Opening the black box: an open-source release of Maxent

Steven J. Phillips[1], Robert P. Anderson[2], Miroslav Dudík[3], Robert E. Schapire[3], Mary E. Blair[1]

1 Center for Biodiversity and Conservation, American Museum of Natural History, Central Park West at 79th Street, New York, NY 10024, USA

2 Dept. of Biology, City College of New York, City Univ. of New York, New York, NY 10031, USA; Program in Biology, Graduate Center, City Univ. of New York, 365 Fifth Avenue, New York, NY 10016 USA; and Div. of Vertebrate Zoology (Mammalogy), American Museum of Natural History, Central Park West at 79th Street, New York, NY 10024

3 Microsoft Research, 641 Avenue of the Americas, 7th floor, New York, NY 10011 USA

Abstract

This software note announces a new open-source release of the Maxent software for modeling species distributions from occurrence records and environmental data, and describes a new R package for fitting such models. The new release (Version 3.4.0) will be hosted online by the American Museum of Natural History, along with future versions. It contains small functional changes, most notably use of a complementary log-log (cloglog) transform to produce an estimate of occurrence probability. The cloglog transform derives from the recently-published interpretation of Maxent as an inhomogeneous Poisson process (IPP), giving it a stronger theoretical justification than the logistic transform which it replaces by default. In addition, the new R package, maxnet, fits Maxent models using the glmnet package for regularized generalized linear models. We discuss the implications of the IPP formulation in terms of model inputs and outputs, treating occurrence records as points rather than grid cells and interpreting the exponential Maxent model (raw output) as as an estimate of relative abundance. With these two open-source developments, we invite others to freely use and contribute to the software.

New hosting and licensing for Maxent

Maxent is a self-contained Java application for species distribution modeling (SDM) based on occurrence records (locations where the species has been found) together with environmental variables such as rainfall and temperature for a surrounding study area (Phillips et al. 2006; Phillips and Dudík 2008). Since performing well in a comparison of species distribution modeling methods (Elith et al. 2006), it has been widely used: Google Scholar reports more than 6000 citations for Phillips et al. (2006) at the time of writing. Until now the software source code has been owned by AT&T, but the application has been freely available and hosted online by Princeton University (www.cs.princeton.edu/%7Eschapire/maxent).

Despite documentation of the underlying mathematics, the software has sometimes been referred to as a black box, since the underlying source code was not available. The source code is now released under the MIT open-source licence, and we invite interested developers to use and contribute to the code. The geospatial community has been a leader in open source and free software development (Bocher and Neteler 2012) and many in the ecology, evolution, and environmental science communities have called for increased openness as not only an ethical imperative but also a necessity to answer key pressing questions about global change (Wolkovich et al. 2012). The Maxent application will henceforth be hosted by the Center for Biodiversity and Conservation (CBC) at the American Museum of Natural History, at biodiversityinformatics.amnh.org/open_source/maxent, extending the role that the CBC has played in fostering the development of Maxent and hosting the New York Species Distribution Modeling Discussion Group for the past 15 years. In addition to the Maxent application, the new site contains the existing tutorial and a few key publications, as well as a link to the source code on GitHub (https://github.com/mrmaxent/maxent).

In addition to the Java source code, we announce a new R package, maxnet (https://CRAN.R-project.org/package=maxnet; Version 0.1.2, https://github.com/mrmaxent/maxent, authored by SJP), which implements Maxent using the glmnet R package (Friedman et al. 2010) for model fitting. This new

package takes advantage of the derivation of Maxent as a form of infinitely-weighted logistic regression (see below). It fits Maxent models using the same feature classes (linear, quadratic, hinge, etc.) and regularization options of the Java version.

Maxent and inhomogeneous Poisson processes

Maxent estimates the distribution (geographic range) of a species by finding the distribution which has maximum entropy (i.e., is closest to geographically uniform) subject to constraints derived from environmental conditions at recorded occurrence locations. The constraints are defined in terms of "features" (environmental variables such as temperature, and simple functions of those variables such as quadratic terms), and require that the mean of each feature should match the sample mean. This formulation is equivalent to maximizing the likelihood of a parametric exponential distribution (Phillips et al. 2004). More recently, it was noted that the exact same maximum likelihood exponential model can be obtained from an inhomogeneous Poisson process (IPP) (Aarts et al. 2012; Fithian and Hastie 2013; Renner and Warton 2013). This development is important for Maxent users, as it yields new interpretations of model inputs and outputs, and allows the use of other software packages for fitting Maxent models. Here we give a very brief overview of the IPP formulation, following Fithian and Hastie (2013); note that in following their notation, some of the same symbols (e.g., λ) are used differently from previous papers describing Maxent.

