Statistical solutions for error and bias in global citizen science datasets

Tomas J. Bird^{1,2,*}, Amanda E. Bates¹, Jonathan Lefcheck³, Nicole Hill¹, Simon
Wotherspoon¹, Martin Krkosek⁴, Rick Stuart-Smith¹, Jemina Stuart-Smith¹,
Gretta Pecl¹, Graham Edgar¹, Russell J. Thomson¹, Neville Barrett¹, Stewart
Frusher¹

¹Institute for Marine and Antarctic Studies, University of Tasmania, Hobart,
Tasmania, 7001 Australia

²School of Botany, University of Melbourne, Parkville, Victoria, 3010 Australia

³Virginia Institute of Marine Science, College of William & Mary, PO Box 1346,
Gloucester Point, Virginia, 23062-1346, USA

⁴Department of Ecology and Evolutionary Biology, Ramsay Wright Zoological
Laboratories, University of Toronto, 25 Harbord St, Toronto, ON, M5S 3G5,
Canada

*Corresponding author: tomasbird@gmail.com

194 Albion Heights Dr, Kingston, Tasmania, 7050, Australia

phone: +61 4 0949 6340

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 now possible to (1) describe how the sampling process 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 further be incorporated into predictions of species distributions and estimates of biodiversity. By detailing how CS data are generated, patterns can be discerned from complex datasets that are unevenly distributed and collected by many observers with varying skill levels. Even so, gaining the full potential from even the best designed CS projects requires meta-data describing the sampling process, reference data to allow for standardization, and insightful modeling suitable to the type of response data of interest.

1. Introduction

Evaluating the conservation implications of human activity on 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', have been central to the collection of broad-scale databases, allowing the scientific community to address questions that would otherwise be logistically or financially infeasible for even the most dedicated scientific team (Dickinson et al., 2010). Consequently, volunteer networks are a practical way to answer conservation-related questions on the broad temporal and spatial scales that are relevant to humanity. 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 or taxa (Bonney et al., 2009).

As well as providing a cost-effective means of addressing large-scale questions in ecology (Gardiner 2012), involving citizens in the collection of data has a number of benefits to conservation-related projects. Being inclusive and engage large numbers of people, CS projects can promote publicity and discourse on conservation issues (Trumbull, 2001), as well as providing people with the opportunity to take an active role in fostering positive environmental change (Crall et al 2012). As well, CS projects can often afford to be more exploratory than more regimented monitoring programs. Consequently CS projects such as Jellywatch and RedMap (www.redmap.org) can include efforts to monitor rare events cannot be

reliably observed through regular sampling. Given these advantages, the capacity to address global-scale conservation may well rest in the realm of citizen science (Silvertown, 2009).

In spite of the proven success and potential for applications of citizen-collected data to 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. As well, significant sources of bias such as under-detection of species or the non-random distribution of effort may be present in the data (Crall et al., 2010). Such concerns have motivated CS projects to maximize the quality of data collected through improved sampling protocols (Edgar and Stuart-Smith, 2009), database management (Crall et al., 2010), 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, most global CS data sets will likely violate many of the assumptions required in simple statistical analyses.

Fortunately, the issues of error and bias that are often present in CS data are not unique or insurmountable; analogous problems exist in datasets across a wide variety of disciplines and are successfully addressed using a suite of analytical approaches. In many cases, CS databases resemble the data collected for meta-analytical and landscape ecology studies where methods

for accurately estimating and incorporating within-study or within-observer variability are a key feature of inference (Hedges et al., 2010). For complex datasets, machine learning approaches are available that can accurately examine the relative importance of large numbers of predictive variables in explaining the response data (Olden et al., 2008) (FINK AND HOCHACHKA 2012). 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. To meet this objective, we split the manuscript into three parts. Using examples from the literature, we first describe the main data quality issues present in CS data and how these issues can be quantified using meta-data that describes key aspects of the sampling process. Second, based on a broad overview of literature in statistical ecology we describe a number of modeling approaches available for use with CS data. Third, we consider how to address specific issues of error and bias using appropriate statistical approaches and meta-data. 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, and many implement volunteer training, data standardization, validation and filtering procedures to remove potential sources of error and bias before, during or after the data are collected (Wiggins et al. 2011b, Bonter and Cooper 2012). Nevertheless, in spite of excellent data quality procedures three primary quality issues may still remain. First, CS data may have little information about the scale or magnitude of the effect observed. Second, CS data may still be prone to greater variability, or error, due to differences in the skills, dedication, and training of volunteer participants. Third, CS data may contain persistent bias. To address these quality issues, it is necessary to carefully consider the type of response data 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 kind of data to collect, as this choice constrains the types of change in communities or species that can be observed over space and time. Inherent in the selection of the type and scope of data collected is a tradeoff between the amount of information contained in each data point and the effort required to collect it. 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, presence-absence, or some measure of quantity (such as abundance, percent cover or biomass),

all recorded over time and space. The relative utility of different data types for detecting change relates to the amount of information each contains about the scale of the observed response.

While the easiest type of data to collect, presence-only data raise difficulties in subsequent analyses, as it can be difficult to discern the relative importance of individual data points without some idea of the scale each point represents (Pearce and Boyce 2006). For example, a presence-only observation may be the result of a brief survey effort or the outcome of a years-long search. Without some information about the sampling effort represented by each data point, both have the same functional relevance. 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. Similarly, abundance (or other measures of quantity) data are required to detect changes in the true size of a population, which are not reflected in presence-absence data.

The collection of more complex data places greater demands on volunteers (Table 1). For example, presence-only data, such as georeferenced reports of species or phenomena (Sullivan et al., 2009) are relatively easy to collect and submit. By contrast, in the case of presence-absence data, volunteers must be able to accurately quantify their search effort in time or space, on top of accurately reporting a species or event. Quantifying search effort can range from searching for a defined time in one location to searching within transects or quadrats (Edgar and Stuart-Smith 2009), but must remain

consistent between samples. With count or percent cover data, there is the additional assumption that volunteers can correctly and exhaustively count or measure a species over the entire sampling area. Depending on the circumstances of the study (such as location, timing, and the species of interest), estimating quantity can require extensive training to execute properly (Edgar and Stuart-Smith 2009).

