Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2013, **4,** 236–243 doi: 10.1111/2041-210x.12004

Presence-only modelling using MAXENT: when can we trust the inferences?

Charles B. Yackulic[1,2,*,†], Richard Chandler[1], Elise F. Zipkin[1], J. Andrew Royle[1], James D. Nichols[1], Evan H. Campbell Grant[1] and Sophie Veran[1]

[1]*U.S. Geological Survey, Patuxent Wildlife Research Center, 12100 Beech Forest Road, Laurel, MD, 20708, USA; and*
[2]*Department of Ecology and Evolutionary Biology, Princeton University, Princeton, NJ, 08544, USA*

Summary

1. Recently, interest in species distribution modelling has increased following the development of new methods for the analysis of presence-only data and the deployment of these methods in user-friendly and powerful computer programs. However, reliable inference from these powerful tools requires that several assumptions be met, including the assumptions that observed presences are the consequence of random or representative sampling and that detectability during sampling does not vary with the covariates that determine occurrence probability.

2. Based on our interactions with researchers using these tools, we hypothesized that many presence-only studies were ignoring important assumptions of presence-only modelling. We tested this hypothesis by reviewing 108 articles published between 2008 and 2012 that used the MAXENT algorithm to analyse empirical (i.e. not simulated) data. We chose to focus on these articles because MAXENT has been the most popular algorithm in recent years for analysing presence-only data.

3. Many articles (87%) were based on data that were likely to suffer from sample selection bias; however, methods to control for sample selection bias were rarely used. In addition, many analyses (36%) discarded absence information by analysing presence–absence data in a presence-only framework, and few articles (14%) mentioned detection probability. We conclude that there are many misconceptions concerning the use of presence-only models, including the misunderstanding that MAXENT, and other presence-only methods, relieve users from the constraints of survey design.

4. In the process of our literature review, we became aware of other factors that raised concerns about the validity of study conclusions. In particular, we observed that 83% of articles studies focused exclusively on model output (i.e. maps) without providing readers with any means to critically examine modelled relationships and that MAXENT's logistic output was frequently (54% of articles) and incorrectly interpreted as occurrence probability.

5. We conclude with a series of recommendations foremost that researchers analyse data in a presence–absence framework whenever possible, because fewer assumptions are required and inferences can be made about clearly defined parameters such as occurrence probability.

Key-words: AUC, detection, occurrence, prevalence, sample selection bias

Introduction

Species distribution modelling has a long tradition in ecology and is becoming increasingly important in applied ecology as researchers and managers seek to understand current species distribution patterns and to predict future distributions in the face of climate change, human-assisted invasions and many other ongoing environmental changes. Numerous methods exist to model species distributions when either repeated (i.e. multiple visits to a subset of specific sites) or single-visit 'presence–absence' data are available (e.g. Austin 1998; Guisan & Zimmermann 2000; MacKenzie *et al.* 2006; Royle & Dorazio 2008). In recent decades, there has also been increasing focus on developing methods to model presence-only data (i.e. data lacking information on surveyed locations where a species did not occur). Many of these methods originate from machine learning, an area of statistics that was unfamiliar to many ecologists until recently (Olden, Lawler & Poff 2008).

Regardless of a method's pedigree, our ability to trust a model's output is directly linked with the degree to which the assumptions of the model are met. In the case of presence-only data, a key assumption is that sampling is either random or representative throughout a landscape (N.B. Presence–absence analyses are less sensitive to this assumption provided that the model is not grossly misspecified. In other words, so long as important covariates are not missing and the range of covariate values that are sampled is similar to the range of covariate

*Correspondence author. E-mail: cyackulic@usgs.gov
†Present address: Grand Canyon Monitoring and Research Center U.S. Geological Survey, 2255 N. Gemini Dr., Flagstaff, AZ, 86001, USA

values in the overall landscape). In the event that sampling does not meet these standards, it may still be possible to reach reasonable inference from presence-only data provided that we have some knowledge of the spatial distribution of sampling effort and have attempted to correct for variation in sampling intensity (Araújo & Guisan 2006; Pearce & Boyce 2006). Presence-only modelling also shares with most single-visit presence–absence studies the assumption that detection probability is constant across sites. This assumption need not be met in occupancy models that use repeat-visit data to estimate detection and occurrence probabilities (MacKenzie *et al.* 2006; Royle & Dorazio 2008).