An IPP is a widely-used model for a random set Z of points falling in some domain D (Cressie 1993; Diggle 2003). To apply it to species distribution modeling, we can use the set of occurrence records for Z, while D is the geographic study area. The IPP can be defined by an intensity function λ which assigns a non-negative real-valued intensity $\lambda(z)$ to each point z in D. It can be thought of as indexing the likelihood that a point (here, an occurrence record of the species) falls at or near z. We can define a probability density over the domain D by:

$$p_\lambda(z) = \lambda(z) / \int_D \lambda(z) dz \qquad \text{(Equation 1)}$$

where the denominator simply makes p_λ sum to 1. An IPP with intensity λ is defined as an independent and identically distributed (i.i.d.) sample from p_λ, whose size (number of points) is a Poisson random variable with mean $\int_D \lambda(z)dz$. Warton and Shepherd (2010) suggested modeling the occurrence records for a species as arising from an IPP whose intensity $\lambda(z)$ is a log-linear function of a vector of real-valued features $\mathbf{x}(z)$:

$$\lambda(z) = \exp(\alpha + \beta' \mathbf{x}(z)) \qquad \text{(Equation 2)}$$

The coefficient α is essentially a normalizing constant, giving no information about the species' distribution – its maximum likelihood value simply ensures that $\int_D \lambda(z)dz$ equals the total number of occurrence records. Conditioned on the number of occurrence records, the likelihood of the IPP is the same as Maxent's likelihood, since p_λ is exactly Maxent's exponential distribution. The maximum likelihood values of the coefficients β are therefore exactly the same as those given by Maxent. This equivalence still holds true when using regularization (as is done by the Maxent software), by which a penalty term (Phillips et al. 2004; Elith et al. 2011) is added to the log likelihood to penalize larger values of the coefficients and thereby produce a simpler model.

Implications for model inputs

While the IPP model can be defined for a finite discrete domain (such as a regular grid), it is perhaps most natural for a continuous D; for SDM, this means that occurrence records are considered to be points (with zero area) in geographic space rather than sites or quadrats of some non-zero area. The IPP then models the process of drawing some point locations randomly from the locations of all individuals of the species (leaving aside complications due to sample selection bias; see Renner et al. 2015; Fithian and Hastie 2013; Phillips et al. 2009). Hence, it

models occurrence records as being obtained with probability proportional to the local abundance of the species. This contrasts with Phillips et al. (2006), who outlined an idealized data model in which the domain D is a finite grid of equal-sized cells, with occurrence records corresponding to grid cells randomly selected from those occupied by the species. In this latter case, a grid cell with a single individual is considered as likely to have an occurrence record as a grid cell in which the species is very abundant; the difference between the two data models is more pronounced for larger grid sizes. Given the realities of biological sampling, the truth surely lies somewhere between these two extremes, and depends strongly on the data-collection methods used in the field (Renner et al. 2015). For example, the density of records derived from incidental sightings will likely be strongly affected by local abundance; in contrast, presence-only data sets of occurrence records from intensive sampling of transects will not distinguish between areas with a few individuals truly present per transect versus those with many.

For the IPP model, we expect to have more occurrence records in areas and environmental conditions where the species is abundant. However, care should be taken when multiple records lie close together, since co-located or nearby records are often an artifact of spatially auto-correlated sampling (for example, records may be clustered around a research station). For these reasons, the occurrence records may need to be thinned (Boria et al. 2014; Aiello-Lammens et al. 2015) to better match the IPP's assumption of independent samples. Note that this particular issue of clustered sampling is separate from that of spatially biased sampling (Fithian et al. 2015) and it may be necessary to apply both thinning and bias correction to the same dataset (Syfert et al. 2013).

Additionally, the primary goal of SDM is often to model and understand the environmental conditions inhabited by the species, rather than simply its geographic distribution. Both for this use and to better estimate the geographic distribution, it is important that (to the degree possible) the occurrence data represent a random sample of suitable conditions in D. In addition to consideration of sampling biases (see above), this requires a careful choice of the study area D –

see for example Renner et al. (2015) and discussion of environmental equilibrium and noise assumptions of Anderson (2013).