The increased focus on training when collecting more complicated datasets means greater effort is required to gain similar coverage as could be achieved using a simpler sampling scheme; in general, it is easier to recruit, train and retain volunteers to perform tasks requiring less effort. In addition, increasing the complexity of tasks required by volunteers leads to greater opportunity for observation error. This tradeoff has two implications for survey design. First, researchers considering the scope of inference for their study should be aware of how the data type they collect will influence the kinds of questions they can ask, what statistical tools will be available to do so, and what additional information should be collected with each response data point for analyses (Wiggins et al., 2011, Wiggins and Crowston, 2011).

2.2. Random Error in Citizen Science Datasets

Where error in data influences the variability of the response, but not the mean, it is referred to as random error. In the context of CS data, random 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. Left unexplained, random error may obscure detection of process-related trends in the response data.

Accounting for sources of random error requires meta-data to provide additional context for each data point, helping to distinguish variability due to the sampling process from the process of interest. 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 in each survey are also useful for standardizing count data by the amount of time searched (Bray and Schram, 2001, Maunder and Punt, 2004). Additionally, environmental conditions can have significant impacts on the success of sampling, particularly for inexperienced volunteers. For instance, in SCUBA surveys, maintaining buoyancy and focus on the task at hand in conditions of high swell is challenging for even expert divers (Edgar and Stuart-Smith, 2009). Thus, recording meta-data on any factors that are likely to influence the effectiveness of surveys can underpin successful analyses.

2.3. Bias

Bias refers to the tendency of some measurements to provide inaccurate estimates of the process being measured. Here we define two kinds: systematic bias and sampling bias.

226

227 Systematic bias occurs when repeated measures of the same process are
228 likely to provide systematic over or under-estimates of the true value. False
229 absences in presence-absence data (i.e., incorrectly stating that a species
230 was absent in a sample when it was actually present) and species
231 misidentification are an example of bias particularly common with CS data,
232 due to the skill and expertise required to observe and identify cryptic species
233 (Royle 2007). Such errors can lead to under-estimates of the area occupied
234 by a species (Farmer et al., 2012). Another example occurs in underwater
235 dive surveys in which individuals are asked to estimate the size of objects by
236 sight alone, as the physical properties of light passing through the air/water of
237 a diver's mask results in viewed objects appearing 25% larger (Jeffreys,
238 2010). If measurement bias is consistent across all aspects of a study, then it
239 can be relatively easy to deal with using corrective formulae. However, when
240 measurement bias varies unpredictably between samples, it can be
241 impossible to extricate from the process of interest. Accounting for
242 measurement bias in CS data requires a way to describe how this difference
243 between a measured response and its true value varies across samples. This
244 usually requires either a sampling design that can account for this bias,
245 calibration samples, or otherwise into the data collection design (see Section
246 3.3; Hierarchical models).

247

248 Sampling bias occurs when some aspects of the underlying distribution are
249 more likely to be sampled than others. Thus, aspects of the underlying
250 distribution that are overrepresented in sampling will dominate apparent

trends. One common source of bias is observer-level variability in sampling. On average, the mean value of a measurement made by observers may be centered on the true value – the “Wisdom of the Crowd” effect is a good example of this (Arlington and Krause, 2013). However, if some observers are more prolific in their observations and their observations are systematically over-or underestimates, then their observations will bias the overall estimate. Commonly, uneven 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 (Boakes et al., 2010). For example bird surveys are often clumped near areas that are more accessible, which may in turn be associated with habitats preferable to certain species (Lawler and O’Connor, 2004; Tulloch et al., 2012). Volunteer effort may also be temporally restricted due to seasonal windows or variability in volunteer availability making it difficult to distinguish seasonal patterns from those created by overall volunteer effort (Seys et al. 2002; Ahrends, 2012).

3. Modeling approaches

The development of modern statistical tools presents options for accounting for many types of error and biases. Here, we describe a variety of such techniques that may be particularly relevant to citizen science data. Supplementary Table 2 provides examples of freely available statistical packages for the open-source program R (R Development Core Team 2012).

3.1. *Linear and Generalized Linear Models*

Linear models describe the relationship between the response data and a series of predictors (such as covariates or meta-data) as an additive combination of linear relationships (Zuur, 2007). For many linear models, the response data are assumed to be normally distributed, and could include factors such as site-level biodiversity or species range sizes. In linear models, the predictor variables are assumed to directly influence the response data and are called fixed-effects. The strengths of relationships between the fixed effects and response the response data are estimated by model parameters. Additive models extend linear models by allowing non-linear relationships between predictors and response data through the use of smoothing functions with multiple parameters (Hastie and Tibshirani, 1990). Note that an additive model with only one parameter for each predictor is a linear model.

Basic linear models assume that the response data follow a normal distribution, which may render them unsuitable for many kinds of CS data as counts of species, percentages or data with a large proportion of zeros often do not fit a standard normal distribution. Generalized linear and additive models (GLMs and GAMs), further extend linear and additive models to allow for other kinds of distributions. Both GLMs and GAMs can be based on the Poisson or negative binomial distribution for count data, or the logistic regression for binary data (Zuur, 2007). Additionally, zero-inflated linear of additive models can be useful for analyzing CS data, which can contain many zeros (Zuur 2007, Liu and Chan 2010). However, linear and additive models

are generally not suitable for presence-only data, unless pseudo-absences are generated, an important consideration for the context of citizen-generated data.

3.2. Mixed-effects models

Mixed-effects models include fixed effects with ‘random-effects’ that estimate the influence of predictors (often groups) that increase variability in the data but do not affect the mean response. As an example, one might expect observers to have differing sampling efficiencies, so we could include observer identifier as a random effect to account for this source of variation before estimating the effects of other predictors in the model (Zuur, 2009). In this case, the mean influence (positive or negative) of each observer on the overall mean is estimated and given a parameter. Random-effects can be included in linear or additive models, resulting in Generalized linear (or additive) mixed models, which are now widely applied in ecology (Bolker et al., 2009).