Based on our experience reviewing presence-only studies for journals and in management contexts (e.g. habitat modelling for endangered species), we hypothesized that the assumptions of presence-only models are frequently violated in applications to real data. We tested this hypothesis by reviewing 108 articles based on empirical data and identified through Web of Science searches. These 108 articles were published during the time period of 2008–2012, with 78 of these papers coming from 2008 to 2010 and 30 published in the first half of 2012. We identified these articles by searching in Web of Science with the terms 'MAXENT' and 'species distribution' and manually filtering out papers that dealt mainly with simulated data. We chose to focus on MAXENT papers, because it was the most popular presence-only method used during the time period we investigated. As such, our conclusions about the prevalence of various practices are strictly limited to MAXENT use over the period of 2008–2012; however, as many of the issues we identify are not particular to MAXENT, but rather to presence-only modelling in general, our recommendations should apply more broadly.

Our main focus was to understand the degree to which the above assumptions were being met. However, we also chose to examine other problematic issues of implementation of presence-only analysis methods including: whether output was being interpreted as occurrence probability, whether studies were reporting the information that would allow readers to critically evaluate modelled relationships (e.g. parameter values or response curves) and lastly how complex were models compared to the available data. Answering these questions sometimes required subjective assessments because relevant information was not explicitly stated. The authors of this paper are responsible for the subjective assessments and are all quantitative ecologists or statisticians with experience in estimation and study design. However, we recognize that we may have erred in our assessments of individual articles, so we include our assessments of individual papers in the Supporting information to be transparent. In the following sections, we first review the assumptions of presence-only modelling and report the findings of our questionnaire and then explain why these issues are relevant to our overarching concern that analyses lead to reliable inferences. We conclude with a series of recommendations that, if followed, should both improve the application of presence-only methods and increase the transparency of results from these models.

Inference from presence-only data requires strong assumptions that are frequently violated

The probability of a given location being included in a presence-only data set is the product of three probabilities: (i) sampling probability–the probability that the location was surveyed, (ii) occurrence (or occupancy) probability–the probability that the location was occupied and (iii) conditional detection probability–the probability that the species was detected given that the location was both occupied and sampled. Ecologists are mainly interested in occurrence probability and seek to model and/or remove (via study design) the influences of sampling and (conditional) detection probabilities. For some types of presence-only analysis, it may be acceptable to ignore sampling and detection probability if they are constant with respect to the environmental covariates that determine occupancy. However, if sampling or detection probabilities vary with key environmental covariates and if this variation cannot be objectively quantified and included in the analysis, then it is impossible to separate the influences of these probabilities from the quantity of interest (i.e. the occurrence probability or the relative probabilities of occurrence).

SAMPLING PROBABILITY

Two of the most common examples of presence-only data sets are museum specimens and herbarium records. These data rarely arise from random or systematic sampling; rather, the data in such collections have generally been collected 'without planned sampling schemes' and 'the intent and methods of collecting are rarely known' (Elith *et al.* 2006). Lack of a systematically planned random or stratified sample can often lead to biases in sampling intensity. Many museum records are produced through sampling efforts that focus on locations where a species is expected to occur or which are most accessible. For example, herbarium records are often collected near roads (Pearce & Boyce 2006), and in many landscapes, roads may be more common in particular habitats or elevations (e.g. roads may be found in valley bottoms near riparian areas in some landscapes or along ridge lines in other areas). Presence-only methods that do not account for these biases will therefore return biased estimates when comparing habitats found near and far from roads. This is concerning because the existence of the vast collections in museums and herbariums is often cited as a key factor motivating the development of MAXENT and other presence-only methods (Elith *et al.* 2006; Phillips, Anderson & Schapire 2006); however, without correcting for the severe sample selection biases in many of these data sets, they are among the least appropriate choices available for analysis using presence-only methods.

Developers of some presence-only software have begun to introduce methods that allow users to correct for sample selection bias when additional information on sampling effort is available (Phillips *et al.* 2009; Chakraborty *et al.* 2011). Phillips *et al.* (2009) show that inference from presence-only modelling is fairly robust to sample selection bias so long as background points have the same sample selection bias as the presence points. Phillips *et al.* (2009) suggest two ways of introducing sample selection bias into the background points. One method consists of creating 'bias grids' based on known biases in sampling, while the other consists of treating points where other species in the same data set were observed, but not the focal species, as background points. Our main concern with the bias grid approach is that the nature of sampling biases is likely to be unknown or only partially known (i.e. some biases may be more obvious than others) in most situations where species data are collected haphazardly, and as such, this approach may have limited applicability. The target group (Phillips *et al.* 2009) approach is likely to more broadly applicable; however, potential users should be aware that it implicitly assumes that species have an equal probability of being recorded at all sites. In other words, it may not solve some other common issues with herbarium and museum data. For example, target group methods may not account for sample selection bias when a collector focused on obtaining a set number of samples of a particular species and stops collecting that species afterwards, while still collecting rarer species, or when a collector focuses on certain microhabitat types, at the expense of others, as a collection matures. Moreover, if the target group is appropriate, then users could simply use presence–absence methods rather than presence-only modelling.

