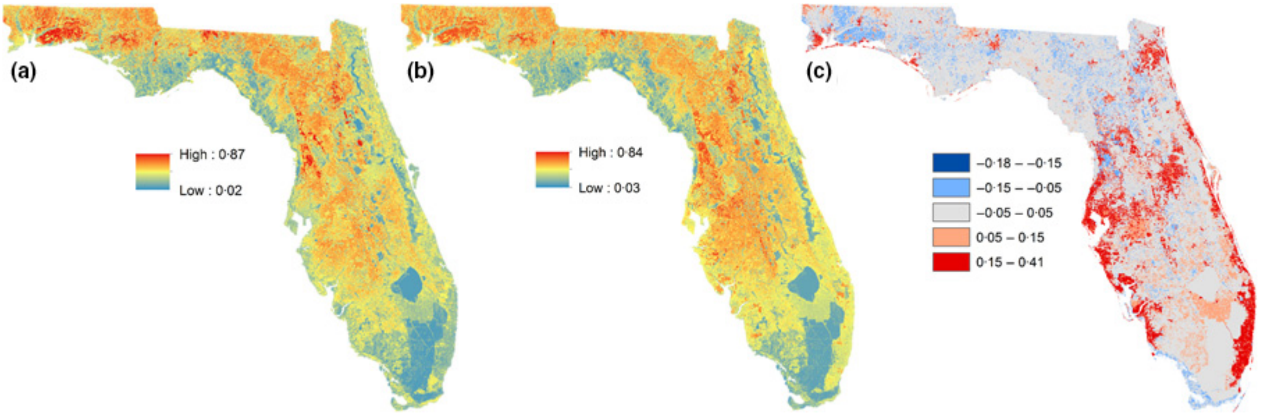

Fig. 5. Predictions from models collected with (a) professionals and (b) citizens, and (c) the difference in predictions between professionals and citizens. [Colour figure can be viewed at wileyonlinelibrary.com].

This might explain why fox squirrels avoided areas with a closed tree canopy and were more active in areas with a heterogeneous forest canopy (Fig. 4), indicative of the Florida's once vast pine savannas (Myers & Ewel 1990). This evolutionary history might also explain squirrel prevalence in areas of low-intensity development. These areas often have a broken canopy, open understorey and savanna-like vegetation structure (McCleery *et al.* 2012). Additionally, fox squirrels may benefit from the supplemental food, ornamental trees and well-watered landscaping in areas of low-level development (Jodice & Humphrey 1992; McCleery *et al.* 2007) that appear to have potential for fox squirrel conservation, as long as they do not replace natural pinelands or become more intensely urbanized (Fig. 3).

Citizens have an unmatched and growing (i.e. cell phone, GPS, cameras) ability to collect occurrence data across broad geographic areas, yet turning these data into useful information can be challenging. Using biased data and presence-only modelling, we were able to turn opportunistic data into important ecological information on the factors influencing the distribution of fox squirrels in Florida. Yet these models had only moderate predictive accuracy when validating against data from independent planned field surveys (Table 1). Moderate accuracy of distribution models has been observed in other situations where truly independent data are used for validation (e.g. McCarthy *et al.* 2012), such that it is unclear whether relatively low performance is from the use of volunteers or the rigorous application of an independent prospective sampling validation set.

For this study, data quality might have been improved if we required participants to zoom to a standard resolution when using the web tool to identify squirrel locations. However, placing constraints on volunteers has been shown to reduce the overall accuracy of citizen science projects (Lukyanenko, Parsons & Wiersma 2014). Similarly, because more data did not improve model performance, future efforts might consider using fewer citizens to collect presence/absence data in specific areas if the difficulty and constraints of this sampling do not jeopardize data quality (Lukyanenko, Parsons & Wiersma 2014). One potential way to improve the predictive accuracy of citizen-generated distribution models may be to integrate it with expert knowledge within a statistical framework (Drew & Perera 2011). Another avenue to enhance the predictive ability of opportunistic data is to model it in conjunction with planned presence–absence surveys. Modelling these two data sources together has been shown to generate more highly predictive distribution models than using the data sets by themselves (Fletcher *et al.* 2015). It precisely these types of synergies between citizens and professionals that are necessary to generate the information needed to develop conservation strategies for the planet's growing biodiversity crisis.

