Biological Conservation 173 (2014) 144–154

Contents lists available at ScienceDirect

Biological Conservation

journal homepage: www.elsevier.com/locate/biocon

Statistical solutions for error and bias in global citizen science datasets

Tomas J. Bird [a,b,*], Amanda E. Bates [b], Jonathan S. Lefcheck [c], Nicole A. Hill [b], Russell J. Thomson [b], Graham J. Edgar [b], Rick D. Stuart-Smith [b], Simon Wotherspoon [b], Martin Krkosek [d], Jemina F. Stuart-Smith [b], Gretta T. Pecl [b], Neville Barrett [b], Stewart Frusher [b]

[a] *School of Botany, University of Melbourne, Parkville, Victoria 3010, Australia*
[b] *Institute for Marine and Antarctic Studies, University of Tasmania, Hobart, Tasmania 7001, Australia*
[c] *Virginia Institute of Marine Science, College of William & Mary, PO Box 1346, Gloucester Point, VA 23062-1346, USA*
[d] *Department of Ecology and Evolutionary Biology, Ramsay Wright Zoological Laboratories, University of Toronto, 25 Harbord St., Toronto, ON M5S 3G5, Canada*

ARTICLE INFO

Article history:
Received 5 February 2013
Received in revised form 23 July 2013
Accepted 28 July 2013
Available online 25 September 2013

Keywords:
Volunteer data
Statistical analysis
Experimental design
Linear models
Additive models
Species distribution models
Biodiversity
Reef life survey

ABSTRACT

Networks of citizen scientists (CS) have the potential to observe biodiversity and species distributions at global scales. Yet the adoption of such datasets in conservation science may be hindered by a perception that the data are of low quality. This perception likely stems from the propensity of data generated by CS to contain greater levels of variability (e.g., measurement error) or bias (e.g., spatio-temporal clustering) in comparison to data collected by scientists or instruments. Modern analytical approaches can account for many types of error and bias typical of CS datasets. It is possible to (1) describe how pseudo-replication in sampling influences the overall variability in response data using mixed-effects modeling, (2) integrate data to explicitly model the sampling process and account for bias using a hierarchical modeling framework, and (3) examine the relative influence of many different or related explanatory factors using machine learning tools. Information from these modeling approaches can be used to predict species distributions and to estimate biodiversity. Even so, achieving the full potential from CS projects requires meta-data describing the sampling process, reference data to allow for standardization, and insightful modeling suitable to the question of interest.

1. Introduction

Evaluating global changes in the distribution and diversity of Earth's biota requires datasets of ambitious proportions where effort is shared over hundreds, or even thousands of individuals (Silvertown, 2009). In recent decades, volunteers, often labeled as 'citizen scientists' (CS), have been central to the collection of broad-scale datasets allowing the scientific community to address questions that would otherwise be logistically or financially unfeasible, even for the most dedicated scientific team (Dickinson et al., 2010; Stuart-Smith et al., in press). Consequently, volunteer networks provide an opportunity to answer conservation-related questions on the broad temporal and spatial scales that are relevant to understanding global biodiversity patterns. As proof of this concept, long-running volunteer monitoring programs have generated thousands of peer-reviewed papers (Sullivan et al., 2009) and can thus offer models for the development of similar programs in novel systems (Bonney et al., 2009).

As well as providing a practical means of addressing large-scale questions in ecology, involving citizens in the collection of data has a number of benefits to conservation-related projects. By being inclusive and engaging large numbers of people, CS projects can bring important publicity and discourse on conservation issues, and provide opportunities for the public to take an active role in management and conservation (Pattengill-Semmens and Semmens, 2003). Additionally, CS projects can often afford to be more exploratory than more regimented monitoring programs, making observations of rare events possible with sightings from large networks of volunteers that span broad spatial scales. Given these advantages, the capacity for addressing global-scale conservation may well rest in the realm of citizen science (Silvertown, 2009).

In spite of the proven success and potential for using CS datasets to address pressing global issues, there has been intense debate over the utility of such data in a scientific framework. Detractors suggest that involving large numbers of individuals with varying skill and commitment will lead to decreased precision in measurements such as in the identification or counting of species. Moreover, significant sources of bias may be present in the data, such as under-detection of species or the non-random distribution of effort (Crall et al., 2011). Such concerns have motivated CS projects to maximize the quality of data collected through

* Corresponding author. Address: School of Botany, University of Melbourne, Parkville, Victoria 3010, Australia. Tel.: +61 4 0949 6340.

E-mail address: tbird@student.unimelb.edu.au (T.J. Bird).

0006-3207/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.biocon.2013.07.037

improved sampling protocols and training (Edgar and Stuart-Smith, 2009), database management (Crall et al., 2011), and filtering or subsampling data to deal with error and uneven effort (Wiggins and Crowston, 2011; Wiggins et al., 2011). However, in many broadly distributed databases it may be impossible to implement rigid protocols or to eliminate all sources of error and bias. Thus, global CS datasets will likely violate the basic assumptions of some statistical analyses.

Fortunately, the issues of error and bias that are often present in CS data are not unique; analogous problems exist in datasets across a wide variety of disciplines and can be addressed using a suite of analytical approaches. In many cases, CS databases resemble the data collected for meta-analytical and landscape ecology syntheses where methods for accurately estimating and incorporating within-study or within-observer variability are key to drawing conclusions from the data (Hedges et al., 2010). For complex datasets, machine learning (ML) approaches are available that can examine the relative importance of large numbers of predictive variables in explaining the response data (Fink and Hochachka, 2012; Olden et al., 2008). Moreover, custom hierarchical analyses can recognize and account for the variable and clustered nature of CS data (Hochachka et al., 2012).

Here, our overall objective is to promote the use of CS data in conservation ecology and policy by highlighting how issues of data quality can be addressed using a suite of relatively new statistical tools. We first provide context by describing the main considerations for identifying and quantifying data quality issues present in CS data. Second, we explore a number of modeling approaches available for use with CS data with case examples to illustrate how specific issues of error and bias can alter understanding of biological patterns when left unaccounted for. Our perspective is that CS data has the potential to describe global patterns in biodiversity and the mechanisms driving change in ecosystems, communities and species. The inferential capacity to do so rests on the continued development and use of modeling approaches to identify and correct for data quality issues.

2. Contextualizing the quality issues present in citizen science data

Most CS projects recognize the potential issues of error and bias present when using large numbers of volunteers to collect data. Volunteer training, data standardization, validation and filtering procedures reduce potential sources of error and bias before, during and after the data are collected (Bonter and Cooper, 2012; Wiggins et al., 2011). In fact, studies comparing data generated by skilled volunteers vs. experts often show comparable estimates (e.g. Delaney et al., 2008; Edgar and Stuart-Smith, 2009). In spite of the best efforts of volunteers and researchers, two primary quality issues may still remain. First, CS data may still be prone to greater variability, or error, due to differences in the skills, dedication, and training of volunteer participants. Second, CS data may contain persistent bias. To address these quality issues, it is necessary to carefully consider the type of response data collected and how potential sources of error and bias might have been introduced during sampling.

2.1. Types of response data

Central to the design of CS studies is the consideration of what type of data to collect, as this will influence the kinds of questions that can be asked, what statistical tools are appropriate, and what additional information should be collected with each response data point for analyses (Wiggins and Crowston, 2011; Wiggins et al., 2011). At the same time, survey design and analysis also should acknowledge the limitations of data collection. For applications of CS data to conservation-related issues, inference is generally focused on describing changes in the locations and abundance of species, populations, and their associated habitats. Thus, response data in CS studies generally fall into the categories of presence-only, presence–absence, or some measure of quantity (such as abundance, percent cover or biomass), all recorded over time and space. Which kind of data are collected will depend on the scope of the study and the challenges associated with collecting the data.

Presence-only data require minimal effort to collect, and are therefore amenable to many CS applications that aim to recruit greater numbers of volunteers. However, the lack of information on where species were absent constrains what questions can be answered and the types of analyses available (Pearce and Boyce, 2006). Most significantly, presence-only samples are not representative of where the species (or event) was not found, which limits the predictive power of inference. For example, consider a walking club that is recruited to report sightings of a species of bird. In general, walkers are more likely to go to aesthetically interesting locations. Thus, the inferred distribution of bird species based solely on presence data will be concentrated at sites preferred by humans, when in fact the real distribution might be uniform in space. As well, because the amount of effort put into sampling is often directly tied to the locations of reported presences, any changes in effort may be interpreted as a change in the true distribution of a species.

By contrast, presence–absence (or occupancy) data provide information on the spatial and/or temporal distribution of a species, allowing for comparison of a species' occupancy status between different areas or times, such as for documenting range contractions associated with population declines (Tulloch et al., 2013). Similarly, abundance (or other measures of quantity) data are required to detect changes in the size of a population. However, presence–absence and abundance data have their limitations as well: in many cases, it is difficult to distinguish imperfect detections (i.e., failing to observe a species that is actually present) from true absences. Similarly, reported abundances often provide an underestimate of the true number of individuals present at a location. We discuss approaches to dealing with error in each of these kinds of data in Section 3.

