Limits of uncertainty about estimates of probability of ecological events

Petr Keil *

Dept. of Ecology & Evolutionary Biology, Yale University, 165 Prospect Street, New Haven, 06511 Connecticut, USA

Center for Theoretical Study, Charles University and Academy of Sciences of the Czech Republic, Jilska 1, Praha 1, Czech Republic

Abstract

Probability (P) of binomial event is a commonly estimated quantity in ecology. Recently, interest has moved to estimation and communication of the associated uncertainty about the estimates of P. Here I use the principle of maximum entropy to introduce truncated exponential probability density function $f(P)$ on a closed interval [0,1] that gives expectation of the uncertainty, given that the only information we have is a single-number estimate P_{single}, which I assume to represent mean μ of an unknown probability density distribution of P. This expectation puts an upper bound on the maximum uncertainty about P. I also present the associated cumulative distribution function, quantile function, and random number generator. I demonstrate the MaxEnt $f(P)$ on a species distribution model predicting probability of a species' occurrence on a geographic map. The MaxEnt $f(P)$ presented here

*E-mail: pkeil@seznam.cz

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.446v1 | CC-BY 4.0 Open Access | received: 21 Jul 2014, published: 21 Jul

can be used to make conservative probabilistic statements about probability statements, and it can be used as an alternative to beta distribution, and as the least informative prior distribution of P in Bayesian modelling.

Keywords: Cumulative distribution function, exponential, MaxEnt, maximum likelihood, MCMC, species distribution model, quantile function.

Introduction

Ecologists are often interested in probability P of binomial event, where the event is, for instance, presence of a species, colonization or extinction event, or survival of an individual. The P is usually estimated as a parameter of binomial probability density function in statistical models. Examples of such models popular in ecology are generalized linear models for binomial or proportion data [17, 3], or more complex hierarchical models that also incorporate (or estimate) probability of detection of the event [4, 21].

With the upsurge of Bayesian modelling ecologists have started to acknowledge the *uncertainty* about parameter estimates, where the uncertainty is expressed as probability density of the parameter values. If P is the estimated parameter, then it can also be assigned its own probability density. This density is usually estimated as posterior conditional density $p(P|data, model)$ by Markov Chain Monte Carlo (MCMC) [19] or by Laplace approximation [22] algorithms.

In many cases the full posterior probability density of P is not available, e.g. when we fit the model by likelihood maximization, and all we have is a single number representing P, which I will hereafter call $\boldsymbol{P_{single}}$. Sometimes P_{single} actually does not come from a formal model at all, or the model that produced P_{single} is unknown, incomprehensible, or P_{single} is an average coming from an ensemble of models [2]. For example, while reading a scientific paper, one can encounter a statement that *"we estimated the probability of the species being present at the locality to be 0.67"*, and no other information is provided on how confident are the authors about such claim. In my field, which is geographical ecology, probabilities of a species occurrence are often mapped to geographical space, conditional on suitable climatic conditions. It is very rare to see maps of uncertainty about the estimated probabilities (but see [10, 7]). Yet it would perhaps be useful to have a way to associate P_{single} with some magnitude of uncertainty, or to bound the possible magnitude uncertainty, even when all we have is just the P_{single}.

In this paper I propose that the P_{single} can be assumed to represent mean μ of an unknown probability density function $f(P)$, and I leave the judgement about appropriateness of this assumption solely up to the reader – yet, this assumption is at the very foundation of my reasoning, and any practical application of the methods presented here will depend on the assumption. I propose that $f(P)$ has to satisfy the following conditions (constraints): (i) $f(P)$ is continuous (ii) μ is known, (iii) possible values of P are bounded between 0 and 1, and (iv) $\int_0^1 f(P)\,\mathrm{d}P = 1$ (the normalization condition). The probability density function $f(P)$ that satisfies these constraints, and at the same time represents our ignorance about all of the other properties of $f(P)$, is the $f(P)$ that gives the maximum value of entropy (H) [8] defined as:

$$H = \int_0^1 f(P) \log f(P)\,\mathrm{d}P \tag{1}$$

In the next section I introduce such function, and I call it *MaxEnt* $f(P)$. I will show that it has properties that enable to put an upper bound on the magnitude of uncertainty about P, given that we only have P_{single}.