Acknowledgements

Funding for this study was provided by the Florida Fish and Wildlife Conservation Commission (Agreement 11427) and the Institute of Food and Agricultural Sciences at the University of Florida. We also that Y. Wiersma, F. Huettmann and an anonymous review for their constructive comments to improve the manuscript.

Data accessibility

Fox squirrels occurrence data are located in Dryad Digital Repository http://dx.doi.org/10.5061/dryad.8t475 (Tye *et al.* 2016). Our road data were obtained from the US Census Bureau's Florida 2013 Topologically Integrated Geographic Encoding and Referencing (TIGER) data layer (https://catalog.data.gov/dataset/tiger-line-shapefile-2013-state-florida-primary-and-secondary-roads-state-based-shapefile). We obtained tree canopy data from the 2011 National Land Cover Database (Homer *et al.* 2015), land cover data from the Florida Natural Areas Inventory (FNAI 2012) and elevation data from the US Geological Survey National Elevation Dataset (NED) for the state of Florida (Gesch *et al.* 2009).

References

Albert, C.H., Graham, C.H., Yoccoz, N.G., Zimmermann, N.E. & Thuiller, W. (2010) Applied sampling in ecology and evolution – integrating questions and designs. *Ecography*, **33**, 1028–1037.

Araújo, M.B., Pearson, R.G., Thuiller, W. & Erhard, M. (2005) Validation of species–climate impact models under climate change. *Global Change Biology*, **11**, 1504–1513.

Barbet-Massin, M., Jiguet, F., Albert, C.H. & Thuiller, W. (2012) Selecting pseudo-absences for species distribution models: how, where and how many? *Methods in Ecology and Evolution*, **3**, 327–338.

Ceballos, G., Ehrlich, P.R., Barnosky, A.D., García, A., Pringle, R.M. & Palmer, T.M. (2015) Accelerated modern human–induced species losses: entering the sixth mass extinction. *Science Advances*, **1**, e1400253.

Conrad, C.C. & Hilchey, K.G. (2011) A review of citizen science and community-based environmental monitoring: issues and opportunities. *Environmental Monitoring and Assessment*, **176**, 273–291.

Crall, A.W., Newman, G.J., Jarnevich, C.S., Stohlgren, T.J., Waller, D.M. & Graham, J. (2010) Improving and integrating data on invasive species collected by citizen scientists. *Biological Invasions*, **12**, 3419–3428.

Dennis, R.L.H., Sparks, T.H. & Hardy, P.B. (1999) Bias in butterfly distribution maps: the effects of sampling effort. *Journal of Insect Conservation*, **3**, 33–42.

Devictor, V., Whittaker, R.J. & Beltrame, C. (2010) Beyond scarcity: citizen science programmes as useful tools for conservation biogeography. *Diversity and Distributions*, **16**, 354–362.

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

Dickinson, J.L., Shirk, J., Bonter, D.N., Bonney, R., Crain, R.L., Martin, J., Phillips, T. & Purcell, K. (2012) The current state of citizen science as a tool for ecological research and public engagement. *Frontiers in Ecology and the Environment*, **10**, 291–297.

Drew, C.A. & Perera, A.H. (2011).Expert knowledge as a basis for landscape ecological predictive models. *Predictive Species and Habitat Modeling in Landscape Ecology* (eds C.A. Drew, Y. Wiersma & F. Huetmann), pp. 229–248. Springer, New York, USA.

Eisenberg, D.A., Noss, R.F., Waterman, J.M. & Main, M.B. (2011) Distribution and habitat use of the big cypress fox squirrel (*Sciurus niger avicennia*). *Southeastern Naturalist*, **9**, 75–84.

Elith, J., Graham, C.H., Anderson, R.P., Dudik, M., Ferrier, S., Guisan, A. *et al.* (2006) Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, **29**, 129–151.

Elith, J., Phillips, S.J., Hastie, T., Dudik, M., Chee, Y.E. & Yates, C.J. (2011) A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions*, **17**, 43–57.

Fielding, A.H. & Bell, J.F. (1997) A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, **24**, 38–49.