2.2. Random error in citizen science datasets

The aim of much of ecological inference is to attribute variation in response data to one or more predictors. Random error is the variation in the response that cannot be described in terms of predictors. While some of this error may be due to factors of interest, sampling-related variability can contribute considerably to the observed response data. In the context of CS data, sampling error is often introduced when observers differ in their ability to detect, identify and quantify species or events. Mistakes can be introduced directly in the observation process, through measuring and recording covariate data (such as associated environmental data), or through variable execution of sampling protocols. If these sources of variation are not accounted for in a model, then they are included in the overall random error, which may obscure trends of interest. Large amounts of random error may not be an issue if the trend of interest is strong, but more usually results in more data being required to detect patterns. Fortunately, the increased quantity of data from CS programs can sometimes offset this issue, in contrast to the sometimes-limited quantity of data from more formal surveys.

Accounting for sources of random error requires measurements of both meta-data and covariates. Meta-data are measurements or classifiers related to sampling which help describe variation in how sampling was performed. As a start, each observation should be attributed an observer identifier. This identifier can then be used to relate metrics (such as observer training, frequency of involvement, or outside experience) to the response data and consequently quantify the overall effectiveness of a particular observer (Snäll et al., 2011). Measures of the effort spent conducting each survey are also useful for standardizing abundance or detection data (Bray and Schramm, 2001; Maunder and Punt, 2004). Covariates, on the other hand, include factors that are outside the realm of survey design, but which might still have significant impacts on the success of sampling. For instance, underwater visibility can greatly affect visual surveys undertaken by SCUBA, regardless of whether observers are experts or novices (Edgar and Stuart-Smith, 2009).

2.3. Bias

Random error can be biased or unbiased. In unbiased data, the random error is centered around zero. Bias occurs when this random error is consistently above or below zero due to some flaw in the data collection or estimation process, resulting in over- or under-estimates of the mean. There are many different ways bias can be introduced to a dataset, and identifying the processes which contribute bias is central to deciding what analytical approach to take. Here we differentiate systematic and sampling biases.

Systematic bias occurs when repeated measures of the same process provide consistent over or under-estimates of the true value. Imperfect detection in presence–absence data and species misidentification are examples of bias particularly common with CS data and they typically lead to incorrect estimates of species abundance and occurrence (Royle et al., 2007). Such biases can be nonintuitive. For example, in a survey in which volunteers identified birds from their calls, volunteers that self-identified as experts were more likely to falsely identify rare species than moderately skilled observers (Farmer et al., 2012). Another example of measurement bias occurs when divers are asked to estimate fish size. Typically, the size of small individuals are under-estimated while the size of large individuals are over-estimated, according to magnification and other factors affecting perception of size underwater. Either attempting to reduce the occurrence of such bias in data collection and/or calibration of data prior to analyses can be used to account for measurement bias. In the case of size estimation by divers, divers can be trained through practice with objects of known size, and/or size data can be transformed using known relationships between true and estimated sizes (Edgar et al., 2004).

By contrast, sampling bias occurs when some aspects of the process of interest are more likely to be sampled than others, so that the mean is overly influenced by these samples. One common source of bias for datasets collected by multiple observers is variability among observers in their sampling effectiveness. On average, the mean of measurements made by observers may be centered on the true value, but some observers may contribute more samples than others. In cases where observations are consistently over- or under-estimated by a particular observer, then considering each observation as an independent sample has the potential to bias the overall estimate of a mean or trend. Also, clustered sampling of a process that is auto-correlated in space or time (i.e., closely spaced observations are more alike than more distant observations) can introduce bias, as eventual understanding of the underlying process is dominated by information from the clustered areas that may tend to be more similar than if sampling was regular in spacing (Boakes et al., 2010). For example, bird surveys are often located in areas that are more accessible, such as sites near roads, which may in turn be associated with habitats preferable to certain species or population subsets (Lawler and O'Connor, 2004; Tulloch and Szabo, 2012). Volunteer effort may change over time due to seasonal windows or declining commitment, making it difficult to distinguish seasonal patterns from those due to effort expended (Ahrends et al., 2011; Seys et al. 2002).

3. Modeling approaches

Modern statistical tools present options for accounting for many types of error and biases. In the following sections, we describe a variety of such techniques that may be particularly relevant to CS data. We aim to indicate where and why one might use each tool, to describe the different approaches and illustrate applications by drawing on examples from the literature. Table 1 provides examples of freely available statistical packages for implementing many of the approaches we describe in the open-source program R (R Core Team, 2013). As well, we provide examples for how error and bias can be accounted for using selected subsets of the detailed global marine biodiversity dataset generated through the Reef Life Survey program (RLS, Edgar and Stuart-Smith, 2009). RLS uses intensively trained volunteer divers to quantify the abundance and diversity of fish and invertebrate species on replicate 50 × 5 m transects on rocky and coral reefs, using standardized visual census methods (details provided in the supplementary materials).

3.1. Linear and generalized linear models and extensions

Linear models and their extensions are some of the more widely used tools for quantifying random error in ecological data. The basic premise behind their use is that changes in the response data can be described as a linear function of predictors of interest, covariates or meta-data, called ‘fixed-effects’. Additive models extend linear models by allowing non-linear relationships between predictors and response data through the use of smoothing functions with multiple degrees of freedom (Hastie and Tibshirani, 1990). Put another way, a simple linear model with a single predictor and multiple covariates asks how much a change in that predictor would influence the response data if all other covariates were held constant. The strength of the relationship between two variables is summarized as a parameter. Thus, linear models and their extensions are often used in CS studies to control for sampling-related covariates when estimating the effects of predictors of interest (Table 2).

Often, a large amount of variation in the response data can be described using simple relationships. However, the response data are rarely fully explained by available predictors and covariates. Any variation that cannot be accounted for using parameters is modeled as though it were the result of a random process that can be described using a probability distribution. The goodness-of-fit of a model can then be described based on this remaining, or residual, variation in the data using likelihood based methods such as Akaike's Information Criterion (AIC).

Basic linear and additive models assume that the response data follow a normal or Gaussian distribution, which are suited to specific kinds of measurement data, but may not be suitable for other kinds of response data. Generalized linear and additive models (GLMs and GAMs) further extend linear and additive models to allow for other kinds of distributions, such as a Poisson or negative binomial regression for count data, or the logistic regression for binary data (Zuur et al., 2007). Many CS ecological datasets contain a large number of zero counts, which can violate the assumptions of the Poisson or negative binomial distributions. In this case, zero-inflated models can be useful for analyzing CS data (Arab et al., 2008). As well, autoregressive regression models, which model the change in similarity between more distant data points, can be used where closely-spaced samples are more likely to be similar

Table 1
Statistical approaches and software packages available for dealing with error and bias in citizen science data.

Method	R package	Package reference[a]
GLM	Base	R core team (2012)
GLMM	MCMCglmm	Hadfield (2010)
	lme4	Bates et al. (2012)
	glmmADMB	Skaug et al. (2011)
GAMM	mgcv	Wood (2011)
	gammSlice	Pham and Wand (2012)
Geographically-weighted regression	spgwr	Bivand (2013)
Spatio-temporal models	stem	Cameletti (2009)
Detection–occupancy	Unmarked	Fiske (2011)
Capture–recapture	Unmarked	
Bayesian hierarchical	R2WinBUGS, R2jags	Sturtz et al. (2005)
		Su and Yajima (2012)
Multiple ML approaches	RWeka	Hornik et al. (2009)
Mixed-effects trees	REEMtree	Sela and Simonoff (2012)
	longRPart	Stewart and Abdolell (2008)
Boosted regression tree (BRT)	gbm	Ridgeway (2013)
Classification and regression tree	tree	Ripley (2012)
	rpart	Therneau (2012)
Neural networks	nnet	Venables and Ripley (2002)
Richness and other indices	vegan	Oksanen (2012)
Ordination (NMDS, CCA, RDA)	vegan	Oksanen (2012)
Indicator species analysis	Indicspecies	De Caceres and Legendre (2009)
Modeling detectability	mrds	Laake et al. (2012)
Species distribution models	Biomod2	Thuiller et al. (2013)
BioClim	dismo	Hijmans et al. (2012)
Bayesian hierarchical SDM	hSDM	Vieilledent et al. (2012)
BRT and random forest mapped predictions	ModelMap	Freeman (2012)

[a] R package citations are available in the supplementary reference material.

to one another than those that are more distant (Legendre et al., 2002).