3.3. Hierarchical models

Hierarchical models allow for the parameters describing relationships between predictors and response data to themselves be described by linear (or other) models (Royle and Dorazio, 2008). In this way, mixed-effects models are hierarchical models where the effect of all observers on the mean response is normally distributed around zero. Hierarchical models are frequently used in ecology to separately describe observation process (i.e. how the response data were generated) and the underlying process of interest. For example,

when surveying for cryptic species, the number of individuals observed is often an under-estimate of the true number that were present (Royle, 2006). A hierarchical model for this situation would assume that the observations were the outcome of a random process that depended on the probability of an observer sighting the species and whether or not the species was present in the site. The true presence (given the number of samples that were likely miscounted) would then be modeled as a function of environmental covariates (MacKenzie, 2006). Hierarchical models are increasingly popular in ecological analyses, and books covering the topic have been published recently, from frequentist (Clark, 2007) and Bayesian perspectives (Royle and Dorazio, 2008).

3.4. Machine Learning

Machine learning approaches use computer algorithms to learn about the most likely relationship between predictors and response data on the basis of simple rules about how the response data should be divided in relation to the predictors (Olden, 2008). These approaches thus provide an indicator of how important each predictor is in determining the response data. Unlike linear and additive modeling approaches, machine learning tools do not assume a particular distribution for the response data, and applications for presence-only, presence-absence, abundance and other data types are available (Elith et al., 2008). As well, many machine learning approaches do not assume that the relationships between responses and predictors are linear (or even smooth). Finally, with machine learning it is possible to estimate the influence of large numbers of predictor variables, which can sometimes make it easier

to unravel relationships in complex ecological data. Approaches that have been applied in an ecological setting include classification and regression trees (CART, De'ath, 2000), boosted regression trees (BRT, Elith et al., 2008), random forests (Cutler et al., 2007), artificial neural networks, and genetic algorithms (Olden et al., 2008).

3.5. Estimating Biodiversity

Biodiversity indices provide a means of describing the relative amount of species diversity within ecological communities. Numerous indices are available ranging from species richness, (the number of species in a site), to more complicated indices incorporating information on species' relative abundances (e.g., Shannon or Simpson), functional traits or phylogenetic relationships (e.g., Swenson., 2011). Depending on the data availability and the study design, biodiversity indices can be treated as either response data (e.g. Ahrends et al., 2011) or as a parameter that is estimated as the outcome of a hierarchical model (Kery et al. 2010a).

3.6. Species Distribution Models

Species distribution modeling (SDM), also referred to as environmental niche modeling or climate envelope modeling, quantifies relationships between environmental factors and response data such as species occurrence, abundance or biodiversity. These relationships can 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). SDM's use a broad range of modeling techniques, including many of the parametric and

machine learning methods discussed above. SDM models are also available to deal with presence-only data through envelope models: Carpenter et al. 1993; Gillsion et al. 1993; Stockwell and Peters, 1999), entropy modeling (MAXENT, Phillips et al. 2006) or maximum likelihood (MAXLIKE, Royle et al, 2012). More recently, mixed models that can model the response of groups of species simultaneously have been developed (Dunstan et al. 2011) (Table 2). Large and broad-scale datasets such as those collected by citizen science programs are attractive for SDM as they can be compared against extensive geographical data sets using GIS. As a consequence SDM is gaining popularity in conservation ecology and policy (Ashcroft et al. 2012; Pattengill-Semmens and Semmens 2003; Sarda-Palomera et al. 2012).

4. Accounting for error and bias

In this section, we explore how some of the previously mentioned sources of error and bias can be explicitly accounted for using the modeling approaches described in section 3. We also provide several examples of analyses, using data drawn from the global Reeflife Survey (RLS) database (Edgar and Stuart-Smith, 2009). A description of RLS and RLS survey methodologies are found in the SOM. Finally, supplementary table 2 provides examples of studies that have used some of the methods we describe in the context of CS data.

4.1. Random Error

To check for sampling-related error in CS data sets, the simplest approach is to first test for relationships between response data and meta-data using

visual plots. The significance of relationships can be tested using generalized linear or additive models. For example, a study on the detectability of wild animals in spotlight surveys used GLMs to model how age and hunting experience accounted for up to a 45% increase in observer effectiveness (Sunde and Jesson, 2012). Similarly many CS project engage in a data 'validation' exercise aimed at comparing data generated by volunteers and experts (e.g. Delaney 2008, Edgar and Stuart-Smith 2009), with the comparisons being made with the aid of GLMs. However, because there can be many competing factors influencing response data, there is a risk of building models with more parameters than is possible to estimate with the data. Thus an important part of inference using linear or additive models is the process of determining which model provides the best fit with as few parameters as possible (Zuur, 2007).

In cases where there are many predictor variables, or where the response data do not follow standard probability distributions, machine learning approaches can be particularly useful (De'ath, 2000).

RUSSELL EXAMPLE HERE

4.2. Bias

4.2.1 Sampling bias

Treatment of bias due to uneven sampling effort can be challenging, and may require exploratory modeling of both the observations and underlying process (Fink et al. 2010). As well, description of error structure requires large amounts of data in order to adequately attribute sources of bias to random factors. Sampling bias that influences the variability of observations but not the mean can be easily accommodated within GLMMs or GAMMs (Bolker et al. 2009). Observer or site-related random effects, as well as interactions with various sampling conditions can often account for a large amount of variability in CS data. Alternatively, if the relationships between the meta-data and response data are numerous, a mixed-effects regression tree (which allows for hierarchical clustering of the response data) may be feasible (Sela and Simnoff, 2012) though this approach has yet to be implemented in statistical packages.