Implications for model outputs

The IPP model gives an estimate $\lambda(z)$ of the intensity of occurrence records at or near the point z. If the sampling effort is unbiased (an unlikely assumption – see Reddy and Dávalos 2003), this is also an estimate of the relative abundance of the species: i.e., it is linearly proportional to the average number of individuals per unit area at or near z. The constant of proportionality (which we will write as c_p) between the model and the true abundance of the species cannot be determined from occurrence records alone (Fithian and Hastie 2013). Rather, some independent measure or estimate of total population size is required to estimate c_p and hence absolute abundance. Maxent models (derived from occurrence records) have indeed been found to show correlation with independently measured local abundance (VanDerWal et al. 2009; Weber et al. 2016). We note, however, that although a linear relationship is theoretically predicted between $\lambda(z)$ (Maxent's "raw" output format) and local abundance, a nonlinear (but still monotonic) relationship is predicted for transformed outputs such as the logistic transform used by VanDerWal et al. (2009).

From relative abundance to probability of presence

The interpretation of Maxent as an IPP allows Maxent's "raw" output format to be used directly as a model of relative abundance. However, many SDM uses call for models of probability of presence. Additionally, maps made from raw output do not often match ecologists' intuition about the (potential or realized) distribution of their study species. For these reasons, Maxent's default output scaling is a model of probability of presence, with an important caveat.

Consider a quadrat within the domain D, and assume the environmental conditions are constant within the quadrat. The IPP estimates the species' absolute abundance in the quadrat as

a Poisson variable with mean:

$$\text{Predicted mean abundance} = c_p A \exp(\alpha + \beta' \mathbf{x}(z)) \qquad \text{(Equation 3)}$$

where A is the area of the quadrat. The probability of presence of the species in the quadrat is the probability that there is at least one individual there, which according to the Poisson distribution is:

$$\text{Probability of presence} = 1 - \exp(-c_p\, A\, \exp(\alpha + \beta' \mathbf{x}(z))) \qquad \text{(Equation 4)}$$

thus yielding a Bernoulli generalized linear model whose link function is termed a complementary log-log (cloglog) link (Fithian et al. 2015). Note, however, that this derivation relies on the species' presence or absence at nearby sites being independent. Therefore, it may not be appropriate when species distributions (including patterns of abundance) exhibit spatial dependence beyond that owing to spatial autocorrelation of the utilized predictor variables. For example, positive autocorrelation of individuals occurs for flocking birds and for plants with limited dispersal abilities, and negative autocorrelation for territorial mammals. Note that the IPP literature includes methods to detect spatial dependence (such as Ripley's K-function) and to incorporate it into the model (e.g., using area-interaction processes) at the cost of increasing modeling complexity (Renner et al. 2015).

Because of the above derivation, a cloglog transform appears to be most appropriate for estimating probability of presence, and Maxent version 3.4.0 now uses it by default. The previous default (a logistic transform, Phillips and Dudík 2008) is now available as an option. For both transforms, the entropy $H = -E_{p_\lambda}[\ln(p_\lambda)]$ of the probability distribution p_λ is used as a constant offset to the linear model; specifically, the new cloglog transformation estimates probability of presence as:

$$\text{Probability of presence} = 1 - \exp(-\exp(H) p_\lambda(z)). \qquad \text{(Equation 5)}$$

Note that this estimate is appropriate for some quadrat size, but we cannot say explicitly what that size is, since it depends on the (unknown) c_p (this is the important caveat mentioned at the beginning of this section). At best, we can define the quadrat size implicitly: consider a point

z whose log probability under p_λ equals the mean log probability, i.e., $\ln(p_\lambda(z)) = E_{p_\lambda}[\ln(p_\lambda)]$.