To show how different types of data can be accomodated using linear modeling, we present a subset of RLS data on sightings of the urchin genus *Holopneustes* along the east coast of Australia (Fig. 1A). We used the counts from the RLS data first as presence-only data (ignoring sites in which the genus was absent), second as presence–absence data (ignoring counts of the species within sites) and third as abundance data. We related each of these three kinds of response data to the maximum sea-surface temperature at each site to describe the range of temperatures occupied by *Holopneustes* spp. Using presence-only data, we find that the range occupied by the genus was between 17 and 25 °C (Fig. 1B). In comparison, using presence–absence data in a logistic regression model, we find that the probability of the genus occupying a site decreases as the sea-surface temperature increases, reflecting its increasing prevalence at more southern sites in Australia (Fig. 1C). Incorporating abundance data in a zero-inflated Poisson model shows that the temperature distribution of the genus displays two distinct peaks, likely corresponding to the gap between the core ranges of the two main species in the *Holopneustes* genus (Fig. 1D).

While attractive for their conceptual simplicity and broad applicability, GLMs and GAMs have limitations in terms of the numbers of predictors and covariates they can accommodate simultaneously. Thus, an important part of inference using linear or additive models (and their extensions, Section 3.2) is the process of determining which model provides the best fit with as few parameters as possible (Zuur et al., 2007). Where large numbers of predictors and covariates may be in play, ML approaches may be more suitable for inference (Section 3.4). As well, linear and additive models are generally not suitable for presence-only data, unless used in the context of species distribution models (SDMs; Section 3.5), an important consideration in the context of citizen-generated data. GLMs and GAMs are generally unreliable when the data are heteroscedastic, that is, the variance within the data is uneven across samples. To account for sampling bias in predictive models, tools such as mixed effects or hierarchical models are required.

3.2. Mixed-effects models

Where CS data are subject to sampling bias, mixed-effects models can be a powerful tool. Mixed-effects models include fixed effects used in linear or additive models with 'random-effects' that estimate the influence of predictors (often groups) that increase variability in the data but do not affect the mean response. For example, some observers in a study may have differing sampling efficiency – i.e., some over and some under-estimating a true value. A mixed-effects model would assume that if each observer contributed one sample, the mean of these observations would be centered on the true mean Zuur, 2009). However, if some observers contribute more samples than others, the contribution of these observers would skew the overall average, an effect that must be accounted for as with pseudo-replication in controlled experiments. Thus, we could use the observer identifiers as an index to model observer-to-observer variability before estimating the effects of other predictors in the model.

To demonstrate how sampling bias can influence inference and one way that this bias may be accounted for using linear mixed-effects modeling, we provide an example of a dataset with high variability among sampling sites and patchy sampling across latitude. In our example we plot species richness data of reef fish against latitude for a subset of the RLS dataset (selected purposely to illustrate uneven variance among groups of samples and differences in the means among sites). In Fig. 2A, we show a dataset that is clustered at two spatial scales; the bulk of the data are from lower latitudes and there is significant site-level pseudo-replication. Applying a linear model to the data (the nlme package in R (Table 1) using the function "lme" and fitted using maximum likelihood) provides a fit (AIC = 3472) with narrow confidence intervals around the model prediction. However, this narrow interval is largely an artifact of the large sample size; examination of the residuals shows a large discrepancy between the variance in different regions, violating the assumption of equal variance required for linear models (Fig. 2B). Including a random effect at the site level gives a marginally better fit (Fig. 2C, AIC = 3470), broader confidence intervals and centers the model predictions (Fig. 2D), however, there is still uneven variance between the high and low latitude sites that were sampled. Finally by using a variance-weighting model that accounts for the error structure among the four dominant regions of the data (Temperate Northern Pacific, Eastern Indo-Pacific, Temperate Northern Atlantic and Tropical Eastern Pacific), we arrive at a better-fitting model (Fig. 2E, AIC = 3381) that does not require the polynomial relationship between latitude and richness and that properly reflects the amount of variability in each region (Fig. 2F). We have therefore improved model fit by taking into account the clustered nature of the data collection and met the assumptions of the approach.

As extensions of GLMs and GAMs, generalized linear and additive mixed models (GLMMs and GAMMs) have proven extremely useful in ecological studies due to their flexibility and predictive power (Bolker et al., 2009). Thus, GLMMs and GAMMs have been used in CS data to accommodate observer bias and spatial cluster-

Table 2
Examples of CS studies that have used methods described in the text. For each study, the general class of method is listed, along with the source of the data (CS or otherwise), type of data and a description of the general class of issue addressed with the modeling approach. We also briefly summarize how the analysis helped inform the study results.

Model type	Source[a]	Data[b]	Issue	Study	Findings
GLM	CS	Size	Measurement error	Butt et al. (2013)	Measurements made by volunteers were not significantly different to those made by experts, after filtering
GLM	CS	PA	Identification	Delaney et al. (2008)	Age and education predicted rates of false identification of invasive crabs
GLM	CS	A	Identification	Crall et al. (2010)	Volunteers that were more confident performed better at species identifications
GLM	CS	PA	Detection	Sunde and Jesson (2013)	Experienced hunters were more likely to detect rabbits in spotlight surveys
GLM	CS	Size	Bias	Edgar et al. (2004)	Divers consistently over-estimated the sizes of fishes
GLMM	CS	P	Spatial clustering	Brunsdon and Comber, 2012	Onset of spring was shown to gradually advance over time when continental-scale spatial clustering was accounted for
GAM	CS	P	Spatial clustering	Fewster et al. (2000)	GAMMs reveal temporal trend in arrival time of bird species based on volunteer data
GLM	CS	P	Presence-only data	Parsons et al. (2009)	Targeted generation of pseudo-absences resulted in presence–absence data suitable for regression modeling.
GWR	CS	P	Spatial clustering	Comber et al. (2013)	Geographically-weighted regressions (GWR) and control data used to infer reliability of volunteered geographic information
Hierarchical	CS	PA	Detection	deSolla et al. (2005)	Survey effort is related to probability of detecting rare frogs from calls
Hierarchical	CS	PA	False-positive	Miller et al. (2011)	False-positive rates of bird classification by calls were related to distance, ambient noise and observer ability
Hierarchical	CS	PA	Spatiotemporal Clustering	Fink et al. (2010)	Modeling effort and detection in space and time led to improved models of species distribution
Hierarchical	DDB	A	Site-level bias	Amano et al. (2012)	Accounting for site-level effects allowed for more accurate estimation of population trends
Hierarchical	CS	PA	Detection	Kery et al. (2010b)	Accounting for detection in SDMs led to a 2-fold increase in estimated site occupancy
Hierarchical	CS	PA	Identification	Conn et al. (in press)	Hierarchical modeling allowed for estimation of species misidentification rates in double-observer surveys
Regression Tree	CS	A	Observer error	Cox et al. 2012	The differences in community similarity values among data collectors were not important
Regression splines	NHC	P	Spatial Clustering	Mateo et al. (2010)	Generating pseudo-absences using targeted rather than random approaches produced more accurate distribution models
MaxEnt	DDB	P	Spatial Clustering	Phillips et al. (2009)	Clustering pseudo-absences at the same scale as occurrence data results in more accurate distribution models
Diversity	CS	P	Detection	Holt et al. (2013)	Hierarchical models show that species richness estimates based on roving diver surveys were higher than those of standardized protocols.

[a] Data Sources include: Citizen Science (CS), Natural History Collections (NHC), Scientific survey (S), and distributed sampling databases (DDB).
[b] Data types include: presence-only (P), presence–absence (PA), abundance (A). For each paper we have included a result that shows how the analysis helped improve inference.

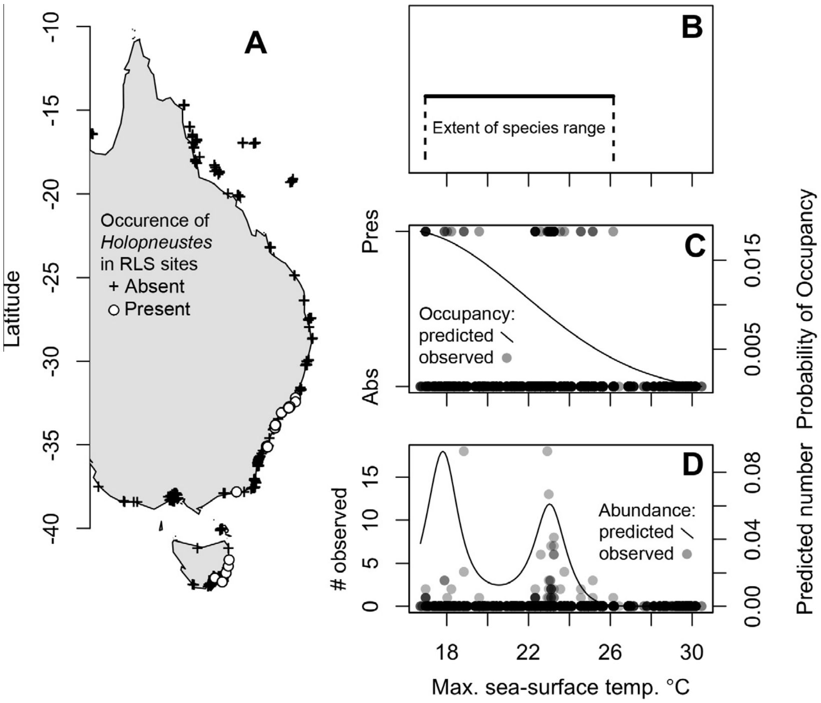

Fig. 1. (A) Occurrence of the urchin genus *Holopneustes* spp. along the east coast of Australia in RLS surveys. (B) The temperature range occupied by these species lies between 17° and 26°. (C) These species occupied 49 of 2008 surveys, leading to low predicted occupancy rates across the range of temperatures examined. (D) The number found per site is generally low, to a maximum of 18 individuals, resulting in low predicted numbers per site.