Accounting for bias due to spatial clustering of sampling is one of the trickier aspects of dealing with CS data, and indeed any large-scale database. Part of the problem lies in the fact that samples that lie closer together in space or time may be more similar. Thus, the first challenge is to recognize whether the underlying ecological data are autocorrelated in space or time (Legendre et al., 2002). Given the importance of spatial and temporal processes in ecology, an increasing body of literature is available describing how to deal with the spatial and temporal distribution of data (Cressie and Wikle, 2011, Zhang 2012). However, sampling bias in spatially or temporally autocorrelated data can often arise due to clustering of sampling relative to an underlying spatial or temporal trend in the data. If enough data are available,

subsampling from the data may be sufficient to examine unbiased trends in the underlying data, at the cost of a large proportion of the (Legendre et al., 2002). An alternative within Linear and additive mixed-effects models is to use variance weighting to reduce the importance of undersampled areas (Crawley, 2012). However, this approach has the disadvantage of being sensitive to the weightings used, which can be difficult to estimate. Hierarchical Bayesian approaches to spatial and temporal modeling have also been used to explicitly model sampling in relation to a latent underlying process (Wikle, 2003). However, such models are not generally available in a ready-to-use format, and require significant effort in terms of parameterization and mathematical understanding.

We briefly note that many machine learning techniques do not allow for modeling spatial auto-correlation. However, a 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). The use of data-mining approaches to the exploration of spatial data is also an area of development in large-scale CS databases that will likely provide novel analyses (Fink and Hochachka, 2012)

4.2.2. Example: Spatial clustering in species richness data

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 dataset in which a strong spatial trend is confounded

475 with uneven sampling. A fundamental hypothesis in ecology is that the
476 diversity of species decreases towards the poles. In our example we plot
477 species richness data of fish against Latitude from a subset of the Reeflife
478 Survey dataset. In figure 2A it is immediately apparent that the bulk of the
479 data available are from lower latitudes. Ignoring this fact, a polynomial fit to
480 the data (using the `gls` function from the `nlme` package in R) seems to provide
481 an adequate fit ($AIC = 3472$), given the narrow confidence intervals around
482 the model prediction. However, this narrow interval is largely an artifact of the
483 large sample size, which is mostly clustered around Hawaii. Examination of
484 the residuals shows a large discrepancy between the variance in different
485 regions, violating the assumption of equal variance required for such linear
486 models (figure 2D). Including a random effect at the SITE? Level gives a
487 better fit ($AIC = 3470$), broader confidence intervals (figure 2B) and helps to
488 center the model predictions (figure 2E), but we are still faced with uneven
489 variance between the two regions of the data. Finally by using a model that
490 estimates the differing levels of variation between the four dominant regions of
491 the data (Temperate Northern Pacific, Eastern Indo-Pacific, Temperate
492 Northern Atlantic and Tropical Eastern Pacific), we arrive at a better-fitting
493 model ($AIC = 3381$) that does away with the polynomial relationship between
494 Latitude and richness (figure 2C) and that properly reflects the amount of
495 variability in each region (figure 2E). Thus accounting for site-level bias and
496 regional differences in sampling variability can improve model fit to the data.
497 However, it should be noted that these approaches assume equal sampling
498 effectiveness across all sites and thus do not account for any spatial variability

in systematic bias. To do so, additional data on the observation process are required.

4.3.1. Systematic Bias

Where there is systematic bias in a dataset, accounting for this error can be as simple as measuring observer data against some known quantity. For example, dealing with consistent over-estimates of size while underwater can be accomplished by calibrating diver's observations against objects of known length (Jeffreys, 2007). However, it is often not the case that a simple correction will account for systematic bias. In this case, more explicit modeling of the data generation process may be required, including modifications of the survey design to estimate bias-generating mechanisms. A number of hierarchical approaches have been published for handling systematic biases common to CS data, including false absences (Farmer 2012) and false presences (i.e., mistakenly recording a species as present when it was in fact absent, Royle and Link, 2006), and species misidentifications (Conn et al, 2013). As in linear or additive models, random effects within the observation model component (such as the rate of false-absence reporting) can be used to address the non-independence of observations within sampling groups such as individual observers (MacKenzie, 2006). Similarly, consistent over- or under-estimates of size and abundance can be accommodated within a hierarchical framework if suitable reference or calibration data are available.

Presence-only data can be thought of as a special case of sampling bias and several techniques have been developed to enhance the value of such data. The first is to change the scope of the study to defining the limits of a species' range Programs such as BIOCLIM (Busby 1986, 1991) and HABITAT (Walker & Cocks 1991) have been designed to help describe the 'Environmental Envelope' of species, which can then be compared against other species. Unfortunately, such envelope methods usually provide an under-estimate of a species' true distribution or niche, especially with rare or cryptic species. Alternatively, Ward et al. (2009) have developed a novel boosted regression tree algorithm to estimate prevalence from presence-only data, though this approach is still under development. If the habitat surrounding a presence-only observation has been detailed (such as from aerial or satellite imagery), it is possible to randomly generate pseudo-absence data and use this information in standard linear models, usually in the context of a SDM framework (Barbet-Massin et al. 2012; Engler et al. 2004, section 4.5).

4.3.2. Example: estimating detection error in transect survey data

Here, we show how accounting for imperfect detection in sampling can result in drastic underestimates of species occurrence, and can be accounted for using replicate sampling. Most species are adapted to live within a certain thermal niche, which influences where on the planet they can live. Here, we look at how the presence or absence of the urchin *Amblypneustes ovum* relates to temperature on the east coast of Australia. A simple logistic regression (Figure 3, red line) assumes that the vector of observed site

occupancy data Y are Bernoulli distributed random variables with $\Pr(y_i=1)=U_i$ and U_i is the probability of occupancy at site i . Further, to allow site-level dependence of occupancy on temperature, U_i has the form $\text{logit}(U_i) = B_0 + BT \times SST_i$, where B_0 and BT are the intercept and slope parameters of the regression and the logit function is $\text{logit}(X) = \frac{\exp(X)}{1+\exp(X)}$. Thus, we see that the probability of finding *A. ovum* is highest at lower temperatures, but that it does not appear occupy more than ~30% of sites.