Complete raw data and codes (with detailed comments) used for this study are provided in Supplements S1 and S2. The latter also provides all of the raw figures and the LaTeX code of the manuscript.

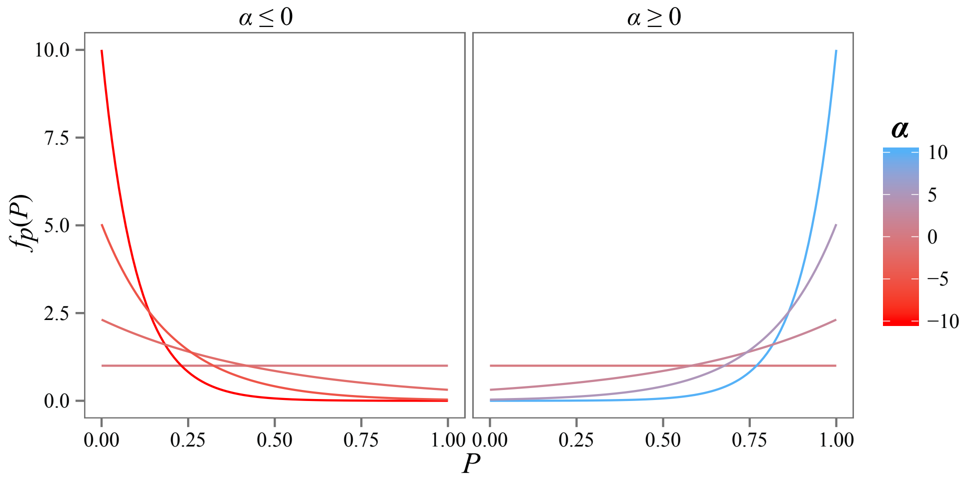

Figure 1: Shapes of the *MaxEnt* $f(P)$ given by eq. 4 with different values of parameter α. If $\alpha = 0$ then $f(P) = 1$ everywhere between 0 and 1 (uniform distribution); otherwise the function is a truncated exponential.

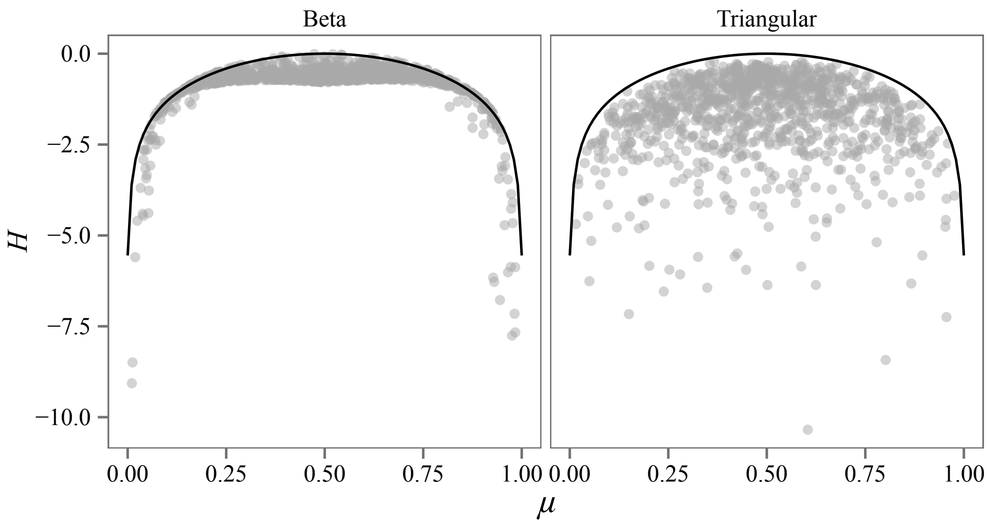

Figure 2: Relationships between mean probability density μ and entropy H of beta and triangular probability density functions. Solid black lines are the *MaxEnt* $f(P)$ described by eq. 4. Grey dots are 1000 beta (left) and triangular (right) density functions with randomly simulated parameter combinations. In the case of beta $f(P)$ the two shape parameters were drawn independently from $Uniform(0.1, 1)$. The triangular $f(P)$ has parameters A, B and C, where A and B define the interval at which $f_p(P) > 0$, and C is the peak of $f(P)$. A and B were chosen from $Uniform(0, 1)$ and C from $Uniform(A, B)$.