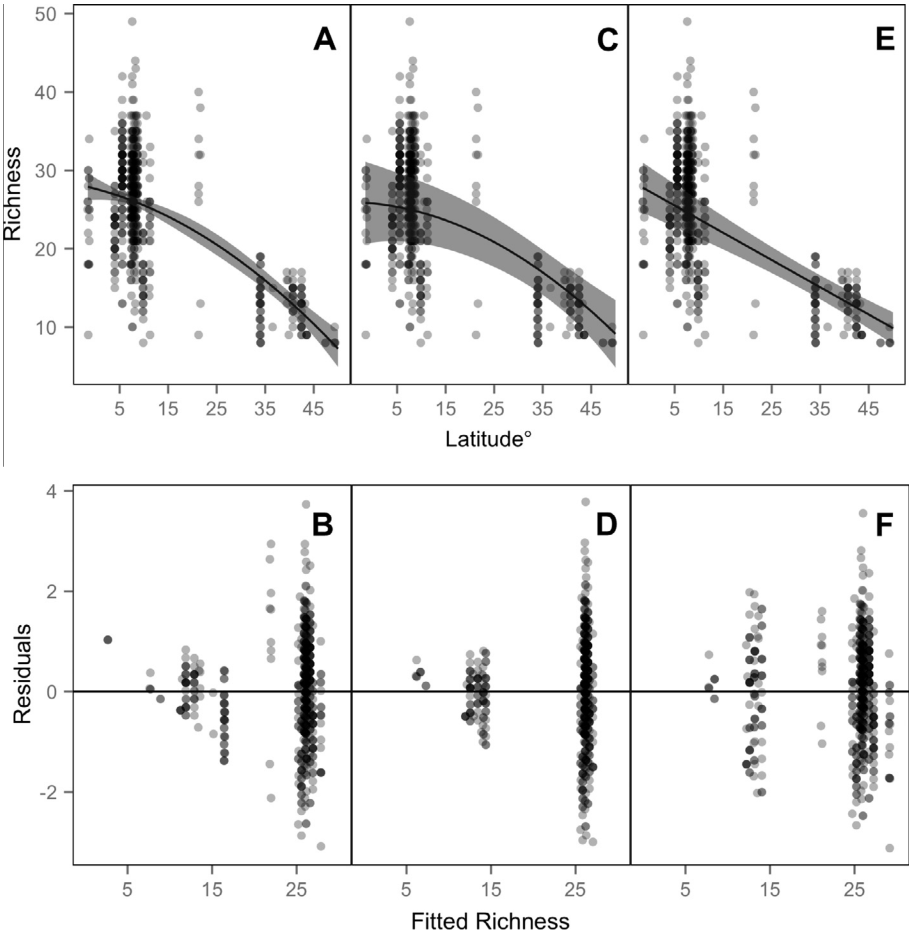

Fig. 2. Species richness of fish in the northern Pacific decreases with increasing latitude: Analysis by using linear model with the package "nlme" in R (A), linear regression with random effects at the site level (C) and with variance weighting (E). Predicted richness values (black line) and 95% confidence intervals (gray) are shown for each model. Residuals of the fitted values for each of the three models are shown in B,D,F. Points are 30% transparent to show areas of high data density.

ing (Table 2). However as in GLMs and GAMs, the number of predictors that can be included in models is limited by the amount of response data available and estimating the influence of random factors can require a great deal of replication within each factor level.

Thus, to avoid over-parameterizing the model, inference using mixed-effects models should include model selection using some measure of model fit such as AIC (Zuur, 2009). Finally, the assumption that random effects influence the variance but not the mean of the data ignores the possibility of measurement bias. We also note that while our example has shown how mixed-effects models can account for some kinds of sampling bias, systematic bias must be dealt with using other approaches. Hierarchical models may therefore be required to deal with sources of bias that cannot be accounted for with fixed or random-effects models.

3.3. Hierarchical models

Hierarchical models are a good choice for modeling CS data when the sampling design has some element of systematic bias that can be measured with data. Hierarchical models are similar to the models described above in that they are used to estimate parameters describing the relationship between predictor and response data using linear (or other) models. However, in hierarchical models the parameters themselves may be described as a function of other predictor variables (Royle and Dorazio, 2008). For example, in the previous section, we saw how sampling variability could be modeled separately between regions. As such, mixed-effects models represent a kind of hierarchical model and many other kinds of models can be adapted to match the specifics of CS surveys. Examples of ways to deal with systematic bias include models for imperfect detection, false-positives, and species misidentification (Table 2). As well, hierarchical Bayesian approaches are available to deal explicitly with spatially or temporally clustered data (Wikle, 2003). Hierarchical models, however, usually require specific sampling designs to accurately describe the sampling process (Royle and Dorazio, 2008).

Here, we show how not accounting for imperfect detection in sampling can result in drastic under-estimates of species occurrence. Again, we subsample from the RLS data to investigate how the presence or absence of the urchin genus *Echinostrephus* relates to maximum sea-surface temperature (Max_SST) on the east coast of Australia. A logistic regression estimates the influence of temperature on the probability of *Echinostrephus* occurring at a site, which is highest (∼60%) at higher temperatures (Fig. 3).

However, this model ignores the possibility that the urchin may have gone undetected in some transects. *Echinostrephus* species are small, burrow, and are patchily distributed at local scales, meaning that patches of few individuals may easily be overlooked. Our hierarchical model takes advantage of the fact that multiple transects were laid at some sites and employs an occupancy-detection model (MacKenzie, 2006) to estimate the probability of detecting these urchins. We do so by assuming that the site-level occupancy of *Echinostrephus* is known to be 1 if it is found at one transect within a site. From this assumption and the known number of transects used within a site, we can estimate the probability of observing the urchin given that it is present. Thus, the observed data at each site now becomes the outcome of one or two attempts to find the urchin, with the number of successes determined by both the product of the probability of occurrence (which we still assume is related to temperature) and the probability of detecting the urchin. We fit this model using Markov-chain Monte-Carlo (MCMC) sampling in the BUGS programming language (Lunn et al., 2000), and find that by accounting for low detection rates, the occupancy rate of *Echinostrephus* is almost double that estimated by the logistic regression (Fig. 3, dashed line).

We note here that in the case of *Echinostrephus* spp. the detection rate that we are estimating at the site-level is confounded with the patchiness of the genus. Thus our example shows how replication can be used to build a hierarchical model, but also demonstrates how different kinds of error can be additive. In our case, site-level replication allows for explicit modeling of the observation process, resulting in a more realistic modeling approach. Statistical packages are available to perform hierarchical analyses using similar syntax to well-known linear and additive models (Table 1), and the development of more complex models can be accommodated using the BUGS programming language.

3.4. Machine learning

In cases where many predictor variables are of interest and may be correlated, ML approaches can be particularly useful (De'ath and Fabricius, 2000). In CS data, there can be many competing factors influencing the response data and there is a risk of building models with more parameters than can be supported by the data. Some ML approaches bypass many of the assumptions required by the models described in Sections 3.1–3.3, by ignoring the need for the response data to fit any particular probability distribution, though, options such as boosted regression trees (BRT) may use different algorithms (and perform better) for different kinds of response data.

Machine learning approaches use heuristic algorithms to learn about the most likely relationship between predictors and response data (Olden et al., 2008). For example, a classification tree might split the proportions of observed presences in presence/absence data based on whether the observer was experienced or novice. Because these rules are not based on rigid probabilistic assumptions about the distribution of the response, ML approaches may be more suited to CS data that were collected under a sampling design that might violate the assumptions of classical experimental design.

Applications of ML are available for presence-only, presence–absence, abundance and other data types (Table 1). As well, many ML approaches do not assume that the relationships between responses and predictors are linear (or even smooth). Many available methods have been applied in an ecological setting, including classification and regression trees (CART, De'ath and Fabricius, 2000), boosted regression trees (BRT, Elith et al., 2008), random forests (RF, Cutler et al., 2007), artificial neural networks, and genetic algorithms (Olden et al., 2008).

In our example, we use a random forests (RF) approach to predict the presence/absence of sharks using RLS data. The worldwide RLS dataset has surveys nested within sites, which are nested within eco-regions. The unmodified RF procedure assumes all observations are independent, ignoring possible bias due to within-site pseudo-replication. It is possible to account for non-independence in the data by aggregating observations up to a higher level (Fig. 4). The Receiver-Operator Curves (ROC) shown in Fig. 4 show how aggregating observations at different levels improve model performance, with curves that have a greater area under curve (AUC – a measure of the discriminatory power of the model) providing greater predictive power. In our case model performance is greatest when samples are grouped at the site-level, albeit with a reduction in sample size. Details of the RF approaches and ROC curves used in Fig. 4 are available in the SOM.