Such a point could be called a "typical" location of the species, as predicted log abundance there is average among all points where individuals of the species are found. The predicted probability of occurrence in a quadrat centered at such a typical point z is $1 - 1/e \approx 0.632$, corresponding to a predicted abundance of one individual per quadrat. This is similar to Maxent's logistic output, which gives a predicted probability of occurrence of 0.5 for such a location. In general, the cloglog transformed output is somewhat greater than the logistic one (Figure 1), especially at higher values. The main effect of using the cloglog rather the logistic transform is that areas of moderately high output (yellow and orange in Figure 2 left) are more strongly predicted (relatively warmer colors in Figure 2 right). We emphasize that the use of entropy as an offset is somewhat arbitrary, but has the advantage of being scale independent (for example, it would not be affected by changing units from meters to kilometers) and produces output values (and hence mapped predictions) with good visual discrimination across the same range of values (0-1) for all species. Importantly, whenever more is known about the species, such as its absolute abundance at some sites or its total population size, the use of entropy as an offset can be avoided by deriving an estimate of c_p and therefore the probability of presence for quadrats of any given size. This is analogous to using addition information about the species' prevalence to derive an appropriate offset for the logistic transform (Guillera-Arroita et al. 2014).

Although the above derivation of the cloglog transform provides a stronger theoretical justification than the robust Bayes argument (Phillips and Dudík 2008) for the logistic transform, the cloglog transform may have only a small effect on model performance. On a large reference data set (that of Elith et al. 2006), the cloglog transform marginally lowered values of model calibration (measured by correlation with 0/1 data encoding observed absences/presences, known as the COR statistic) relative to the logistic transform for models made with random background data (Table 1). In contrast, and more importantly, it improved this measure of model performance when target-group background (Phillips et al. 2009) was

used to reduce the effects of sample selection bias. The raw output (the exponential model of Eqns 1 and 2) substantially underperformed cloglog (and other output formats), which is as expected given that COR measures ability to predict probability of presence rather than abundance. We note that rank-based statistics such as AUC (area under the receiver operating characteristic curve) are unaffected by the logistic and cloglog transforms.

Maxent as infinitely-weighted logistic regression (IWLR): the maxnet package

Because Maxent is an IPP, standard generalized linear modeling software can be used to fit Maxent models via Poisson regression (Renner and Warton 2013), or even more conveniently, using standard logistic regression Fithian and Hastie (2013). Specifically, the latter authors showed that the coefficients β of the Maxent or IPP model can be fitted via a weighting process they call infinitely-weighted logistic regression (IWLR). The idea is to fit a logistic model to occurrence records (with response variable $y = 1$) and background data (points chosen randomly from the domain D, with response variable $y = 0$). This process yields coefficients β for an exponential model, and has been used in studies of resource selection by animals (Manly et al. 2002), but may not produce the same values of the coefficients as Maxent. The novel contribution of Fithian and Hastie (2013) was to give a large weight W to all the background data and to show that the limit (as W tends to infinity) of the resulting vector of logistic regression coefficients equals the Maxent (and IPP) coefficients. This allows Maxent (and IPP) models to be fitted using standard GLM software. The new R package for fitting Maxent models (maxnet, available at https://CRAN.R-project.org/package=maxnet) does just this – leveraging the glmnet R package (Friedman et al. 2010) to fit an l_1-regularized logistic regression model with a large weight W. A weight of $W = 100$ is used by default, in contrast to a weight of 1 for occurrence records.

Instead of upweighting background points, we may equivalently downweight presence points (Renner et al. 2015). In addition, the intercept term (α, above) can be manipulated by

choosing the background weights based on the area of the study region, so that the resulting IPP is scaled in units of occurrence records per unit area (Renner et al. 2015). Given the area of the study region, this weighting scheme could easily be used with maxnet, though it would not affect any of the standard output formats (raw, cloglog etc.) since none of them use α.

There is a variety of R packages available for fitting point process models (Renner et al. 2015), so why introduce another? The novel contribution of maxnet is to implement all the derived feature classes (especially hinge features) and default tuned regularization values of the Maxent Java application, so that Maxent models can be fitted natively and easily in R. The package is brief – about 200 lines of code implementing feature classes and regularization parameters, model fitting, predicting from a model, and plotting of response curves. Additionally, it provides some simple use examples based on the Bradypus variegatus (brown-throated three-toed sloth) data set from Phillips et al. (2006). The purpose of the package is to replicate the behaviour of the Maxent Java application by using the equivalence with IPPs; this complements Renner et al. (2015), who show (Appendix Section 6) how to adjust default settings in the Maxent application in order to fit an IPP.