By contrast, the hierarchical model uses the fact that sampling was replicated twice within a site to separately model 1) the chance of a site being occupied given temperature and 2) the number of times the species was detected given whether the site was occupied and the number of replicates. The vector N , the number of times that the species was observed at each site is a binomial random variable of size 2 and probability $U_i \times D$. Here, D is the probability of detection, which in this simple model is assumed to be constant across all sites. Thus we see that the temperature-related occupancy rate of *A. ovum* is almost double that estimated by the simple logistic regression. While this analysis does account for important underlying aspects of the distribution of this species, it does demonstrate the importance of accounting for imperfect detection in survey data.

4.4 Quantifying biodiversity given imperfect sampling

Because species vary in their susceptibility to being observed, biodiversity indices can be heavily influenced by variation in sampling and detectability. How error and bias are accounted for within the context of biodiversity

depends on whether they are treated as a type of response data, or as a parameter derived after correcting for error and bias in the data for the component species. Alternatively the modeling approaches described above can be used to deal with error and bias at the species level and biodiversity calculated after the fact. If biodiversity is the response data of interest, then error and bias correction mechanisms can be used in the calculation of the index data themselves.

Very simply, the choice of index used may also help minimize issues of detectability, and 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 in calculating overall community diversity. Alternatively, one can model the detectability of each species and incorporate this information into the calculation of diversity (Gelfand et al., 2005), which requires a detection-occupancy or similar design to estimate detectability.

Rarefaction is often used on biodiversity data to account for uneven sampling effort. Traditional rarefaction generates species accumulation curves, then reduces the largest samples until they are equivalent in size to the smallest (Gotelli and Colwell, 2011). Newer methods referred to as ‘fixed-coverage’ or ‘shareholder quorum subsampling’ extrapolate richness from the same curves, then scales back to the lowest level of sample ‘coverage’, a measure of sample completeness (Alroy, 2009; Chao and Jost, 2012). The latter methods are less biased, use all available data, and minimize the amount of sampling effort.

598

599 *4.4.2 Example: Rarefaction of species data to highlight sampling coverage*

600 Comparing the diversity of communities can be challenging when the
601 sampling conditions vary markedly between sites. In particular, the presence
602 of poorly-observed species in samples can bias estimates of diversity. Thus,
603 one important way remove the impact of rare species on site-to site
604 comparisons of diversity is to use rarefaction to show how each individual in
605 the sample contributes to the overall sample richness. In figure 4A, the Goat
606 Island site (red triangle) has much higher species richness and abundance
607 than the Waiwiri site (black circle). However, the rarefaction curves show that
608 the two sites arrive at their respective richness levels at different rates, with
609 the Waiwiri site accumulating richness at a higher rate per individual.
610 Additionally, extrapolating to the theoretical impact on richness of further
611 sampling shows that further sampling at the Goat Island site would add
612 richness, but not so at Waiwiri (figure 4B). Examining rarefaction curves like
613 this can therefore provide a measure of the relative effectiveness of sampling
614 across different areas, or between observers. While such metrics do not
615 provide predictive power as in other modeling approaches, they do give a
616 useful tool for comparison.

617

618 *4.5 Predicting species distributions with imperfect data.*

619 Given that most SDMs use linear, additive or machine-learning models to
620 make predictions into unsampled space it is possible to address random error
621 appropriate for each method using meta-data and covariates where possible.
622 However, this approach may be limited for use in predictive SDMs because

the sampling-related may not be defined in the space for which predictions are being made. Therefore, random error of the kind encountered in CS data is often reduced as much as possible by screening the data before analysis. However, detection-occupancy modeling has been used successfully within SDMs (Kery et al., 2010) 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.

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, 2005; Latimer, 2006).

Highly clustered presence-only data, which seem to represent the worst of all possible scenarios, have received recent attention in SDMs. Often ‘pseudo-absences’ are generated to enable modeling with parametric and machine learning methods, or ‘background’ samples are defined as part of the

modeling procedure (e.g. MAXENT; Phillips et al. 2006). However, presence-only methods are particularly sensitive to sampling bias. Recent work suggests that generating pseudo-absence data that are spatiotemporally biased in the same way as the observation data significantly improves the performance of predictive models (Phillips 2009).

5. Recommendations

Taking an overview of recent techniques used in research based on citizen-science data (Table 3), we have extracted a few basic recommendations:

- 1) *Understand the tradeoff between volunteer effort, data quality and data quantity:* Researchers must clearly understand the scope of sampling they wish to undertake and how to recruit, train and maintain volunteers across the areas of interest.
- 2) *Record survey information:* Meta-data will help define sources of and account for random error and bias in the data.
- 3) *Plot the data to identify potential data issues:* Understanding what kinds of quality issues are likely present will help in handling them.
- 4) *Network with statisticians:* Considering analytical approaches prior to the start of sampling will help decide what information are needed from surveys.

To call in the statistician after the experiment is done may be no more than asking [them] to perform a post-mortem examination: he may be able to say what the experiment died of. ~ Sir Ronald Aylmer Fisher

672 5) *Keep it simple*: While it may be tempting to build increasingly complex
673 models, practitioners often favor simpler, more robust approaches
674 more likely to give reliable outcomes.

675 In closing, the recent growth of citizen-science datasets and proliferation of
676 exciting statistical tools have combined to create unprecedented opportunities
677 to understanding global-scale patterns and changes in species distributions
678 and biodiversity.

682 **Acknowledgments**

683 T.J.B. was supported by an Australian Research Council Linkage doctoral
684 fellowship. Salary for A.E.B was by the Australian National Network in Marine
685 Science, a collaborative funding initiative between James Cook University,
686 The University of Tasmania, and The University of Western Australia.