The ROC curves in Fig. 4 were obtained from a cross-validation technique that is part of the RF method, so that predictions at a survey are independent from the models developed using a particular survey. However, predictions at a survey location could be based on nearby surveys, which could introduce a spatial bias. Consequently the performance of non-aggregated methods could be over-estimated. In spite of this, the RF method used here shows how site aggregation can be used to remove pseudo-replication.

A drawback of ML approaches is that they generally do not provide easy ways to deal explicitly with uncertainty in the model, data or parameters. As such, it can be difficult to determine the

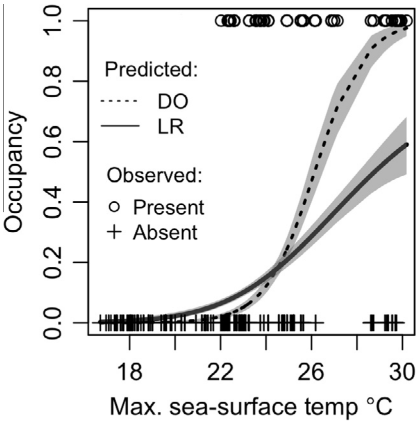

Fig. 3. Relationships between estimated occupancy rates and maximum sea-surface temperature for the sea urchin genus *Echinostrephus* found in RLS surveys along the east coast of Australia. Solid line indicates an estimate based on a logistic regression (LR) between Max SST and occupancy, while the dashed line is the estimated probability of occupancy from a detection–occupancy (DO) model which takes into account failure to detect the genus given that it was present at a site. Gray shading indicates 95% Bayesian credible intervals around the estimated trend. Points indicate temperatures at which the urchin was (o) or was not (+) found.

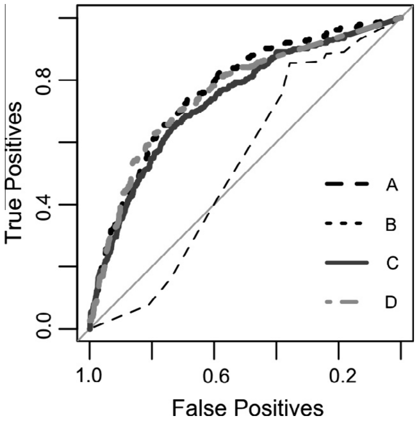

Fig. 4. Receiver Operator Characteristic (ROC) curves for estimated presence/absence of sharks found in RLS sites worldwide using random forests (RF) at different scales. (A) Regression RF on the average presence at an eco-region, area under the curve (AUC, (95% CI*) = 0.649 (0.62–0.67)). (B) Regression RF on the average presence at a site (AUC = 0.814 (0.80–0.83)). (C) Classification RF on the presence/absence at a site, where one survey (with depth closest to 6 m) is sampled for each site (AUC = 0.78 (0.76–0.8)). (D) Classification RF on the presence/absence at a survey (AUC = 0.809 (0.79–0.83)).

reliability of results derived by ML methods that do not provide confidence intervals or standard errors. BRT approaches have been developed to allow a more probabilistic style of inference using ML (Elith et al., 2008). Several novel approaches for dealing with bias are also being developed, including mixed-effects regression tree (which allows for hierarchical clustering of the response data) (Sela and Simnoff, 2012). Another novel approach to dealing with clustered data is a spatio-temporal exploratory model (STEM) framework which breaks the data into discrete but overlapping spatial and temporal units that are modeled locally (using bagged trees in this instance) and then aggregated (Fink et al., 2010). Alternatively, pseudo-replication can be accounted for by altering the bootstrapping step in random forests, so that the bootstrap sampling is at a higher level (Karpievitch et al., 2009). Interestingly, when this method was used on a dataset that was cluster-correlated as CS data often are, Karpievitch et al. found no difference in classification accuracy over the unmodified random forest model, but a significant improvement in predictive ability, a result that highlights the importance of checking whether particular approaches are suitable to each dataset.

3.5. Estimating biodiversity

One common aim in many large-scale CS projects is to compare different habitats in terms of their species composition. Biodiversity indices describe species (and functional/phylogenetic) diversity within ecological communities. Numerous indices are available ranging from species richness (the number of species in a site or sample), to more complicated indices incorporating information on species' relative abundances (e.g., Shannon or Simpson), functional traits (Petchey and Gaston, 2006) or phylogenetic relationships (e.g., Cadotte et al., 2010).

Some species are more cryptic than others and as a consequence biodiversity indices can be heavily influenced by variation in sampling and detectability. To account for error and bias in biodiversity measures, the calculated indices can be treated as response data, as in Fig. 2, and analyzed using approaches such as linear modeling. Alternatively, error and bias correction measures can be applied at the species level in a hierarchical model (such as by using a detection–occupancy model) and the diversity indices calculated as a derived parameter (Gelfand et al., 2005; Holt et al., 2013; Kery et al., 2010a).

Various diversity indices also emphasize the contributions of rare species differently, and the choice of index used may also help minimize issues of detectability. A simple solution is to choose a metric that emphasizes abundant species (e.g., Simpson index) to down-weight the influence of rare or poorly detected species. Additionally, rarefaction is often used on biodiversity data to account for uneven sampling effort. Traditional rarefaction generates species accumulation curves, and then reduces the largest samples until they are equivalent in size to the smallest (Gotelli and Colwell, 2001).

New methods employ what is called "shareholder quorum subsampling" (Alroy, 2010) or "fixed coverage subsampling" (Chao and Jost, 2012), which extrapolate richness outwards and then scale back based on a measure of sample 'completeness.' These methods are less biased, have ideal mathematical properties, and minimize the amount of discarded data and sampling effort. Recent work has extended this framework to include effective numbers, which are increasingly being used to compare different dimensions of biodiversity (Chao et al., 2013).

In Fig. 5, we present species richness of fish aggregated within two RLS sites in New Zealand. The Shortland Bluff site has much greater richness ($S = 54$) compared to the Goat Island site ($S = 18$, Fig. 5A). Taking the traditional rarefaction approach, we scale richness back to the fewest number of observed individuals: 68, in the Goat Island sample. In this case, the estimated richness for the Shortland Bluff site is approximately equal to that in the Goat Island Site: $S = 22$ vs. 18, respectively. Taking a coverage-based approach, we first extrapolate outwards (dashed line, Fig. 5A) and calculate the coverage, or proportion of individuals in the sample that belong to species in the sample. Subtracting the coverage from unity yields the probability that a new species would be found if an additional individual was sampled, and is equivalent to the final slope of the rarefaction curve in Fig. 5A. Scaling back to the lowest degree of coverage (approximately 93%, Fig. 5B), we see that the estimated richness for Shortland Bluff is now twice that of Goat Island: $S = 39$ vs. $S = 18$, respectively. Using the coverage-based approach, we have used more of the available data, and provided a less biased interpretation of the difference in richness between the two sites.

3.6. Species distribution models

Species distribution models (SDMs) use spatial occurrence or abundance datasets to describe or predict species' distributions in unsampled space. The basic premise is to use one of the modeling approaches described above to characterize the relationship between species data and a series of environmental predictor variables. This model can then be used to predict the likely distribution of species (or communities) in unsampled space or time (Elith et al., 2006; Ferrier and Guisan, 2006; Franklin, 2009). A broad range of modeling techniques are applied to SDMs, including many of the parametric and ML methods discussed above. Large and broad-scale datasets such as those collected by citizen science programs are a natural place to use SDMs as they can be compared against extensive geographical datasets using GIS. As a consequence SDMs are gaining popularity in conservation ecology (Ashcroft et al., 2012; Sarda-Palomera et al., 2012).

Given that most SDMs use linear, additive or ML models to make predictions into unsampled space, it is possible to address random error and bias appropriate for each method using metadata and covariates where possible. However, this approach may be limited for use in predictive SDMs because the sampling-related fixed and random effects may not be defined in the space for which predictions are being made. Occupancy or abundance predictions can be made by (1) averaging across values for each sampling-related effect (representing, for example, predictions across the typical observer or survey period), (2) omitting them (random effects only) or (3) a combination of the two (Welham et al., 2004). In practice, however, random error of the kind encountered in CS data is often reduced as much as possible by screening the data before analysis. Detection–occupancy modeling has been used successfully within SDMs (Kery et al., 2010b) to account for imperfect detection rates where repeat observations are available. Additional research is needed on how best to account for observation errors in SDMs where the underlying data do not have repeat observations (Monk, 2013).