The *MaxEnt* $f(P)$

Conrad [5] provides a proof that the general form of maximum entropy probability density function $f(x)$ for any x on the interval $[a, b]$ with mean μ is a truncated exponential function:

$$f(x) = \frac{\alpha e^{\alpha x}}{e^{\alpha b} - e^{\alpha a}}, x \in [a, b] \tag{2}$$

where the mean μ is given by

$$\mu = \frac{\int_a^b \alpha P e^{\alpha P} \, \mathrm{d}P}{e^{\alpha b} - e^{\alpha a}} = \frac{b e^{\alpha b} - a e^{\alpha a}}{e^{\alpha b} - e^{\alpha a}} - \frac{1}{\alpha}. \tag{3}$$

As probability is defined between 0 and 1, I simplified eqs. 2 and 3 by setting $a = 0$ and $b = 1$ in order to get the $f(P)$. When using P instead of x, we get:

$$f(P) = \frac{\alpha e^{\alpha P}}{e^{\alpha} - 1}, P \in [0, 1] \tag{4}$$

and

$$\mu = \frac{e^{\alpha}}{e^{\alpha} - 1} - \frac{1}{\alpha}. \tag{5}$$

The function in eq. 4 is the *MaxEnt* $f(P)$. Fig. 1 illustrates how the shape of $f(P)$ varies with parameter α. The value of α associated with a given value of μ can be found using eq. 5 and numerical optimization (see the Supplement S1). I propose that $-700 < \alpha < 700$ or $0.001 < P < 0.999$ is a reasonable range for applied ecological purposes; more extreme values only complicate the optimization. Fig. 2 shows how the entropy of $f(P)$ compares with entropies of beta and triangular probability density functions. It is clear that, contrary to some opinions [1], our maximum entropy $f(P)$ satisfying the μ, $a = 0$ and $b = 1$ constraints is not beta distribution (Fig. 2).

I derived the *cumulative distribution function* $F(p < P)$ to be:

$$F(p < P) = \int_0^P f(\pi \to p)\,\mathrm{d}\pi = \frac{e^{\alpha P} - 1}{e^{\alpha} - 1}. \tag{6}$$

I was unable to come up with a closed-form solution of the inverse of $F(p < P)$, which is the *quantile function* (F^{-}) necessary for the estimation of the quantiles (i.e. uncertainty) about P, given μ. Hence, I have calculated the inverse numerically (see the Supplement S1). Finally, to *generate a random number* P^* from $f(P)$ we can use the inverse transform sampling: $P^* = F^{-}(U)$ where $U \sim Uniform(0, 1)$.

I provide R [20] code for the probability density function, cumulative distribution function, quantile function, random number generator and the procedures to switch between α and μ in the Supplements S1, and a more detailed version in Supplement S2.

Uncertainty from *MaxEnt* $f(P)$

In ecology, uncertainty about estimates of P has been expressed in a variety of ways. The classical measure of uncertainty is the span of the 95% confidence interval [6, 9]; Newcombe [18] reviews several ways to calculate it for proportions between 0 and 1. Alternatives are standard deviation of posterior density [21, 12] classification of uncertainty into classes [15], or weighted indices of uncertainty [10]. Marcot [16] reviews some other uncertainty measures applicable for posterior $p(P)$ in Bayesian modelling such as 95% credible interval width, posterior probability certainty index or certainty envelope.