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. *Diversity and Distributions* 17, 191-200.
- Alroy, J., 2009. A deconstruction of Sepkoski's Phanerozoic marine evolutionary faunas based on new diversity estimates. Page 507 in Geological Society of America abstracts with programs. GSA annual meeting, Portland, Oregon, USA. Geological Society of America, Boulder, Colorado, USA.
- 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. *Ecological Indicators* 13, 243-252.
- 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. *Biological 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 in ecology and evolution*, 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

712 and temporal bias in species occurrence data. PLoS biology 8,
713 e1000385.

714 Bolker, B., M., Brooks, M., E. , Clarkm Connie, J., Geange, S., W. , Poulsen,
715 J., R. , Stevens, H.H., White, J.-S., S. , 2009. Review: Generalized
716 linear mixed models: a practical guide for ecology and evolution.
717 Trends in Ecology & Evolution 24, 127-135.

718 Bonney, R., Cooper, C.B., Dickinson, J., Kelling, S., Phillips, T., Rosenberg,
719 K.V., Shirk, J., 2009. Citizen Science: A Developing Tool for Expanding
720 Science Knowledge and Scientific Literacy. BioScience 59, 977-984.

721 Bonter, N. D. and Cooper, C. B., 2012. Data validation in citizen science: a
722 case study from Project FeederWatch. Frontiers in Ecology and the
723 Environment, 10:6, 305–307.

724 Bray, B.S. and Schramm, J.L. Jr., 2001. Evaluation of a statewide volunteer
725 angler diary program for use as a fishery assessment tool. North
726 American Journal of Fisheries Management, 21:3, 606-615.

727 Brunsten, C., and Comber, L., 2012. Assessing the changing flowering date of
728 the common lilac in North America: a random coefficient model
729 approach. Geoinformatica, 16, 675-690.

730 Chao, A., Jost, L., 2012. Coverage-based rarefaction and extrapolation:
731 standardizing samples by completeness rather than size. Ecology 93,
732 2533-2547.

733 Clark, J.S., 2007. Models for ecological data: an introduction. Princeton
734 University Press, Oxford.

735 Crall, A., Newman, G., Jarnevich, C., Stohlgren, T., Waller, D., Graham, J.,
736 2010. Improving and integrating data on invasive species collected by
737 citizen scientists. *Biological invasions* 12, 3419-3428.

738 Crawley, M.J., 2012. *The R book*. Wiley

739 Cressie, N., & Wikle, C.K., 2011. *Statistics for spatio-temporal data*. Wiley.

740 Conn, P., McClintock, B.T., Cameron, M., Johnson, D.S., Moreland, E. and
741 Boveng, P.L., 2012. Accomodating species identification errors in
742 transect surveys. *Ecology*, 2013 (online).

743 Cox, T.E., Philippoff, J., Baumgartner, E., Smith, C.M., 2012. Expert variability
744 provides perspectives on the strengths and weaknesses of citizen-
745 driven intertidal monitoring program. *Ecological Applications* 22:4,
746 1201-1212.

747 Cutler, A., Cutler, D.R., Edwards, T.C., Beard, K.H., Hess, K.T., Gibson, J.,
748 Lawler, J.J., 2007. Random Forests for Classification in Ecology.
749 *Ecology* 88, 2783-2792.

750 De'ath, G., Fabricius, K.E., 2000. Classification and regression trees: A
751 powerful yet simple technique for ecological data analysis. *Ecology* 81,
752 3178-3192.

753 deSolla, S.R., Shirose, L.J., Fernie, K.J., Barrett, G.C., Brousseau, C.S.,
754 Bishop, C.A., 2005. Effect of sampling effort and species detectability
755 on volunteer based anuran monitoring programs. *Biological*
756 *Conservation* 121, 585-594.

757 Delaney, D., Sperling, C., Adams, C., Leung, B., 2008. Marine invasive
758 species: validation of citizen science and implications for national
759 monitoring networks. *Biological Invasions* 10, 117-128.

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

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

768 Dudík, M., Phillips, S.J., Schapire, R.E., 2005. Correcting sample selection
 769 bias in maximum entropy density estimation., In *Advances in neural*
 770 *information processing systems*. pp. 323-330. MIT Press, Cambridge,
 771 Massachusetts, USA.

772 Dunstan, P.K., Foster, S.D., Darnell, R., 2011. Model based grouping of
 773 species across environmental gradients. *Ecological Modelling* 222,
 774 955-963.

775 Edgar, G.J., Stuart-Smith, R.D., 2009. Ecological effects of marine protected
 776 areas on rocky reef communities;a continental-scale analysis. *Marine*
 777 *Ecology Progress Series* 388, 51-62.

778 Elith, J., Leathwick, J.R., Hastie, T., 2008. A Working Guide to Boosted
 779 Regression Trees. *Journal of Animal Ecology*, 802.

780 Elith, J., Lehmann, A., Li, J., Lohmann, L.G., Loiselle, B.A., Manion, G.,
 781 Moritz, C., Nakamura, M., Nakazawa, Y., Peterson, A.T., Phillips, S.J.,
 782 Graham, C.H., Richardson, K., Scachetti-Pereira, R., Schapire, R.E.,
 783 Soberón, J., Williams, S., Wisz, M.S., Zimmermann, N.E., Araujo, M.,
 784 Anderson, R.P., Dudík, M., Ferrier, S., Guisan, A., Hijmans, R.J.,

785 Huettmann, F., Leathwick, J.R., 2006. Novel Methods Improve
 786 Prediction of Species' Distributions from Occurrence Data. *Ecography*
 787 29, 129-151.

788 Engler, R., Guisan, A., Rechsteiner, L., 2004. An improved approach for
 789 predicting the distribution of rare and endangered species from
 790 occurrence and pseudo-absence data. *Journal of Applied Ecology*, 41,
 791 263-274.

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

797 Ferrier, S., Guisan, A., 2006. Spatial modelling of biodiversity at the
 798 community level. *Journal of Applied Ecology* 43, 393-404.

799 Fewster, R.M., Buckland, S.T., Siriwardena, G.M., Baillie, S.R., Wilson, J.D.,
 800 2000. Analysis of population trends for farmland birds using
 801 generalized additive models. *Ecology* 81, 1970-19.