Our literature review demonstrates that the random sampling assumption of presence-only modelling is rarely met. Of 108 papers, 76 were based on data that were not likely to have come from random sampling, 21 did not provide enough information to judge and 11 seemed to have come from appropriate sampling methods (Table 1). Fortunately, a greater percentage of papers published in 2012 were likely to have come from random sampling as opposed to the earlier period (2008–2010), suggesting that perhaps reviewers and authors are becoming more aware of the importance of this assumption; however, the rate in 2012 was still very low (24%). Many articles did not describe the sampling design in sufficient detail to allow for objective classification. We took a conservative approach in our subjective assessments, including the category, 'did not provide enough information to judge', as an option. We also note that 'bias grids' were not used in any of these studies, while target groups were used in at least three papers (Mateo *et al.* 2010; Milanovich *et al.* 2010; Urbina-Cardona & Flores-Villela 2010).

In the process of our review, we noted many statements suggesting that MAXENT is specifically designed for analysing data that do not result from random sampling. As has been stressed elsewhere, the only way to remove sampling bias is through study design and/or modelling of the bias (e.g. through target groups when appropriate). Yet, as previously mentioned, we question whether the degree of sampling bias can ever be known for many data sets that arise from museum or herbarium collections.

DETECTION PROBABILITY

A large body of evidence demonstrates that detection probability often varies with the same covariates that determine occurrence probability and that failure to account for this can mislead inference (Tyre *et al.* 2003; MacKenzie *et al.* 2006; Dorazio in press). Detection probability and occurrence probability might both increase or decrease with a given covariate if population density responds to the covariate (e.g. Royle,

Table 1. Questions and summary responses based on 78 articles published between 2008 and 2010 and 30 articles published in the first half of 2012 [Correction added after online publication 6 December 2012: responses for question 1 have been changed]

		Frequency (% of clear responses)		
Questions	Response	2008–2010	2012	Total
1. Is it likely that the presence-only data suffers from sample selection bias (nonrandom sampling)?	Yes (Y)	57 (92%)	19 (76%)	76 (87%)
	Unclear (–)	16	5	21
	No (N)	5 (8%)	6 (24%)	11 (13%)
2. Does article acknowledge detectability and/or heterogeneity in detectability? (No articles discussed heterogeneity in detectability.)	Mentioned detectability (Y)	12 (15%)	3 (10%)	15 (14%)
3. Were absence data available and discarded (i.e. could they have done a PA analysis)?	Yes (Y)	27 (36%)	9 (35%)	36 (36%)
	No absence data (N)	47 (64%)	17 (65%)	64 (64%)
	Unclear/Used for comparison (–)	4	4	8
4. Was MAXENT's output interpreted as an occurrence probability? (Possible answers: (a) Yes and interpretation of results relied heavily on this assumption, (b) Yes but results not dependent on assumption, (c) No.)	Yes (a or b)	34 (44%)	24 (83%)	58 (54%)
	(a)	20 (26%)	15 (52%)	35 (33%)
	(b)	14 (18%)	9 (31%)	23 (21%)
	No (N)	44 (56%)	5 (17%)	49 (46%)
	Unclear		1	1
5. Were response curves or betas reported? (Possible answers: (a) Response curves, (b) Beta values, (c) Signs of betas, (d) No.)	(a)	11 (14%)	4 (13%)	15 (14%)
	(b)	0	1 (3%)	1 (1%)
	(c)	0	2 (7%)	2 (2%)
	No (N)	67 (86%)	23 (77%)	90 (83%)
6. How many presences were used?	See Fig. 1			
7. How many covariates were tested?	See Fig. 1			