In this paper use the word *uncertainty* to refer an to an inter-quantile range (or span) of (1) posterior density of samples coming from MCMC sampling, or (2) probability density function *MaxEnt* $f(P)$ (Figs. 3, 5, 6), where the latter is calculated by the quantile function provided in the Supplements S1 and S2. Although these two things are not equivalent, they relate to the

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.446v1 | CC-BY 4.0 Open Access | received: 21 Jul 2014, published: 21 Jul

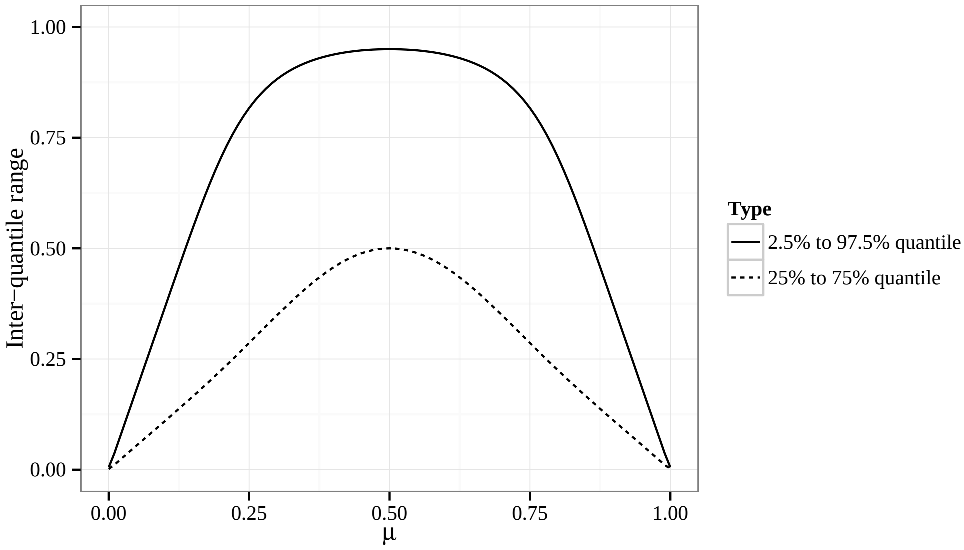

Figure 3: Relationship between the inter-quantile range and mean μ of the *MaxEnt* $f(P)$ given by eq. 4. Inter-quantile range is a commonly used measure of uncertainty in parameter estimates in statistical ecology.

same idea (sensu Plato) of uncertainty.

The *MaxEnt* $f(P)$ gives a hump shaped relationship between uncertainty (the inter-quantile range) and μ. The range is very broad for any $\mu \approx 0.5$, but it quickly decreases as μ approaches 0 or 1. This has fundamental implications: it means that, under the assumption that $P_{single} = \mu$, any reported $P_{single} > 0.85$ or $P_{single} < 0.15$ automatically brings low[1] uncertainty about the P value. In other words, it is impossible to say that something has high (or low) probability, and at the same time be uncertain about such statement; in contrast, statement that $P_{single} \approx 0.5$ allows for uncertainty.