When run on the data set of Elith et al. (2006), maxnet has similar performance to the Maxent Java application. Small differences are likely due primarily to different implementations of hinge features and different random choices of background data. In order to limit computation time, the maxnet implementation of hinge features uses 50 hinge features per environmental variable by default, with evenly spaced knots, in contrast to Maxent which may use one knot per unique value of the environmental variable. The scripts used to run maxnet on that data set appear in the online Appendix, along with further examples of usage.

A change in default feature classes

Both the previous and current releases of Maxent allow the use of quadratic, product (or interaction), threshold (or step-function) and hinge (piecewise linear) features, in addition to the original environmental variables (linear features). The selection of feature classes for use in the

model depends only on sample size, though l_1-regularization forces many coefficients to zero. Given enough occurrence records (80+), all of the derived feature classes were previously considered by the model. Version 3.4.0 differs by omitting threshold features by default (although they are available as an option), since this appears to improve model performance generally and results in models that are smoother and simpler, and hence likely to be more realistic. Avoiding use of threshold features makes a small but useful improvement to the performance of Maxent (Table 1) on the data set of Elith et al. (2006). The differences in AUC and COR are similar to differences within the three groupings of modeling methods in Elith et al. (2006), but smaller than differences among groupings. Apparently hinge features, which were introduced to Maxent later than threshold features (Phillips and Dudík 2008), are best used as a replacement for threshold features rather than as a complement. Hinge features provide at least as much flexibility in the fitted response to predictor variables as threshold features, while tending to reduce over-fitting to the training data. The scripts used to run Maxent with various settings for Table 1 appear in the online Appendix. Unfortunately, the data set of Elith et al. (2006) is not yet publicly available, but when it is publicly released, it will be added to the online Appendix too, so that the values in Table 1 can be easily replicated.

We emphasize that the settings used in the analyses reported in Table 1 are merely defaults, and the best choice for other data sets may be different (Merow et al. 2013; Radosavljevic and Anderson 2014). We have also found that product features barely improve average performance on the data set of Elith et al. (2006) (not shown), and could usually be omitted in order to make simpler and more easily interpreted models. Importantly, avoiding product features enables the use of the Explain tool to interactively explore model predictions (Elith et al. 2010; Renner et al. 2015).

Future directions

Species distribution models based on Maxent and IPP remain an active area of research, as new methods are developed to accommodate the challenging nature of occurrence data and

species distributions (Fithian et al. 2015; Merow et al. 2016). The open-source release of the Maxent Java code, together with the maxnet R package, will facilitate the work of others in improving the science of modeling species distributions. Similarly, we hope that it will facilitate the practical and public use of Maxent for mapping and preserving biodiversity, as done by the Atlas of Living Australia (http://www.ala.org.au/).

The IPP perspective on Maxent input requirements and interpretation of model outputs should provide new direction for research regarding studies of population abundance. One conclusion is that under certain assumptions regarding the occurrence data, Maxent's raw (exponential) output can be interpreted as a model of relative abundance (Renner et al. 2015). A natural question arising from this is whether in practice, raw output indeed better correlates with local abundance than logistic output (as used by VanDerWal et al. 2009). This should also inform meta-analyses of relationships between SDM output and abundance and other measures of population performance (Weber et al. 2016).

The IPP model also offers new insights into spatial dependence for Maxent models, along with some tools for detecting spatial trends in residuals (Renner et al. 2015). When we use the IPP intensity to infer a Poisson distribution for abundance within a quadrat (Eqn. 3) and thereby determine probability of presence therein, we make a strong independence assumption, namely that presence of the species at nearby points within the quadrat is conditionally independent given the predictor variables. This assumption may often be violated on fine spatial scales, for example because individuals in close proximity interact through competition, reproduction, etc., and disturbances such as fire impose spatial signatures on species' patterns of distribution (including abundance). Outstanding questions include: How important is this violation in practice when using IPP / Maxent for modeling abundance or probability of presence from occurrence data? In what circumstances should modellers resort to more complex variants of IPPs (such as Gibbs or Cox processes; Renner et al. 2015) to explicitly model spatial dependence? Similar issues arise when modeling abundance from count data: patterns of occurrence and abundance may be affected by different processes, resulting in excess zeroes in

the abundance data (Wenger and Freeman 2008). This is likely related to the finding that Maxent models predict a "potential/maximal abundance", which may not be attained at all sites (VanDerWal et al. 2009). Zero-inflated models are often used in place of simpler Poisson models when modeling abundance from count data that exhibit excess zeroes (Barry and Welsh 2002). Perhaps similar ideas can be applied to occurrence data.