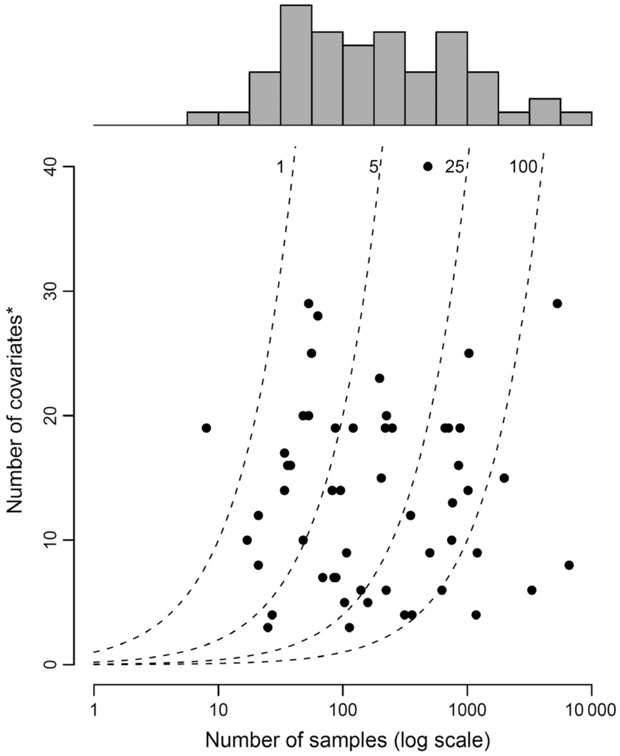

Fig. 1. Scatterplot and marginal histograms detailing number of covariates and number of presences for 54 articles that reported both values. Points to the left of the dashed lines have fewer than that many data points per covariate. *These are the numbers of covariates put into the model. The actual number of effective parameters was not reported for most articles and could be substantially more (multiple features per covariate) or less (weight for feature estimated at zero).

Nichols & Kery 2005). As a result, portions of the geographic range where the population is most dynamic (low densities, high turnover, etc.) are precisely where nondetection is most likely to occur. Such regions of rapid change in occupancy with corresponding low-detection probabilities are expected to occur at the fronts of invasions, edges of ranges, areas of range contractions and areas of range overlap of competing species (Doherty, Boulinier & Nichols 2003). Because all presence-only methods and most presence–absence methods do not separate detection probability and occurrence probability, their estimates will tend to overestimate the importance of a covariate in determining occupancy when occupancy and detection are both linearly related to a covariate with the same sign. On the other hand, detection probability and occupancy probabilities can also have opposite signs when species prefer habitats that are difficult to survey. This will lead to underestimation of the covariate effect on occurrence probability when detection probability and occupancy are not modelled separately (see MacKenzie *et al.* 2006, Fig. 2.3).

In our survey, only 15 of 108 articles acknowledged the potential for detectability issues, and no articles included discussion of how detectability might vary with habitat covariates used in their models. Finally, we note that heterogeneity in detection probability creates problems for presence-only modelling regardless of whether space is randomly sampled (Dorazio in press). Readers should be aware that methods exist to account for detection probability even when data come from nontraditional sources including, for example, independent expert surveys (Karanth *et al.* 2009) or species checklists (Kery, Gardner & Monnerat 2010).

A NOTE ON PRESENCE–ABSENCE DATA

As has been noted elsewhere, absence data should be used whenever they are available, and discarding them involves a loss of information (Brotons *et al.* 2004; Ward *et al.* 2009). In particular, when detection probability is 1, the investigator is removing unequivocal information about known absences. Even when detection probability is less than one, presence–absence is preferable under most circumstances, particularly when there is sample selection bias, and the model is not grossly misspecified (Phillips *et al.* 2009; Table 2). As the data in many of the surveyed papers came from nonrandom sampling and presence–absence analyses are clearly preferable under these conditions (Table 2), it was curious that 36 papers conducted presence-only analyses when absence data were available. It is possible that investigators chose presence-only methods, in general, and MAXENT, in particular, because of the misconception that MAXENT is designed for analysing data from nonrandom sampling. Alternatively, users may have mistakenly believed that removing potentially 'false absences' (i.e. nondetections when the species is present) from the data solves the problem associated with imperfect detection.

Table 2. Corrections required for both presence–absence and presence-only analyses under various assumptions

	Detection probability		
	Equal to one	Less than one and constant	Varies*
Sampling probability			
Constant	Presence–absence analysis preferable. Presence-only allowable, but many methods only yield relative occurrence probability	Relative measures of occurrence possible using both presence–absence and presence-only; Royle *et al.* (2012) allows estimation of occurrence probability provided that there is a relationship between occurrence and covariates. Presence–absence methods yield occurrence probability when provided with information on detection probability	Presence–absence analysis only; requires estimating relationship between detection probability and environmental covariates [e.g. through multiple visits to some sites and use of programs such as PRESENCE (freely available online)]
Varies*	Presence-only modelling requires that sampling intensity can be standardized objectively through modelling or subsampling of data. Presence–absence analysis provides unbiased estimates of occupancy conditional on sampled areas without covariates, but requires covariates and a reasonably well-specified model for unbiased estimates of occurrence probability across a landscape	Presence-only modelling requires that sampling intensity can be standardized objectively through modelling or subsampling of data. Requires correction for detection probability in addition to covariates for unbiased estimates across a landscape	Presence-absence analysis only; requires estimating relationship between detection probability and environmental covariates. If users want to make inferences about average occupancy across a landscape (as opposed to inferences about relationship to covariates), this estimate must be based on covariate values across the landscape. In other words, average occupancy from a non-representative sample will not be equal to average occupancy across the l andscape without additional steps

*Varies here is shorthand for varies with respect to environmental covariates that are also related to occupancy patterns.