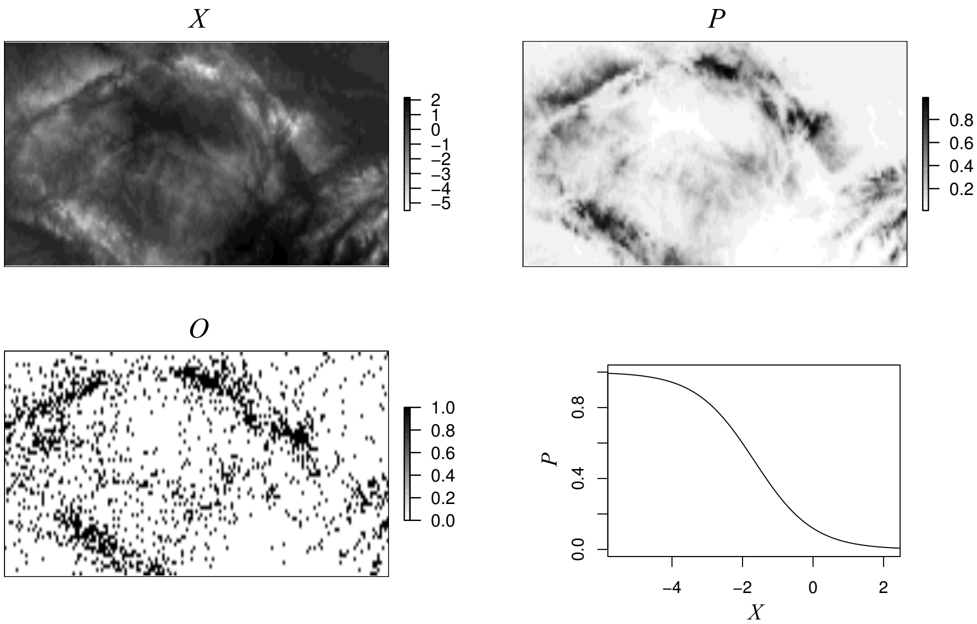

Figure 4: Geographic distribution of a predictor X, probability P, and the binary outcomes O. The sigmoidal function describes the relationship between X and P (eq. 7). For exact geographic coordinates of the maps see Fig. 6.

Example: *MaxEnt* $f(P)$ in species distribution modelling

To show how *MaxEnt* $f(P)$ could be used in ecological modelling I created a dataset consisting of a predictor vector X, which is standardized mean annual temperature extracted from [11], and aggregated over 5×5 km grid in the Czech Republic (Fig. 4). The predictor X consists of 9943 grid cells (elements) indexed by integer i where $i \in [1, 9943]$. I modelled the true probability of species occurrence P_i in each grid cell i as a deterministic sigmoidal function of X_i:

$$P_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_i)}} \tag{7}$$

[1]'Low' and 'high' are subjective terms, and depend on arbitrarily selected reference, such as a level of significance (α). They should be understood in a relative sense.

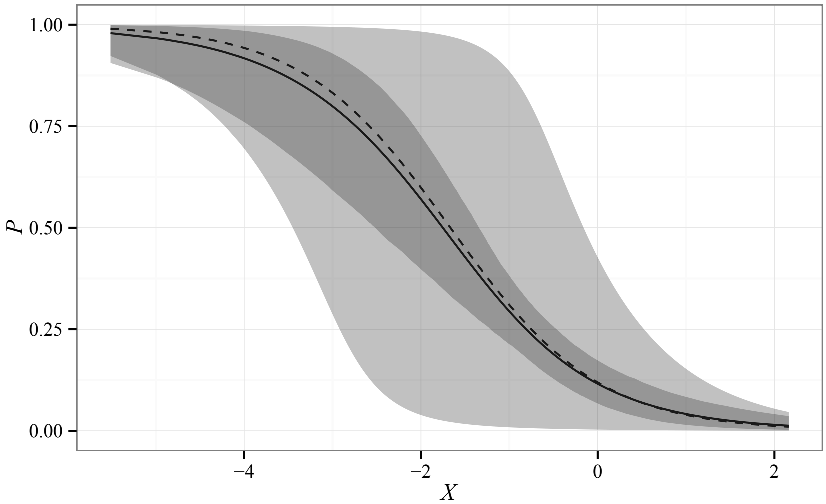

Figure 5: Relationship between predictor X and probability of species' occurrence P, where the dashed line is the "true" relationship (eq. 7), the solid line is the μ estimated by MCMC sampling, dark grey area is a 95% credible interval of μ coming from the MCMC sampling, and light grey area is the 95% inter-quantile range estimated by *MaxEnt* $f(P)$. The same information is visualized in Fig. 6 in a spatially-explicit way.

I set $\beta_0 = -2$ and $\beta_1 = -1.2$. Further, I modelled the actual realized occurrences O_i of the species as a Bernoulli-distributed random variable:

$$O_i \sim Bernoulli(P_i). \tag{8}$$

Figure 4 shows the geographical distribution of the predictor X, the probability of occurrence P, and the outcomes O in each grid cell i.