The maxnet R package encodes feature classes and regularization defaults in order to fit the same models as the Maxent Java application, opening up new ways to better integrate Maxent modeling with the wide variety of visualization and analysis tools available in R. Future contributions to maxnet could facilitate this integration, for example by contributing code and/or a vignette that links maxnet with the dismo package or the ENMeval package (Muscarella et al. 2014) which manage data preparation, modelling and evaluation for SDM. A test suite would be very helpful to ensure consistency as new functionality gets added to the package. Only some capabilities of glmnet are used by maxnet, and others (such as elastic net regularization and data-driven feature selection) could be incorporated. Alternatively, glmnet or other IPP packages (such as ppmlasso; Renner et al. 2015) could be used directly on the data set of Elith et al. (2006) (i.e., not via maxnet) and compared with the performance of maxnet described here in order to determine the most effective use of IPPs on species occurrence data. In addition, the standard collection of statistical analyses and maps produced as html output by the Maxent Java application could be assembled in R, and would then be available for any species distribution modeling method with an implementation in R.

Finally, we invite developers and modelers to imagine new capabilities for the Maxent Java application, and to contribute to its development. Free and open access to data, software, tools, publications, and other resources facilitate key steps towards more informed and powerful models and more inclusive research outcomes (Soberón and Peterson 2004). Through these open-source releases we aim to empower Maxent users as a community to use, contribute, and innovate towards improvement of the software, and to generate new open-source software resources and tools.

Acknowledgements

We thank the Center for Biodiversity and Conservation at the American Museum of Natural History for hosting the open-source release of Maxent, especially via the efforts of Eleanor Sterling, Peter Ersts and Ned Horning. RPA acknowledges the support of the U.S. National Science Foundation (NSF DEB-1119915 and DBI-1650241). We appreciate the thoughtful and helpful comments made by reviewers of the first draft of this paper.

Online Appendix

An online appendix contains the scripts used to generate Table 1 and Figures 1 and 2, and further examples of usage for maxnet.

References

Aarts, G. et al. 2012. Comparative interpretation of count, presence-absence and point methods for species distribution models. – Methods in Ecology and Evolution 3: 177–187.

Aiello-Lammens, M. E. et al. 2015. spThin: an R package for spatial thinning of species occurrence records for use in ecological niche models. – Ecography 38: 541–545.

Anderson, R. P. 2013. A framework for using niche models to estimate impacts of climate change on species distributions. – Annals of the New York Academy of Sciences 1297: 8–28.

Barry, S. C. and Welsh, A. 2002. Generalized additive modelling and zero inflated count data. – Ecological Modelling 157: 179–188.

Bocher, E. and Neteler, M. (Eds.) 2012. Geospatial Free and Open Source Software in the 21st Century - Proceedings of the first Open Source Geospatial Research Symposium, OGRS 2009, Nantes, France, 8-10 July, 2009, Lecture Notes in Geoinformation and Cartography. Springer.

Boria, R. A. et al. 2014. Spatial filtering to reduce sampling bias can improve the performance of ecological niche models. – Ecological Modelling 275: 73–77.

Cressie, N. 1993. Statistics for spatial data. Wiley series in probability and mathematical

statistics: Applied probability and statistics. J. Wiley.

Diggle, P. 2003. Statistical Analysis of Spatial Point Patterns. Mathematics in biology. Arnold.

Elith, J. et al. 2006. Novel methods improve prediction of species' distributions from occurrence data. – Ecography 29: 129–151.

Elith, J. et al. 2010. The art of modelling range-shifting species. – Methods in Ecology and Evolution 1: 330–342.

Elith, J. et al. 2011. A statistical explanation of MaxEnt for ecologists. – Diversity and Distributions 17: 43–57.

Fithian, W. et al. 2015. Bias correction in species distribution models: pooling survey and collection data for multiple species. – Methods in Ecology and Evolution 6: 424–438.

Fithian, W. and Hastie, T. 2013. Finite-sample equivalence in statistical models for presence-only data. – The Annals of Applied Statistics 7: 1917–1939.

Friedman, J. et al. 2010. Regularization paths for generalized linear models via coordinate descent. – Journal of Statistical Software 33: 1–22.