Other problematic practices in the presence-only literature

In the process of our literature review, we became aware of two common practices that users should reconsider in analysing and reporting the outputs of presence-only modelling. Here, we briefly review why we believe they are worrisome and document their prevalence in our literature review.

OCCURRENCE PROBABILITY

A natural way to describe a species' distribution is to model the species occurrence probability as a function of covariates. Occurrence probability can be estimated from presence–absence data under a variety of conditions (Table 2). Some presence-only methods are capable of estimating occurrence probability directly from data (e.g. Lele & Keim 2006; Royle *et al.* 2012); however, inference is only possible under a limited set of conditions (random or representative sampling, detection probability constant and occurrence related only to continuous covariates) that are rarely satisfied in empirical data sets that could not also be analysed by presence–absence methods. Other methods, including MAXENT, estimate a different quantity (referred to as the 'raw output'), which is proportional to the expected number of presences per unit area (i.e. proportional to density), with similar assumptions (random or representative sampling and detection probability constant). MAXENT can also estimate occupancy (referred to as 'logistic output') if the above assumptions are met, and users have additional knowledge of the occurrence probability of a species under 'average' conditions (Phillips & Dudík 2008). Unfortunately, this information is rarely available, as evidenced by the observation that none of the 108 articles we reviewed actually reported this information. In the absence of this information, MAXENT still provides maps of its logistic output; however, it arbitrarily assumes that probability of occupancy at an average site is 0·5 (Elith *et al.* 2011). As the occupancy probability at an average site is unlikely to be exactly 0·5, the resulting probability predictions will necessarily be biased. Elith *et al.* (2011) argue that occupancy probability at average sites of a given spatial extent will always be 0·5 over some time period and that all other estimated probabilities will be accurate over this same temporal extent. We find this argument difficult to follow and accept. Even if this argument was convincing, however, it is unclear how it would clarify interpretation, as this temporal extent is unknowable without outside information about the quantities of focal interest.

Given the narrow set of circumstances that allow for unbiased estimation of occurrence from presence-only data and the near ubiquitous lack of outside information on the occurrence probability under average conditions, we expected that studies would instead present MAXENT's raw output and refrain from referring to the logistic output as occurrence probability. Unfortunately, 58 of the 108 papers used the logistic output and referred to it as occurrence probability. Moreover, the interpretation and/or discussions in 35 of these 58 papers relied heavily on their misinterpretation that the logistic output was an unbiased equivalent to occurrence probability. Lastly, the proportion of papers using the logistic output and referring to it as occurrence probability increased from 44% to 54% between the earlier and later samples of articles.

MODELLED RELATIONSHIPS

Many presence-only models, including MAXENT, are derived from a branch of statistics referred to as machine learning, whose proponents have argued that it is pointless to try to explain nature and that we should instead focus on developing the most efficient 'black box' methods for approximating the black box that is nature (Breiman 2001). An alternative approach to modelling, advocated by Cox in his comments on Breiman (2001), focuses on considering the important aspects of each new problem and considering, 'possible biases arising from the method of ascertainment (of data), the possible presence of major distorting measurement errors and the nature of processes underlying missing and incomplete data and data that evolve in time in a way involving complex interdependencies'. In the context of presence-only modelling, we have attempted to raise some of these same issues. While the creators of popular presence-only modelling methods, such as MAXENT, have also sought to address one important source of bias, sample selection bias, users do not use these features frequently and instead seem to be treating presence-only techniques as black boxes.