I then randomly sampled 200 grid cells in the map, assuming that this sample represents a typical ecological data of predictor X_i and response O_i, where $i \in [1, 200]$. I did this sub-sampling in order to invoke some extra uncertainty about P caused by the small sample size. I then used these data and eqs. 7 and 8 to estimate posterior probability densities of β_0, β_1 and P_i. I used Markov Chain Monte Carlo (MCMC) sampler in JAGS [19](3 chains, 2000 iterations as burn-in, 2000 samples for inference in each chain) to numerically estimate the full posteriors. I estimated μ_i as the mean of the posterior distribution in each grid cell i, together with 95% MCMC quantiles

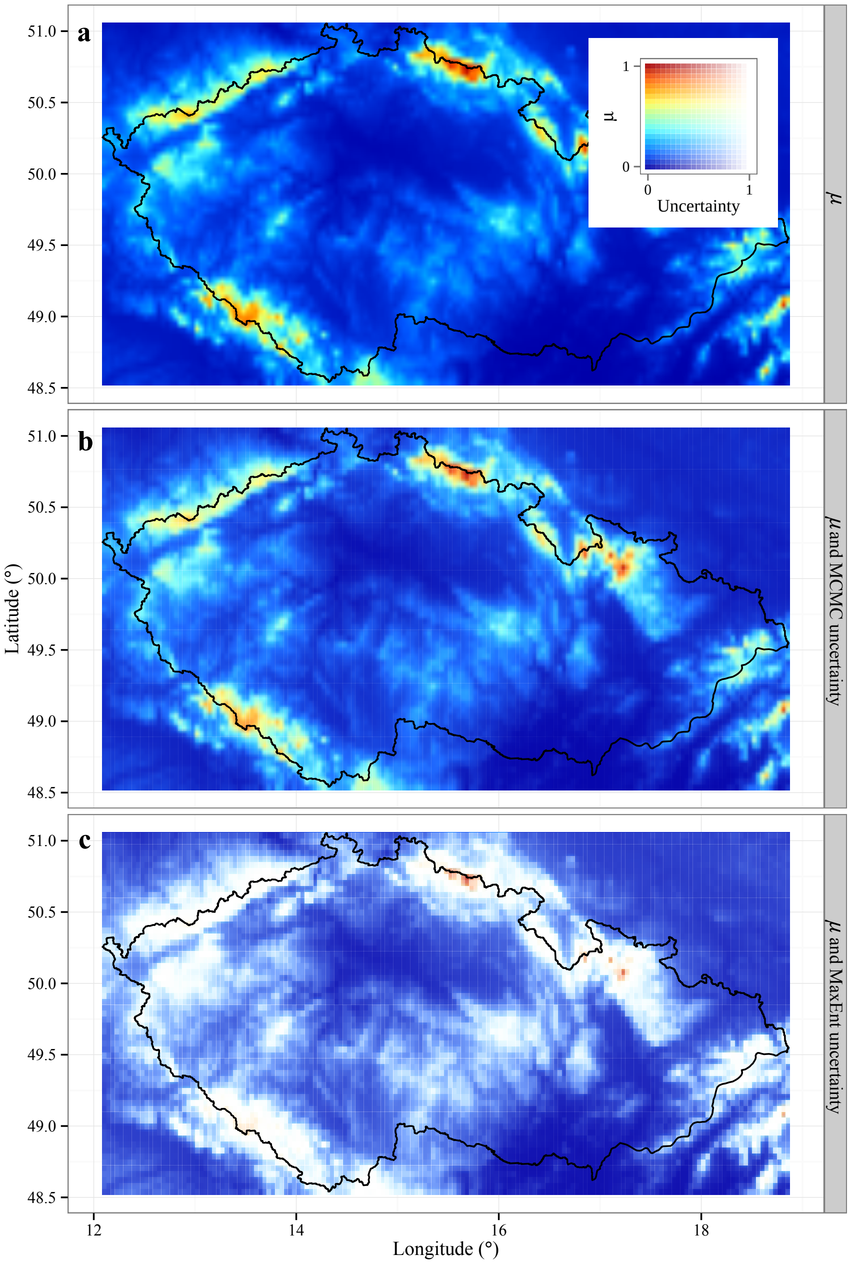

Figure 6: Estimated probabilities of occurrence of a species (colour gradient), together with the uncertainty about the probabilities (transparency gradient). The three maps show the same information as Fig. 5, but in a spatially-explicit way. Panel (a) is μ estimated by MCMC, with zero uncertainty. Panel (b) shows μ values together with the uncertainty about P, which is the 95% credible interval of P estimated by MCMC. Panel (c) shows μ together with the MaxEnt uncertainty, which is MaxEnt 95% inter-quantile range given by $f(P)$.

of the density of P. Finally, I used the *MaxEnt* $f(P)$ and the estimated μ_i values to calculate 95% MaxEnt quantiles.