Guillera-Arroita, G. et al. 2014. Maxent is not a presenceabsence method: a comment on Thibaud et al. – Methods in Ecology and Evolution 5: 1192–1197.

Manly, B. et al. 2002. Resource Selection by Animals: Statistical Design and Analysis for Field Studies, 2nd Edition. New York: Kluwer Press.

Merow, C. et al. 2016. Improving niche and range estimates with maxent and point process models by integrating spatially explicit information. – Global Ecology and Biogeography 25: 1022–1036.

Merow, C. et al. 2013. A practical guide to maxent for modeling species distributions: what it does, and why inputs and settings matter. – Ecography 36: 1058–1069.

Muscarella, R. et al. 2014. ENMeval: An R package for conducting spatially independent evaluations and estimating optimal model complexity for maxent ecological niche models. – Methods in Ecology and Evolution 5: 1198–1205.

Phillips, S. J. et al. 2006. Maximum entropy modeling of species geographic distributions.

– Ecological Modelling 190: 231–259.

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

Phillips, S. J. et al. 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. – Ecological Applications 19: 181–197.

Phillips, S. J. et al. 2004. A maximum entropy approach to species distribution modeling. In Proceedings of the Twenty-First International Conference on Machine Learning, pp. 472–486. New York: ACM Press.

Radosavljevic, A. and Anderson, R. P. 2014. Making better maxent models of species distributions: complexity, overfitting and evaluation. – Journal of Biogeography 41: 629–643.

Reddy, S. and Dávalos, L. M. 2003. Geographical sampling bias and its implications for conservation priorities in Africa. – Journal of Biogeography 30: 1719–1727.

Renner, I. W. et al. 2015. Point process models for presence-only analysis. – Methods in Ecology and Evolution 6: 366–379.

Renner, I. W. and Warton, D. I. 2013. Equivalence of MAXENT and Poisson point process models for species distribution modeling in ecology. – Biometrics 69: 274–281.

Soberón, J. and Peterson, A. T. 2004. Biodiversity informatics: managing and applying primary biodiversity data. Philosophical Transactions of the Royal Society of London B 359: 689–698.

Syfert, M. M. et al. 2013. The effects of sampling bias and model complexity on the predictive performance of MaxEnt species distribution models. – PLoS ONE 8.

VanDerWal, J. et al. 2009. Abundance and the environmental niche: Environmental suitability estimated from niche models predicts the upper limit of local abundance. – The American Naturalist 174: 282–291.

Warton, D. and Shepherd, L. 2010. Poisson point process models solve the “pseudo-absence problem” for presence-only data in ecology. – Annals of Applied Statistics 4: 1383–1402.

Weber, M. et al. 2016. Is there a correlation between abundance and environmental suitability derived from ecological niche modelling? a meta-analysis. – Ecography. to appear.

Wenger, S. J. and Freeman, M. C. 2008. Estimating species occurrence, abundance, and detection probability using zero-inflated distributions. – Ecology 89: 2953–2959.

Wolkovich, E. M. et al. 2012. Advances in global change research require open science by individual researchers. – Global Change Biology 18: 2102–2110.

Table 1. Comparison of performance of Maxent models with varying choices of feature classes and output transforms, for a reference data set of occurrence records of 226 species, and presence/absence data for model evaluation (Elith et al. 2006). Maxent feature classes are abbreviated as “l” (linear), “q” (quadratic), “p” (product), “t” (threshold) and “h” (hinge). Results are shown for analyses run with random background pixels, as well as for those implementing a target-group background (Phillips et al. 2009). The AUC statistic measures area under the receiver operating characteristic curve, while COR measures the correlation between model output and 0/1 data representing observed absence/presence.

Feature Classes	Output Scaling	Study with these defaults	Random background		Target-group background	
			AUC	COR	AUC	COR
Lqpt	Cumulative	Elith et al. 2006	0.7220	0.1989	0.7534	0.2368
Lqpth	Logistic	Phillips and Dudík 2008	0.7282	0.2110	0.7575	0.2447
Lqph	Raw		0.7296	0.1855	0.7593	0.2404
Lqph	Logistic		0.7296	0.2125	0.7593	0.2465
Lqph	Cloglog	Present paper: Maxent	0.7296	0.2120	0.7593	0.2479
Lqph	Cloglog	Present paper: maxnet	0.7271	0.2100	0.7587	0.2490