Perhaps, there is justification in just using these black boxes and ignoring potential biases. After all, many of these novel methods have performed well on independent data sets (Elith *et al.* 2006). On the other hand, when Phillips *et al.* (2009) addressed sample selection bias using the same data sets, they found that predictive ability increased and in some cases increased substantially over performances reported in Elith *et al.* (2006). We cannot help but wonder whether simpler approaches, potentially with lower predictive ability, would have revealed the importance of sample selection bias in the original analyses. For example, users might find it useful to make *a priori* predictions of how relative or absolute occupancy might respond to covariates and how biases might affect the realized responses and then compare modelled responses with *a priori* predictions as a simple check on whether models make sense. For example, if we knew a particular herbarium collection was biased towards river valleys near roads and we found that a presence-only analysis of a particular species suggested that occurrence probability was negatively related to soil moisture, we might be confident in the sign of the relationship (because sampling was biased positively with respect to soil moisture) but recognize that the actual parameter estimate was probably biased. On the other hand, if our analysis showed that occurrence probability was positively related to soil moisture, we would reasonably conclude that a specific sampling design (e.g. systematic data collection along a soil moisture gradient) was necessary to determine whether our result was produced by sampling bias as opposed to an actual relationship.

Most papers reported some basic information about covariates (i.e. which covariates were provided to the program, which covariates were selected and what was the relative importance of each covariate in the model); however, even this information was missing or unclear in a few instances. Moreover, additional information beyond these were lacking from many papers. For example, very few papers contained predictions about how specific covariates were expected to influence species occurrence (let alone sampling bias or detection probability) and only 18 of the 108 papers reported any details about the model that was fit. Of these 18 papers, one provided the actual parameter estimates, two reported the signs of the parameter estimates, and the remaining 15 papers showed response curves depicting changes in some form of MAXENT output as a function of covariates. Moreover, in most instances, users did not even report the number of features included in models making it impossible to accurately judge model complexity, except by coarse measures of the number of covariates and data points (Fig. 1). We could place more confidence in modelled outputs (e.g. maps) if more time was put into developing *a priori* predictions (especially for data that are nonrandomly sampled or where detection probability is likely to vary with environmental covariates), and if modelled relationships were examined critically. We believe that critical examination of modelled relationships would likely lead many users to choose less complicated response functions that are readily interpretable, recognizing that more complex models are as likely to be responding to sampling biases as to actual ecological relationships.

A FEW COMMENTS ON THE USE OF THE RECEIVER OPERATOR CHARACTERISTIC AND AREA UNDER THE ROC CURVE

The area under the ROC curve (AUC) is the statistic most frequently used to characterize model performance in the articles we reviewed. The shortcomings of AUC have been detailed extensively elsewhere, and users of AUC should be aware of these issues (Lobo, Jimenez-Valverde & Real 2008; Hanczar *et al.* 2010). In addition, users should be aware that the AUC value that is calculated by MAXENT is not AUC as it was originally defined. In their standard definitions, receiver operator characteristic (ROC) and AUC are used for the problem of classifying presences vs. absences, whereas MAXENT's versions are used for the problem of classifying presences vs. background points (which may or may not be true absences; Phillips, Anderson & Schapire 2006).The distinction between these classification problems is important and was recognized by MAXENT's creators but seems to be ignored by many of its users. MAXENT's creators should perhaps consider renaming their output (e.g. presence-only AUC – $\mathrm{AUC_{PO}}$) so that the distinction is even clearer to users. Moreover, users should be aware that if their analyses are based on haphazardly collected data and/or if detection probability varies with the covariates that determine relative occupancy levels, then $\mathrm{AUC_{PO}}$ is actually addressing the problem of classifying species detections (which are themselves a product of true presence, variation in sampling intensity and detection probability) vs. background points.

Lastly, there was a common misconception in the articles we reviewed that a particular value of $\mathrm{AUC_{PO}}$ signified whether a model was 'good' or not. For example, Brown, Spector & Wu (2008, p. 1641) argue that an $\mathrm{AUC_{PO}}$ of 0·85 is 'a baseline for model accuracy' (N.B., they cite an article referring to the traditional AUC in justifying this value). $\mathrm{AUC_{PO}}$ is a relative value for comparing the performance of different models based on the same data, and we are not aware of, nor do we think it is possible to construct, any objective argument for a particular threshold for all situations. A perfect model of a species niche may have a low $\mathrm{AUC_{PO}}$ value if the species is limited by dispersal or experiences frequent local extinctions, while a model with a high $\mathrm{AUC_{PO}}$ could be based on trivial distinctions (e.g. an analysis of the distribution of a species restricted to riparian areas in a desert that uses a large proportion of nonriparian habitat in the background sample). In addition, $\mathrm{AUC_{PO}}$ values are dependent on the ratio of prevalence to background points such that $\mathrm{AUC_{PO}}$ can be improved simply by increasing the number of background points. We suggest that the heavy reliance on $\mathrm{AUC_{PO}}$ that we observed in our literature review is not well deserved, and we hope that in coming years presence-only modelling will not rely so singularly on this questionable statistic.