Fitzpatrick, M.C., Preisser, E.L., Ellison, A.M. & Elkinton, J.S. (2009) Observer bias and the detection of low-density populations. *Ecological Applications*, **19**, 1673–1679.

Fletcher, R.J., McCleery, R.A., Greene, D.A. & Tye, C. (2015) Integrated models that unite local and regional data reveal larger-scale environmental relationships and improve predictions of species distributions. *Landscape Ecology*, doi:10.1007/s10980-015-0327-9.

Florida Fish and Wildlife Conservation Commission. (2012). *Florida's Wildlife Legacy Initiative. Florida's Comprehensive Wildlife Conservation Strategy*. Florida Fish and Wildlife Conservation Commission. http://myfwc.com/media/134715/legacy_strategy.pdf.

Florida Fish and Wildlife Conservation Commission. (2013). *A Species Action Plan for the Sherman's Fox Squirrel Sciurus niger shermani*. Florida Fish and Wildlife Conservation Commission http://myfwc.com/media/2738277/Shermans-Fox-Squirrel-Species-Action-Plan-Final.pdf.

Florida Natural Areas Inventory. (2012). *Guide to the Natural Communities of Florida: 2010 edition*. Florida Natural Areas Inventory, Tallahassee, Florida, USA.

Fourcade, Y., Engler, J.O., Roedder, D. & Secondi, J. (2014) Mapping species distributions with MAXENT using a geographically biased sample of presence data: a performance assessment of methods for correcting sampling bias. *PLoS ONE*, **9**, e97122.

Gesch, D., Evans, G., Mauck, J., Hutchinson, J. & Carswell, W.J. Jr. (2009) The National Map—Elevation: U.S. Geological Survey Fact Sheet 2009-3053.

Grand, J., Cummings, M.P., Rebelo, T.G., Ricketts, T.H. & Neel, M.C. (2007) Biased data reduce efficiency and effectiveness of conservation reserve networks. *Ecology Letters*, **10**, 364–374.

Hastie, T. & Fithian, W. (2013) Inference from presence-only data; the ongoing controversy. *Ecography*, **36**, 864–867.

Hernandez, P.A., Graham, C.H., Master, L.L. & Albert, D.L. (2006) The effect of sample size and species characteristics on performance of different species distribution modeling methods. *Ecography*, **29**, 773–785.

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

Homer, C.G., Dewitz, J.A., Yang, L., Jin, S., Danielson, P., Xian, G. et al. (2015) Completion of the 2011 National Land Cover Database for the conterminous United States-Representing a decade of land cover change information. *Photogrammetric Engineering and Remote Sensing*, **81**, 345–354.

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

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

Jodice, P.G. & Humphrey, S.R. (1992) Activity and diet of an urban population of Big Cypress fox squirrels. *The Journal of Wildlife Management*, **56**, 685–692.

Kadmon, R., Farber, O. & Danin, A. (2004) Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. *Ecological Applications*, **14**, 401–413.

Kadoya, T., Ishii, H.S., Kikuchi, R., Suda, S. & Washitani, I. (2009) Using monitoring data gathered by volunteers to predict the potential distribution of the invasive alien bumblebee *Bombus terrestris*. *Biological Conservation*, **142**, 1011–1017.

Kantola, A.T. & Humphrey, S.R. (1990) Habitat use by Sherman's fox squirrel (*Sciurus niger shermani*) in Florida. *Journal of Mammalogy*, **71**, 411–419.

Kery, M., Gardner, B. & Monnerat, C. (2010) Predicting species distributions from checklist data using site-occupancy models. *Journal of Biogeography*, **37**, 1851–1862.

Kremen, C., Cameron, A., Moilanen, A., Phillips, S.J., Thomas, C.D., Beentje, H. *et al.* (2008) Aligning conservation priorities across taxa in Madagascar with high-resolution planning tools. *Science*, **320**, 222–226.

Liu, C., White, M. & Newell, G. (2011) Measuring and comparing the accuracy of species distribution models with presence–absence data. *Ecography*, **34**, 232–243.

Liu, C., White, M. & Newell, G. (2013) Selecting thresholds for the prediction of species occurrence with presence-only data. *Journal of Biogeography*, **40**, 778–789.