802 Fink, D. and Hochachka, W.M., 2012. Using data mining to discover
 803 biological patterns in citizen science observations. *In* *Citizen Science:*
 804 *Public Participation in Environmental Research*. Dickinson, J.L. and
 805 Bonney, R., eds. Comstock Publishing Associates.

806 Fink, D., Hochachka, W.M., Zuckerberg, B., Winkler, D.W., Shaby, B.,
 807 Munson, M.A., Hooker, G., Riedewald, M., Sheldon, D., Kelling, S.,
 808 2010. Spatiotemporal exploratory models for broad-scale survey data.
 809 *Ecological Applications* 20, 2131-2147.

810 Franklin, J., 2009. Mapping species distributions. Cambridge University Press,
811 New York.

812 Gelfand, A.E., Schmidt, A.M., Wu, S., Silander, J.A., Latimer, A., Rebelo,
813 A.G., 2005. Modelling species diversity through species level
814 hierarchical modelling. Journal of the Royal Statistical Society: Series
815 C (Applied Statistics) 54, 1-20.

816 Gotelli, N.J. and Colwell, R.J., 2001. Quantifying biodiversity: procedures and
817 pitfalls in the measurement and comparison of species richness.
818 Ecology Letters, 4, 379-391.

819 Hastie, T. and Tibshirani, R.J., 1990. Generalized additive models. Chapman
820 and Hall, London.

821 Hedges, L.V., Tipton, E., Johnson, M.C., 2010. Robust variance estimation in
822 meta-regression with dependent effect size estimates. Research
823 Synthesis Methods 1, 39-65.

824 Hochachka, W.M., Fink, D., Hutchinson, R.A., Sheldon, D., Wong, W.-K.,
825 Kelling, S., 2012. Data-intensive science applied to broad-scale citizen
826 science. Trends in ecology & evolution (Personal edition) 27, 130-137.

827 Jeffreys 2007. PhD Thesis

828 Kery, M., Gardner, B., Monnerat, C., 2010a. Predicting species distributions
829 from checklist data using site-occupancy models. Journal of
830 Biogeography 37, 1851-1862.

831 Kery, M., Royle, J.A., Schmid, H., Schaub, M., Volet, B., Hafliger, G., Zbinden,
832 N., 2010b. Site-Occupancy Distribution Modeling to Correct Population-
833 Trend Estimates Derived from Opportunistic Observations.
834 Conservation Biology 24, 1388-1397.

835 Latimer, A.M., Wu, S., Gelfand, A.E., Silander, J.J.A., 2006. Building statistical
 836 models to analyze species distributions. Ecological applications : a
 837 publication of the Ecological Society of America 16, 33-50.
 838

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

842 Leathwick, J.R., Elith, J., Francis, M.P., Hastie, T., Taylor, P., 2006. Variation
 843 in demersal fish species richness in the oceans surrounding New Zealand: an
 844 analysis using boosted regression trees. Marine Ecology Progress Series
 845 321, 267-281.

846 Legendre, P., Fortin, M.-J., 1989. Spatial Pattern and Ecological Analysis.
 847 Vegetatio 80, 107-138.

848 Legendre, P., Fortin, M.-J., Gurevitch, J., Hohn, M., Myers, D., 2002. The
 849 Consequences of Spatial Structure for the Design and Analysis of
 850 Ecological Field Surveys. Ecography 25, 601-615.

851 MacKenzie, D.I., 2006. Occupancy estimation and modeling: inferring patterns
 852 and dynamics of species occurrence. Elsevier, Burlington, MA.

853 Mateo, R.G., Felicísimo, Á.M., Muñoz, J., 2010. Effects of the number of
 854 presences on reliability and stability of MARS species distribution
 855 models: the importance of regional niche variation and ecological
 856 heterogeneity. Journal of Vegetation Science 21, 908-922.

857 Maunder, M.N., Punt, A.E., 2004. Standardizing catch and effort data: a
 858 review of recent approaches. Fisheries Research 70, 141-159.

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

863 Olden, Julian D., Lawler, Joshua J., Poff, N.L., 2008. Machine Learning
864 Methods Without Tears: A Primer for Ecologists. The Quarterly Review
865 of Biology, 171.

866 Pattengill-Semmens, C.V., Semmens, B.X., 2003. Conservation and
867 management applications of the reef volunteer fish monitoring
868 program. Environmental Monitoring and Assessment 81, 43-50.

869 Parsons, B., Short, J., Roberts, D.D., 2009. Using community observations to
870 predict occurrence of malleefowl (*Leipoa ocellata*) in the Western
871 Australian wheatbelt. Biological Conservation 142, 364-374.

872 Pearce, J.L. and Boyce, M.S., 2006. Modelling distribution and abundance
873 with presence-only data. Journal of Applied Ecology, 43:3, 405-412.

874 Phillips, S.J., Anderson, R.P., Schapire, R.E., 2006. Maximum entropy
875 modeling of species geographic distributions. Ecological Modelling 190,
876 231-259.

877 Phillips, S.J., Dudík, M., Elith, J., Graham, C.H., Lehmann, A., Leathwick, J.,
878 Ferrier, S., 2009. Sample selection bias for presence-only distribution
879 models: implications for background and pseudo-absence data.
880 Ecological Applications, 19:1, 181-197.

881 Royle, J. A., Chandler, R. B., Yackulic, R. B. and Nichols, J. D., 2012.
882 Likelihood analysis of species occurrence probability from presence-

883 only data for modelling species distributions. *Methods in Ecology and*
884 *Evolution*, 3:3, 545-554.

885 Royle, J.A., Dorazio, R.M., 2008. Hierarchical modeling and inference in
886 ecology : the analysis of data from populations, metapopulations and
887 communities. Amsterdam ; Boston : Academic, c2008. 1st ed.

888 Royle, J.A., Kéry, M. Gautier, R., Schmid, H., 2007. Hierarchical spatial
889 models of abundance and occurrence from imperfect survey data.
890 *Ecological Monographs*, 77:3, 465-481.