Approaches for dealing with sampling biases in CS data for SDM applications have focused on addressing uneven spatial and temporal sampling effort, and include subsampling to reduce the overall variability in sampling effort (Segurado et al., 2006), potentially at the expense of large amounts of data, or down-weighting heavily sampled areas to reduce their influence in models (Dudík et al., 2005). Alternatively, autoregressive models and other spatially explicit models may be useful for dealing with these biases (Dormann et al., 2007). Similarly, hierarchical models can incorporate spatial structures and extensions of detection/occupancy models are possible to simultaneously account for both observation error and spatial and/or temporal bias (Gelfand et al., 2005; Latimer et al., 2006).

Predictive SDM models are also available to deal with presence-only data through programs such as BIOCLIM (Busby, 1991) and

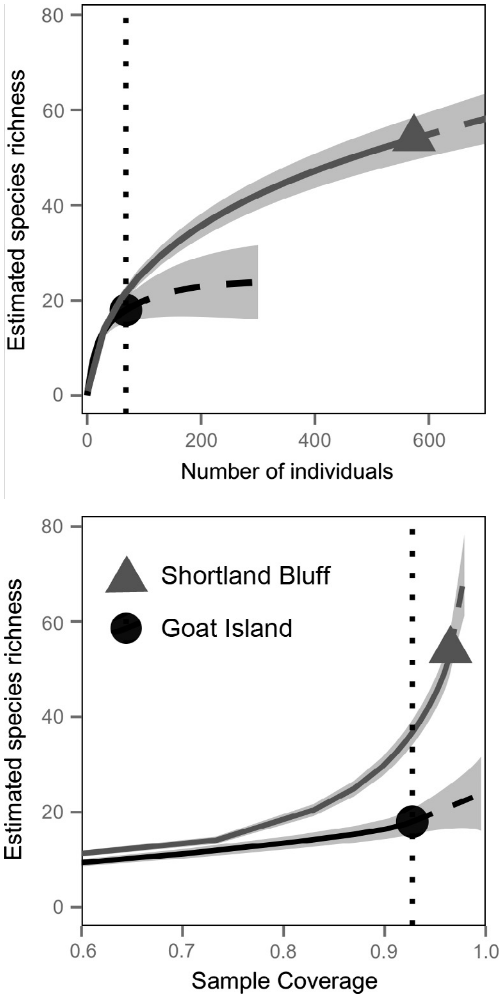

Fig. 5. Estimated species richness for two sites from the Reef Life Survey: Goat Island and Shortland Bluff. (A) Traditional rarefaction scales estimates back to number of individuals in the smallest sample (vertical dotted line). Dashed lines indicate extrapolated richness (i.e., species accumulation curves). (B) Coverage-based rarefaction scales estimates back to the lowest level of sample coverage (vertical dotted line). In both panels, shaded areas represent 95% confidence intervals.

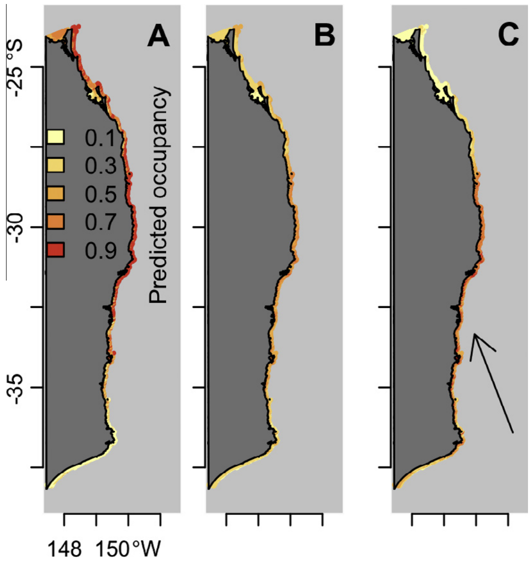

Fig. 6. Predicted probability (and likelihood in the case of presence-only models) of occurrence of *P. unifasciata* using three different modeling scenarios; (A) presence–absence data, (B) presence-only data with pseudo-absences drawn from the study region weighted by their probability of being sampled based on the distribution of sampled sites (targeted background) and (C) presence-only data with pseudo-absences drawn from the study region at random (random background). Arrow indicates north, and figures have been rotated to optimize space usage.

HABITAT (Walker and Cocks, 1991), which calculate the likely environmental limits of a species. Alternatively, SDMs based on presence-only data have used entropy modeling (MAXENT, Phillips et al., 2006) or maximum likelihood (MAXLIKE, Royle et al., 2012) to generate pseudo-absences to compare against observed presences in something like a logistic regression. Highly clustered presence-only data, which are particularly prone to bias, have received recent attention in SDMs. Presence-only methods such as MAXENT are particularly sensitive to sampling bias (Yackulic et al., 2013). Recent work suggests that generating pseudo-absence data that are spatiotemporally biased in the same way as the observation data may improve the performance of predictive models (Barbet-Massin et al., 2012; Phillips et al., 2009). However, care needs to be taken when interpreting the outputs of presence-only models as unless additional data on prevalence are available, then models represent relative (rather than absolute) probability of presence (Phillips and Elith, 2013).

In Fig. 6 we use boosted regression trees to predict the occurrence of a common shallow, rocky-reef fish, *Parma unifasciata* on the East coast of Australia based on environmental covariates (Table S1). We take RLS data and create three modeling scenarios; one where we have presence–absence data (PA), another where we keep only the presence data (PO) and randomly select pseudo-absences from all available sites in the study region (random background) and the third where we use PO data and weight our random selection of pseudo-absences using an additional model that describes the likelihood that a site is sampled (targeted background), thus simulating the biases present in the original dataset (following Phillips et al., 2009). We generated 100 datasets for each PO modeling scenario (Fig. S1) and evaluated each against 30% of the data set aside for validation. Using both AUC and correlations between predicted and observed presence–absence data, we found that the presence–absence model performs the best, followed by the PO model with a targeted selection of background pseudo-absences, although the values for both PO scenarios are similar and lower than the PA scenario (Table S2). Maps of the predicted distribution of *P. unifasciata* show that it is most likely to occur in the center of the study region in all models (Fig. 6). Probability of occurrence is also relatively high in the PA model at several northern sites (Fig. 6A), which, relatively speaking, is captured better by the targeted background PO model (Fig. 6B), and may account for more of the original bias in sampling site distribution than the PO random background model (Fig. 6C).

4. Recommendations

There is great potential for the use of CS data as a mainstream tool to address the important ecological and conservation questions of our time. However, in order to do so, researchers will need to consider some basic principles of data collection, management and analysis. Taking an overview of recent techniques used in research based on CS data (Table 2) and incorporating the advice found in Zuur et al. (2010), we have extracted a few recommendations.

First, working with both statisticians and volunteers will help build an understanding of the likely constraints around sampling, and may require some trial and error. Given the broad array of possible modeling approaches available, it is important to consider the main issues with the dataset, how they will affect the question being asked and then to choose the best method to deal with those issues. Ideally researchers using CS datasets would design their sampling program to collect that data needed to account for such issues ahead of time. At the same time, the design of CS studies must meet the needs of the question being asked, while acknowledging tradeoffs between data quality and quantity that are likely to occur with CS data.

Next, it is vital to record data on aspects of the environment or survey execution (such as observer i.d.) that are likely to influence the results. While standardized data collection procedures will help ensure that volunteers are, to the best of their abilities, collecting data in the same way, true uniformity in sampling is unlikely. Recording meta-data can also help account for pseudoreplication due to clustered sampling.

Finally, where measurement bias is a potential issue, it is important to consider whether it is possible to collect data that will allow characterization of this bias. Using such data, it may be possible to use validation approaches within data collection, or hierarchical modeling to correct or account for such bias. Useful procedures might include re-sampling areas with known quantities, using training datasets, or performing multiple-observer surveys.

In closing, the challenges associated with analyzing CS databases present an exciting opportunity for collaboration between statisticians and conservation scientists. We anticipate the development of novel statistical approaches and survey designs that will break new ground in overcoming some of the problems we have outlined in this paper.

Acknowledgments

T.J.B. was supported by an Australian Research Council Linkage doctoral fellowship and the Australian Research Council Centre for Excellence in Environmental Decisions. Salary for A.E.B was by the Fisheries Research Development Corporation, the National Climate Change Adaptation Facility, and the Australian National Network in Marine Science (a collaborative funding initiative between James Cook University, the University of Tasmania, and the University of Western Australia). Development of the RLS dataset was supported by the former Commonwealth Environment Research Facilities Program, while analyses of this dataset were supported by the Marine Biodiversity Hub, a collaborative partnership supported through the Australian Government's National Environmental Research Program. In addition, we thank the many Reef Life Survey volunteers who contributed their time and energy to collecting these data, as well as several anonymous reviewers for their thoughtful comments.

Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.biocon.2013.07.037.

References

Ahrends, A., Fanning, E., Rahbek, C., Burgess, N.D., Gereau, R.E., Marchant, R., Bulling, M.T., Lovett, J.C., Platts, P.J., Wilkins Kindemba, V., Owen, N., 2011. Funding begets biodiversity. Divers. Distrib. 17, 191–200.