Concluding thoughts and recommendations

All maps are partial truths, and when confronted with maps built from data that were obtained without any sort of sampling design and using complicated functional relationships, it is almost impossible to judge how well the map approximates reality. Some methods (target groups, bias grids) have been introduced to account for sampling selection bias; however, these methods are not being widely used and are also subject to assumptions (e.g. bias grids assume sources of bias in haphazard data sets are well known, and target groups assume species are equally likely to be detected and reported at all sites). Moreover, heterogeneity in detection probability across sample locations, a documented phenomenon for many species, is not addressed by these methods. So while recent advances in presence-only modelling allow users to fit increasingly complex response curves, precisely approximating trends in the data, it is unclear whether these trends are related to the process of interest (relative or absolute occupancy) or to biases that are not being addressed. As a consequence, many studies are potentially presenting 'precise answers to the wrong question'. That is, they may be doing a poor job of approximating the distribution of occurrence probability, but an excellent job of depicting how the product of occurrence probability, sampling probability and conditional detection probability would be distributed across a landscape if investigators continued to sample in the same haphazard way they did in collecting the data.

Important practical decisions (e.g. where to place nature reserves) and our basic understanding of species distributions would be better served if potential presence-only modellers followed these recommendations:

1 Consider the data sources. Is it possible to analyse the data in a presence–absence framework? If not, were the data collected via a standardized sampling scheme or haphazardly? If the latter, are there reasonable ways to account for sampling biases through subsampling or modelling? If an approximate answer is acceptable, does the sampling bias run counter to the expected occurrence pattern? Would time and resources be better spent developing and implementing a sampling scheme designed to address the question of interest?

2 Consider the target species and sampling method. Could detection probability be less than one? If so, could detection probability vary with the environmental covariates that determine occurrence probability? If data are amenable to presence–absence modelling and detection probability is less than one and likely to vary, a limited amount of additional sampling to estimate how detection probability varies may permit reasonable estimation of occurrence probability. If data are not amenable to presence–absence analysis, but detection probability does not vary and space was sampled randomly or representatively, the methods described in Lele & Keim (2006) and Royle *et al.* (2012) should allow direct estimation of occurrence probability, but with far greater uncertainty than one would expect from a presence–absence analysis. In the event that detection probability varies with respect to covariates that also determine occupancy, data are not amenable to presence–absence analysis and data were collected through a standardized sampling scheme, it may still be possible to estimate occurrence probability if additional data (e.g. multiple visits to a subset of sites) are collected to estimate how detection probability varies with the covariates that determine occupancy.

3 Consider potential *a priori* hypotheses. If sampling probability and detection probability are constant or have been controlled for, consider constructing *a priori* hypotheses of how occurrence varies with environmental covariates. If sampling probability and detection probability are likely to vary and are not controlled for, consider *a priori* hypotheses of how both occurrence and biases are likely to vary with environmental covariates. Are the expectations confounded, such that nothing will be learned through the modelling process?

4 Critically examine modelled relationships. If they do not agree with *a priori* hypotheses, develop alternative *a posteriori* hypotheses that could be tested through additional data collection.

5 Provide readers with the necessary information to critically evaluate your results. A hallmark of scientific reporting is that future researchers should be able to compare results of their studies to yours. Maps alone do not provide sufficient output to allow for this, and inclusion of estimated response curves and parameters, either in the bodies of studies or in appendices, would greatly improve the transparency and usefulness of presence-only studies. Moreover, authors should be encouraged to make original data available, when legally appropriate, through online appendices or data repositories.

Presence-only data are widely available, offering a great opportunity to learn about species distributions and their relationships to environmental covariates. At the same time, when data are not collected according to a structured sampling design and variation in species detection probability is not accounted for, these sampling issues greatly limit our ability to draw inferences about species distribution. Ecologists should not avoid these data simply because they were not collected using formal sampling designs; however, at the same time, we should be cautious and modest in our expectations for inference.

Acknowledgements

CBY was supported by the National Science Foundation under Award No. 1003221. This is contribution 2 of the PWRC-WMS research program and contribution 419 of the USGS Amphibian Research and Monitoring Initiative (ARMI). Any use of trade, product, or firm names is for descriptive purposes only and does not imply endorsement by the US government.

References

Araújo, M.B. & Guisan, A. (2006) Five (or so) challenges for species distribution modelling. *Journal of Biogeography*, **33**, 1677–1688.

Austin, M.P. (1998) An ecological perspective on biodiversity investigations: examples from Australian eucalypt forests. *Annals of the Missouri Botanical Garden*, **85**, 2–17.