891 Royle, J.A., Link, W.A., 2006. Generalized site occupancy models allowing for
892 false positive and false negative errors. *Ecology* 87, 835-841.

893 Sarda-Palomera, F., Brotons, L., Villero, D., Sierdsema, H., Newson, S.E.,
894 Jiguet, F., 2012. Mapping from heterogeneous biodiversity monitoring
895 data sources. *Biodiversity and Conservation* 21, 2927-2948.

896 Sauer, J.R., LINK, W.A., 2002. Hierarchical modeling of population stability
897 and species group attributes from survey data. *Ecology* 86:1743–1751

898 Sela, S.J. and Simnoff, J.S., 2012. RE-EM trees: a data mining approach for
899 longitudinal and clustered data. *Machine Learning*, 86: 169-207.

900 Segurado, P., Araújo, M.B., Kunin, W.E., 2006. Consequences of spatial
901 autocorrelation for niche-based models. *Journal of Applied Ecology*,
902 43, 433-444.

903 Seys, J., Offringa, H., Van Waeyenberge, J., Meire, P., Kuijken, E., 2002. An
904 evaluation of beached bird monitoring approaches. *Marine pollution*
905 bulletin 44, 322-333.

906 Silvertown, J., 2009. A new dawn for citizen science. *Trends in ecology &*
907 *evolution (Personal edition)* 24, 467-471.

908 Snäll, T., Kindvall, O., Nilsson, J., Pärt, T., 2011. Evaluating citizen-based
 909 presence data for bird monitoring. *Biological Conservation* 144, 804-
 910 810.

911 Stockwell, D. and Peters, D., 1999. The GARP modeling system: problems
 912 and solutions to automated spatial prediction. *International Journal of*
 913 *Geographic Information Science*, 13:2, 143-158.

914 Sullivan, B.L., Wood, C.L., Iliff, M.J., Bonney, R.E., Fink, D., Kelling, S., 2009.
 915 eBird: A citizen-based bird observation network in the biological
 916 sciences. *Biological Conservation*, 142, 2282-2292.

917 Swenson, N.G., 2011. The role of evolutionary processes in producing
 918 biodiversity patterns, and the interrelationships between taxonomic,
 919 functional and genetic diversity. *American journal of botany* 88:3, 472-
 920 480.

921 Tulloch, A. I. T. and Szabo, J. T., 2012. A behavioural ecology approach to
 922 understand volunteer surveying for citizen science datasets. *Emu*, 112,
 923 313-335.

924 Ward, T., Hastie, T., Barry, S., Elith, J., and Leathwick, J.R., 2009. Presence-
 925 only data and the EM algorithm. *Biometrics*, 65, 554-563.

926 Wikle, C. K., 2003. Hierarchical Bayesian models for predicting the spread of
 927 ecological processes. *Ecology*, 84:6, 1382-1394.

928 Wiggins, A., Crowston, K., 2011. From Conservation to Crowdsourcing: A
 929 Typology of Citizen Science, In *System Sciences (HICSS)*, 2011 44th
 930 Hawaii International Conference on. pp. 1-10.

931 Wiggins, A., Newman, G., Stevenson, R.D., Crowston, K., 2011. Mechanisms
 932 for Data Quality and Validation in Citizen Science, In *e-Science*

933 Workshops (eScienceW), 2011 IEEE Seventh International Conference
934 on. pp. 14-19.

935 Zhang, Hao. On estimation and prediction for spatial generalized linear mixed
936 models. *Biometrics* 58.1 (2002): 129-136.

937 Zuur, A.F., 2009. *Mixed effects models and extensions in ecology with R*.
938 Springer New York, New York.

939 Zuur, A.F., Smith, G.M., Ieno, E.N., 2007. *Analysing ecological data*. Springer,
940 New York.

941

942

Figure Captions

Figure 1. RUSSEL ML ANALYSIS

Figure 2.

Amanda LME example

Figure 3.

Tom hierarchical example

Figure 4.

Jon rarefaction example

961 **TO BE REMOVED**

962 **Table 1.** Types of data collected in citizen-science biodiversity databases.

Data type	Presence -only	Presence- absence	Abundance	Covariates
Correct identification	X	X	X	X
Location recorded	X	X	X	X
Time recorded	X	X	X	X
Independence		X	X	X
Effort measured		X	X	
Consistent effort		X	X	X
Consistent measurement			X	X

TO PUT IN SOM

Table 2. Statistical approaches and software packages available for dealing with error and bias in citizen science data.

	Method	R package	Package reference ^a
models	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 Abdoell, 2008
	Boosted Regression Tree	gbm	Ridgeway, 2013
	CART	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, P., 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)

^aR package citations are available in the supplementary reference material

IN SOM

Table 3. Studies that have used modern statistical approaches to address sampling issues in citizen science databases.

Model type	Source ^a	Data ^b	Issue	Study	Findings
GLM	CS	PA	Identification	Delaney, 2008	Age and education predicted rates of false identification of invasive crabs
GLM	CS	A	Identification	Crall et al., 2011	Volunteers that were more confident performed better at species identifications
GLMM	CS	P	Spatial clustering	Brunsdon and Comber, 2012	Onset of spring was shown to gradually advanced 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.
Hierarchical	CS	PA	Detection	deSolla, 2005	Survey effort is related to probability of detecting rare frogs from calls
Hierarchical	CS	PA	Detection	Kery, et al., 2010	Accounting for detection in SDMs led a 2-fold increase in estimated site occupancy
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	False-positive	Farmer et al., 2012	Experts were prone to over-counting of rare bird species, novices were prone to over-counting common species
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	Armano et al., 2012	Accounting for site-level effects allowed for more accurate estimation of population trends
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

^a Data Sources include: Citizen Science (CS), Natural History Collections (NHC) and distributed sampling databases (DDB). ^b Data types include:

(P), presence-absence (PA), abundance (A), biodiversity (B) or derived parameters (D). For each paper we have included a result that shows how the analysis helped improve inference.