Loeb, S.C. & Moncrief, N.D. 1993. The biology of fox squirrels in the Southeast: a review. *Proceedings of the Second Symposium of Southeastern Fox Squirrels, Sciurus niger* (eds N.D. Moncrief, J.W. Edwards & P.A. Tappe), pp. 1–19. Special Publication 1, Virginia Museum of Natural History, Martinsville, Virginia, USA.

Lukyanenko, R., Parsons, J. & Wiersma, Y.F. (2014) The IQ of the crowd: understanding and improving information quality in structured user-generated content. *Information Systems Research*, **25**, 669–689.

McCarthy, K.P., Fletcher, R.J. Jr, Rota, C.T. & Hutto, R.L. (2012) Predicting species distributions from samples collected along roadsides. *Conservation Biology*, **26**, 68–77.

McCleery, R.A., Lopez, R.R., Silvy, N.J. & Kahlick, S.N. (2007) Habitat use of fox squirrels in an urban environment. *Journal of Wildlife Management*, **71**, 1149–1157.

McCleery, R.A., Moorman, C.E., Wallace, M.C. & Drake, D. (2012) Urban wildlife research and management. *Wildlife Techniques and Management* (ed N.J. Silvy), pp. 169–191. John's Hopkins Press, Baltimore.

Merow, C., Smith, M.J. & Silander, J.A. (2013) A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. *Ecography*, **36**, 1058–1069.

Myers, R.L. & Ewel, J.J. (1990) *Ecosystems of Florida*. University of Central Florida Press, Orlando, Florida, USA.

Nov, O., Arazy, O. & Anderson, D. (2011). Dusting for science: motivation and participation of digital citizen science volunteers. In *Proceedings of the 2011 iConference*. pp. 68–74.

Pearce, J.L. & Boyce, M.S. (2006) Modelling distribution and abundance with presence-only data. *Journal of Applied Ecology*, **43**, 405–412.

Phillips, S.J., Anderson, R.P. & Schapire, R.E. (2006) Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, **190**, 231–259.

Phillips, S.J. & Dudik, M. (2008) Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. *Ecography*, **31**, 161–175.

Phillips, S.J., Dudík, M., Elith, J., Graham, C.H., Lehmann, A., Leathwick, J. & Ferrier, S. (2009) Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological Applications*, **19**, 181–197.

Silvertown, J. (2009) A new dawn for citizen science. *Trends in Ecology & Evolution*, **24**, 467–471.

Steele, M.A. & Koprowski, J.L. (2001) *North America Tree Squirrels*. Smithsonian Books, Washington, District of Columbia, USA.

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

Tye, C.A., Greene, D.U., Giuliano, W.M. & Mccleery, R.A. (2015) Using camera-trap photographs to identify individual fox squirrels (*Sciurus niger*) in the Southeastern United States. *Wildlife Society Bulletin*, **39**, 645–650.

Tye, C.A., McCleery, R.A., Fletcher, R.J. Jr, Greene, D.U. & Butryn, R.S. (2016) Data from: Evaluating citizen vs. professional data for modelling distributions of a rare squirrel. *Dryad Digital Repository*, http://dx.doi.org/10.5061/dryad.8t475.

Warton, D.I., Renner, I.W. & Ramp, D. (2013) Model-based control of observer bias for the analysis of presence-only data in ecology. *PLoS ONE*, **8**, 1–9.

Weigl, P.D., Steele, M.A., Sherman, L.J., Ha, J.C. & Sharpe, T.L. (1989) The ecology of the fox squirrel (*Sciurus niger*) in North Carolina: implications for survival in the Southeast. *Bulletin of Tall Timbers Research Station*, **24**, 1–94.

Williams, K.S. & Humphrey, S.R. (1979) Distribution and status of the endangered Big Cypress fox squirrel (*Sciurus niger avicennia*) in Florida. *Florida Science*, **42**, 201–205.

Received 29 February 2016; accepted 29 April 2016
Handling Editor: Yolanda Wiersma

Supporting Information

Additional Supporting Information may be found in the online version of this article.

Appendix S1. Original Florida Natural Areas Inventories [FNAI] land covers and their corresponding consolidated classification.