Breiman, L. (2001) Statistical modeling: the two cultures (with comments and a rejoinder by the author). *Statistical Science*, **16**, 199–231.

Brotons, L., Thuiller, W., Araújo, M.B. & Hirzel, A.H. (2004) Presence-absence versus presence-only modelling methods for predicting bird habitat suitability. *Ecography*, **27**, 437–448.

Brown, K.A., Spector, S. & Wu, W. (2008) Multi-scale analysis of species introductions: combining landscape and demographic models to improve management decisions about non-native species. *Journal of Applied Ecology*, **45**, 1639–1648.

Chakraborty, A., Gelfand, A.E., Wilson, A.M., Latimer, A.M. & Silander, J.A. (2011) Point pattern modelling for degraded presence-only data over large regions. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **60**, 757–776.

Doherty, P.F., Boulinier, T. & Nichols, J.D. (2003) Local extinction and turnover rates at the edge and interior of species' ranges. *Annales Zoologici Fennici*, **40**, 145–153.

Dorazio, R.M. (in press) Predicting the geographic distribution of a species from presence-only data subject to detection error. *Biometrics*, doi: 10.1111/j.1541-0420.2012.01779.x.

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

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

Guisan, A. & Zimmermann, N.E. (2000) Predictive habitat distribution models in ecology. *Ecological Modelling*, **135**, 147–186.

Hanczar, B., Hua, J., Sima, C., Weinstein, J., Bittner, M. & Dougherty, E.R. (2010) Small-sample precision of ROC-related estimates. *Bioinformatics*, **26**, 822–830.

Karanth, K.K., Nichols, J.D., Hines, J.E., Karanth, K.U. & Christensen, N.L. (2009) Patterns and determinants of mammal species occurrence in India. *Journal of Applied Ecology*, **46**, 1189–1200.

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

Lele, S.R. & Keim, J.L. (2006) Weighted distributions and estimation of resource selection probability functions. *Ecology*, **87**, 3021–3028.

Lobo, J.M., Jimenez-Valverde, A. & Real, R. (2008) AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, **17**, 145–151.

MacKenzie, D.I., Nichols, J.D., Royle, J.A., Pollock, K.H., Bailey, L.L. & Hines, J.E. (2006) *Occupancy Estimation and Modeling: Inferring Patterns and Dynamics of Species Occurence*. Elsevier, Amsterdam.

Mateo, R.G., Croat, T.B., Felicisimo, A.M. & Munoz, J. (2010) Profile or group discriminative techniques? Generating reliable species distribution models using pseudo-absences and target-group absences from natural history collections. *Diversity and Distributions*, **16**, 84–94.

Milanovich, J.R., Peterman, W.E., Nibbelink, N.P. & Maerz, J.C. (2010) Projected loss of a salamander diversity hotspot as a consequence of projected global climate change. *PLoS ONE*, **5**, e12189, 12181–12110.

Olden, J.D., Lawler, J.J. & Poff, N.L. (2008) Machine learning methods without tears: a primer for ecologists. *The Quarterly Review of Biology*, **83**, 171–193.

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

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

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

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

Royle, J.A. & Dorazio, R.M. (2008) *Hierarchical Modeling and Inference in Ecology: The Analysis of Data from Populations, Metapopulations and Communities*. Academic Press, San Diego, California, USA.

Royle, J.A., Nichols, J.D. & Kery, M. (2005) Modelling occurrence and abundance of species when detection is imperfect. *Oikos*, **110**, 353–359.

Royle, J.A., Chandler, R.B., Yackulic, C. & Nichols, J.D. (2012) Likelihood analysis of species occurrence probability from presence-only data for modelling species distributions. *Methods in Ecology and Evolution*, **3**, 545–554.

Tyre, A.J., Tenhumberg, B., Field, S.A., Niejalke, D., Parris, K. & Possingham, H.P. (2003) Improving precision and reducing bias in biological surveys: estimating false-negative error rates. *Ecological Applications*, **13**, 1790–1801.

Urbina-Cardona, J.N. & Flores-Villela, O. (2010) Ecological-niche modeling and prioritization of conservation-area networks for Mexican Herpetofauna. *Conservation Biology*, **24**, 1031–1041.

Ward, G., Hastie, T., Barry, S., Elith, J. & Leathwick, J.R. (2009) Presence-only data and the EM algorithm. *Biometrics*, **65**, 554–563.

Received 25 June 2012; accepted 24 September 2012
Handling Editor: Robert B. O'Hara

Supporting Information

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

Table S1. Results of literature review by article.

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials may be re-organized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.