Alroy, J., 2010. Geographical, environmental and intrinsic biotic controls on Phanerozoic marine diversification. Palaeontology 53 (6), 1211–1235.

Amano, T., Okamura, H., Carrizo, S.F., Sutherland, W.J., 2012. Hierarchical models for smoothed population indices: the importance of considering variations in trends of count data among sites. Ecol. Ind. 13, 243–252.

Arab, A., Wildhaber, M.L., Wikle, C.K., Gentry, C.N., 2008. Zero-inflated modeling of fish catch per unit area resulting from multiple gears: application to channel catfish and shovelnose sturgeon in the Missouri River. N. Am. J. Fish. Manage. 28 (4), 1044–1058.

Ashcroft, M.B., Gollan, J.R., Batley, M., 2012. Combining citizen science, bioclimatic envelope models and observed habitat preferences to determine the distribution of an inconspicuous, recently detected introduced bee (*Halictus smaragdulus* Vachal Hymenoptera: Halictidae) in Australia. Biol. Invasions 14, 515–527.

Barbet-Massin, M., Jiguet, F., Elbert, C.H., Thuiller, W., 2012. Selecting pseudo-absences for species distribution models: how, where and how many? Methods Ecol. Evol. 3, 327–338.

Boakes, E.H., McGowan, P.J.K., Fuller, R.A., Chang-qing, D., Clark, N.E., O'Connor, K., Mace, G.M., 2010. Distorted views of biodiversity: spatial and temporal bias in species occurrence data. PLoS Biol. 8, e1000385.

Bolker, B.M., Brooks, M.E., Clark, M., Connie, J., Geange, S.W., Poulsen, J.R., Stevens, H.H., White, J.-S.S., 2009. Review: generalized linear mixed models: a practical guide for ecology and evolution. Trends Ecol. Evol. 24, 127–135.

Bonney, R., Cooper, C.B., Dickinson, J., Kelling, S., Phillips, T., Rosenberg, K.V., Shirk, J., 2009. Citizen science: a developing tool for expanding science knowledge and scientific literacy. Bioscience 59, 977–984.

Bonter, N.D., Cooper, C.B., 2012. Data validation in citizen science: a case study from project FeederWatch. Front. Ecol. Environ. 10 (6), 305–307.

Bray, B.S., Schramm Jr., J.L., 2001. Evaluation of a statewide volunteer angler diary program for use as a fishery assessment tool. N. Am. J. Fish. Manag. 21 (3), 606–615.

Brunsdon, C., Comber, L., 2012. Assessing the changing flowering date of the common lilac in North America: a random coefficient model approach. Geoinformatica 16, 675–690.

Busby, J.R., 1991. BIOCLIM — A bioclimate analysis and prediction system. Nature conservation: cost effective biological surveys and data analysis. In: Margules, C.R., Austin, M.P. (Eds.), Nature Conservation. CSIRO, Melbourne, pp. 64–68.

Butt, N., Slade, E., Thompson, J., Malhi, Y., Riutta, T., 2013. Quantifying the sampling error in tree census measurements by volunteers and its effect on carbon stock estimates. Ecol. Appl. 23 (4), 936–943.

Cadotte, M.W., Davies, J.T., Regetz, J., Kembel, S.W., Cleland, E., Oakley, T.H., 2010. Phylogenetic diversity metrics for ecological communities: integrating species richness, abundance and evolutionary history. Ecol. Lett. 13 (1), 96–105.

Chao, A., Jost, L., 2012. Coverage-based rarefaction and extrapolation: standardizing samples by completeness rather than size. Ecology 93, 2533–2547.

Chao, A., Gotelli, N., Hsieh, T.C., Sander, E., Ma, K.H., Colwell, R.K., Ellison, A.M., 2013. Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies. Ecol. Monogr..

Comber, A., See, L., Fritz, S., Van der Velde, M., Perger, C., Foody, G., 2013. Using control data to determine the reliability of volunteered geographic information about land cover. Int. J. Appl. Earth Obs. Geoinf. 23, 37–48.

Conn, P., McClintock, B.T., Cameron, M., Johnson, D.S., Moreland, E., Boveng, P.L., in press. Accomodating species identification errors in transect surveys. Ecology.

Cox, T.E., Philippoff, J., Baumgartner, E., Smith, C.M., 2012. Expert variability provides perspectives on the strengths and weaknesses of citizen-driven intertidal monitoring program. Ecol. Appl. 22 (4), 1201–1212.

Crall, A.W., Newman, G., Stohlgren, T.J., Holfelder, K.A., Graham, J., Waller, D.M., 2011. Assessing citizen science data quality: a case study. Conserv. Lett. 4, 433–442.

Cutler, A., Cutler, D.R., Edwards, T.C., Beard, K.H., Hess, K.T., Gibson, J., Lawler, J.J., 2007. Random forests for classification in ecology. Ecology 88, 2783–2792.

De'ath, G., Fabricius, K.E., 2000. Classification and regression trees: a powerful yet simple technique for ecological data analysis. Ecology 81, 3178–3192.

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

deSolla, S.R., Shirose, L.J., Fernie, K.J., Barrett, G.C., Brousseau, C.S., Bishop, C.A., 2005. Effect of sampling effort and species detectability on volunteer based anuran monitoring programs. Biol. Conserv. 121, 585–594.

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

Dormann, F.D., McPherson, J.M., Araújo, M.B., Bivand, R., Bolliger, J., Gudrun, C., Davies, R.G., Hirzel, A., Jetz, W., Kissling, D., Kühn, I., Ohlemü ller, R., Peres-Neto, P.R., Reineking, B., Schröder, B., 2007. Methods to account for spatial autocorrelation in the analysis of species distribution data: a review. Ecography 30, 609–628.

Dudík, M., Phillips, S.J., Schapire, R.E., 2005. Correcting sample selection bias in maximum entropy density estimation. In: Advances in Neural Information Processing Systems. MIT Press, Cambridge, Massachusetts, USA, pp. 323–330.

Edgar, G.J., Stuart-Smith, R.D., 2009. Ecological effects of marine protected areas on rocky reef communities; a continental-scale analysis. Mar. Ecol. Prog. Ser. 388, 51–62.

Edgar, G.J., Barrett, N.S., Morton, A.J., 2004. Biases associated with the use of underwater visual census techniques to quantify the density and size-structure of fish populations. J. Exp. Mar. Biol. Ecol. 308 (2), 269–290.

Elith, J., Lehmann, A., Li, J., Lohmann, L.G., Loiselle, B.A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Peterson, A.T., Phillips, S.J., Graham, C.H., Richardson, K., Scachetti-Pereira, R., Schapire, R.E., Soberón, J., Williams, S., Wisz, M.S., Zimmermann, N.E., Araujo, M., Anderson, R.P., Dudík, M., Ferrier, S., Guisan, A., Hijmans, R.J., Huettmann, F., Leathwick, J.R., 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography 29, 129–151.

Elith, J., Leathwick, J.R., Hastie, T., 2008. A working guide to boosted regression trees. J. Anim. Ecol., 802.

Farmer, R.G., Leonard, M.L., Horn, A.G., 2012. Observer effects and avian-call-count survey quality: rare-species biases and overconfidence/effets des observateurs et qualite des inventaires par le denombrement des chants: biais sur les especes rares et exces de confiance. Auk 129, 76.

Ferrier, S., Guisan, A., 2006. Spatial modelling of biodiversity at the community level. J. Appl. Ecol. 43, 393–404.

Fewster, R.M., Buckland, S.T., Siriwardena, G.M., Baillie, S.R., Wilson, J.D., 2000. Analysis of population trends for farmland birds using generalized additive models. Ecology 81, 1970–1984.

Fink, D., Hochachka, W.M., 2012. Using data mining to discover biological patterns in citizen science observations. In: Dickinson, J.L., Bonney, R. (Eds.), Citizen Science: Public Participation in Environmental Research. Comstock Publishing Associates.

Fink, D., Hochachka, W.M., Zuckerberg, B., Winkler, D.W., Shaby, B., Munson, M.A., Hooker, G., Riedewald, M., Sheldon, D., Kelling, S., 2010. Spatiotemporal exploratory models for broad-scale survey data. Ecol. Appl. 20, 2131–2147.

Franklin, J., 2009. Mapping Species Distributions. Cambridge University Press, New York.

Gelfand, A.E., Schmidt, A.M., Wu, S., Silander, J.A., Latimer, A., Rebelo, A.G., 2005. Modelling species diversity through species level hierarchical modelling. J. Roy. Stat. Soc.: Ser. C (Appl. Stat.) 54, 1–20.

Gotelli, N.J., Colwell, R.J., 2001. Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. Ecol. Lett. 4, 379–391.

Hastie, T., Tibshirani, R.J., 1990. Generalized Additive Models. Chapman and Hall, London.

Hedges, L.V., Tipton, E., Johnson, M.C., 2010. Robust variance estimation in meta-regression with dependent effect size estimates. Res. Synth. Meth. 1, 39–65.

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 Ecol. Evol. (Pers. Ed.) 27, 130–137.

Holt, B.G., Rioja-Nieto, R., MacNeil, M.A., Lupton, J., Rahbek, C., 2013. Comparing diversity data collected using a protocol designed for volunteers with results from a professional alternative. Methods Ecol. Evol. 4 (4), 383–392.

Karpievitch, Y.V., Hill, E.G., Leclerc, A.P., Dabney, A.R., Almeida, J.S., 2009. An introspective comparison of random forest-based classifiers for the analysis of cluster-correlated data by way of RF++. PLoS One. 2009 Sep 18, 4:9.

Kery, M., Gardner, B., Monnerat, C., 2010a. Predicting species distributions from checklist data using site-occupancy models. J. Biogeogr. 37, 1851–1862.

Kery, M., Royle, J.A., Schmid, H., Schaub, M., Volet, B., Hafliger, G., Zbinden, N., 2010b. Site-occupancy distribution modeling to correct population-trend estimates derived from opportunistic observations. Conserv. Biol. 24, 1388–1397.

Latimer, A.M., Wu, S., Gelfand, A.E., Silander, J.J.A., 2006. Building statistical models to analyze species distributions. Ecol. Appl. 16, 33–50.

Lawler, J.L., O'Connor, R.R., 2004. How well do consistently monitoried breeding bird survey routes represent the environments of the conterminous United States? Condor 106, 801–814.

Legendre, P., Fortin, M.-J., Gurevitch, J., Hohn, M., Myers, D., 2002. The consequences of spatial structure for the design and analysis of ecological field surveys. Ecography 25, 601–615.

Lunn, D.J., Thomas, T., Best, N., Spiegelhalter, D., 2000. WinBUGS – a Bayesian modeling framework: concepts, structure, and extensibility. Stat. Comput. 10 (4), 325–337.

MacKenzie, D.I., 2006. Occupancy Estimation and Modeling: Inferring Patterns and Dynamics of Species Occurrence. Elsevier, Burlington, MA.

Mateo, R.G., Felicisímo, Á.M., Muñoz, J., 2010. Effects of the number of presences on reliability and stability of MARS species distribution models: the importance of regional niche variation and ecological heterogeneity. J. Veg. Sci. 21, 908–922.

Maunder, M.N., Punt, A.E., 2004. Standardizing catch and effort data: a review of recent approaches. Fish. Res. 70, 141–159.

Miller, D.A., Nichols, J.D., McClintock, B.T., Grant, E.H.C., Bailey, L.L., Weir, L.A., 2011. Improving occupancy estimation when two types of observational error occur: non-detection and species misidentification. Ecology 92 (7), 1422–1428.

Monk, J., 2013. How long should we ignore imperfect detection of species in the marine environment when modelling their distribution? Fish Fisheries, 1467–2979.

Olden, Julian D., Lawler, Joshua J., Poff, N.L., 2008. Machine learning methods without tears: a primer for ecologists. Quart. Rev. Biol, 171.

Parsons, B., Short, J., Roberts, D.D., 2009. Using community observations to predict occurrence of malleefowl (*Leipoa ocellata*) in the Western Australian wheatbelt. Biol. Conserv. 142, 364–374.

Pattengill-Semmens, C.V., Semmens, B.X., 2003. Conservation and management applications of the reef volunteer fish monitoring program. Environ. Monit. Assess. 81, 43–50.

Pearce, J.L., Boyce, M.S., 2006. Modelling distribution and abundance with presence-only data. J. Appl. Ecol. 43 (3), 405–412.

Petchey, O.L., Gaston, K.J., 2006. Functional diversity: back to basics and looking forward. Ecol. Lett. 9, 741–758.

Phillips, S.J., Elith, J., 2013. On estimating probability of presence from use–availability or presence–background data. Ecology 94 (6), 1409–1419.

Phillips, S.J., Anderson, R.P., Schapired, R.E., 2006. Maximum entropy modeling of species geographic distributions. Ecol. Model. 190, 231–259.

Phillips, S.J., Dudık, M., Elith, J., Graham, C.H., Lehmann, A., Leathwick, J., Ferrier, S., 2009. Sample selection bias for presence-only distribution models: implications for background and pseudo-absence data. Ecol. Appl. 19 (1), 181–197.

R Core Team, 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. <http://www.R-project.org/>.

Royle, J.A., Dorazio, R.M., 2008. Hierarchical Modeling and Inference in Ecology: the Analysis of Data from Populations, Metapopulations and Communities, first ed. Academic, Amsterdam, Boston.

Royle, J.A., Kéry, M., Gautier, R., Schmid, H., 2007. Hierarchical spatial models of abundance and occurrence from imperfect survey data. Ecol. Monogr. 77 (3), 465–481.

Royle, J.A., Chandler, R.B., Yackulic, R.B., Nichols, J.D., 2012. Likelihood analysis of species occurrence probability from presence-only data for modelling species distributions. Methods Ecol. Evol. 3 (3), 545–554.

Sarda-Palomera, F., Brotons, L., Villero, D., Sierdsema, H., Newson, S.E., Jiguet, F., 2012. Mapping from heterogeneous biodiversity monitoring data sources. Biodivers. Conserv. 21, 2927–2948.

Segurado, P., Araújo, M.B., Kunin, W.E., 2006. Consequences of spatial autocorrelation for niche-based models. J. Appl. Ecol. 43, 433–444.

Sela, S.J., Simnoff, J.S., 2012. RE-EM trees: a data mining approach for longitudinal and clustered data. Mach. Learn. 86, 169–207.

Seys, J., Offringa, H., Van Waeyenberge, J., Meire, P., Kuijken, E., 2002. An evaluation of beached bird monitoring approaches. Mar. Pollut. Bull. 44, 322–333.

Silvertown, J., 2009. A new dawn for citizen science. Trends Ecol. Evol. (Pers. Ed.) 24, 467–471.

Snäll, T., Kindvall, O., Nilsson, J., Pärt, T., 2011. Evaluating citizen-based presence data for bird monitoring. Biol. Conserv. 144, 804–810.

Stuart-Smith, R.D., Bates, A.E., Lefcheck, J.S., Duffy, J.E., Baker, S.C., Thomson, R.J., Stuart-Smith, J.F., Hill, N.A., Kininmonth, S.J., Airoldi, L., Becerro, M.A., Campbell, S.J., Dawson, T.P., Navarrete, S.A., Soler, G., Strain, E.M.A., Willis, T.J., Edgar, G.J., in press. Integrating abundance and functional traits reveals new global hotspots of fish diversity. Nature.

Sullivan, B.L., Wood, C.L., Iliff, M.J., Bonney, R.E., Fink, D., Kelling, S., 2009. EBird: a citizen-based bird observation network in the biological sciences. Biol. Conserv. 142, 2282–2292.

Sunde, P., Jesson, L., 2013. It counts who counts: an experimental evaluation of the importance of observer effects on spotlight count estimates. Eur. J. Wildl. Res. 1612–4642, 1–9.

Tulloch, A.I.T., Szabo, J.T., 2012. A behavioral ecology approach to understand volunteer surveying for citizen science datasets. Emu 112, 313–335.

Tulloch, A.I.T., Possingham, H.P., Joseph, L.N., Szabo, J., Martin, T.G., 2013. Realizing the full potential of citizen science monitoring schemes. Biol. Conserv. 165, 128–138.

Walker, P.A., Cocks, K.D., 1991. HABITAT: a procedure for modelling a disjoint environmental envelope for a plant or animal species. Global Ecol. Biogeogr. Lett. 1, 108–118.

Welham, S., Cullis, B., Gogel, B., Gilmour, A., Thompson, R., 2004. Prediction in linear models. Aust. NZ J. Stat. 46 (3), 325–347.

Wiggins, A., Crowston, K., 2011. From conservation to crowd sourcing: a typology of citizen science. In: System Sciences (HICSS), 2011 44th Hawaii International Conference on, pp. 1–10.

Wiggins, A., Newman, G., Stevenson, R.D., Crowston, K., 2011. Mechanisms for data quality and validation in citizen science. In: e-Science Workshops (eScienceW), 2011 IEEE Seventh International Conference on, pp. 14–19.

Wikle, C.K., 2003. Hierarchical Bayesian models for predicting the spread of ecological processes. Ecology 84 (6), 1382–1394.

Yackulic, C.B., Chandler, R., Zipkin, E.F., Royle, J.A., Nichols, J.D., Campbell Grant, E.H., Veran, S., 2013. Presence-only modelling using MAXENT: when can we trust the inferences? Methods Ecol. Evol..

Zuur, A.F., 2009. Mixed Effects Models and Extensions in Ecology with R. Springer New York, New York.

Zuur, A.F., Smith, G.M., Ieno, E.N., 2007. Analysing Ecological Data. Springer, New York.

Zuur, A.F., Elena, N.I., Elphick, C.S., 2010. A protocol for data exploration to avoid zcommon statistical problems. Methods Ecol. Evol. 1 (1), 3–14.