The difference between the MCMC 95% quantiles and 95% MaxEnt quantiles is illustrated in Figure 5. The range between the MCMC quantiles (dark grey) is much narrower than the range between the MaxEnt quantiles – the reason is that the MCMC posterior distributions of P are constrained by the information (the signal) in the data, whilst the MaxEnt quantiles are only constrained by μ.

Visualizing uncertainty in two-dimensional maps has always been a challenge. Here I have chosen the solution of Golding [7] who mapped the estimates of P as a colour gradient, while he expressed the uncertainty as saturation of the colour. While this may be visually confusing when two unbounded continuous variables are mapped, it works well with two bounded qualities on the $[0, 1]$ interval, where colour saturation (or transparency) is a visual metaphor for uncertainty.

Putting both μ and the uncertainty on a map (Fig. 6) we can see that the *MaxEnt* $f(P)$ eliminated all of the intermediate colours (orange, yellow and green), retaining only the extremes close to 0 and 1. Hence, it effectively divided the map into three classes of model predictions: Species highly likely to be present (red), species highly unlikely to be present (blue), and areas where any probabilistic statement about the species presence is uncertain (white). This classification can be useful in situations when we want to be really conservative and careful in our inference. Also, it can serve as a basis for development of new probability thresholding techniques [14].

General utility of the *MaxEnt* $f(P)$

I have shown that, given only P_{single} and the assumption that $P_{single} = \mu$, our maximum uncertainty about the probability density of P is bounded and decreasing as μ approaches 0 or 1. If we assume that a P_{single} reported in scientific literature represents μ, then we can make statements about the latent distribution $f(P)$, even under total ignorance about anything but P_{single}. Specifically:

- We can use $f(P)$ to calculate the 95% quantiles (credible intervals) of P, given $P_{single} = \mu$. For example, we can encounter a statement that "*probability of the species being present at a given locality on our map is 0.87*". This information is sufficient to calculate the MaxEnt 95% quantiles, which are 0.521 and 0.997 for $P_{single} = 0.87$. These give the most conservative span of 95% credible interval of P. This is the maximum uncertainty we are able to get, given the assumptions above, and hence any other extra information that we will bring into the estimation of probability density of P will only reduce the uncertainty.

- We can use $f(P)$ to make conservative probabilistic statements about probabilistic statements, and we can also set arbitrary probability threshold. Using the example above, we can state that "*the true probability of the species' presence at the locality is higher than 0.5*" and, given $P_{single} = \mu = 0.87$, the probability of such statement being false is smaller than 0.021 (eq. 6).

- The span of 95% quantiles of *MaxEnt* $f(P)$ is very broad for any $\mu \approx 0.5$, but it quickly decreases as μ approaches 0 or 1. Hence, any reported $P_{single} > 0.85$ or $P_{single} < 0.15$ automatically implies relatively low uncertainty about a statement – probability cannot be high (or low) and uncertain at the same time.

- Perhaps most importantly, the $f(P)$ can be used as the least informative prior distribution for probability in every instance where we can reasonably assume that $P_{single} = \mu$. Johnson ([13], p. 236) argues that there is a poor theoretical basis for using beta distribution is the best and "natural" non-informative prior distribution of P. He then cites number of studies that attempted to find the least informative combinations of the two parameters of beta distribution. Interestingly, none

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.446v1 | CC-BY 4.0 Open Access | received: 21 Jul 2014, published: 21 Jul

of the authors steps out of the realm of beta distribution. Here I show (Fig. 2) that beta is definitely not the least informative distribution when its mean is known (μ); in such situation the *MaxEnt* $f(P)$ can be used as the un-informative prior.

Loose ends

In practice, the *MaxEnt* $f(P)$ can be excessively broad. However, if P_{single} represents an estimate, then there is probably sufficient information that the distribution will be unimodal about the estimate. This can potentially be incorporated as another constraint in the derivation of the MaxEnt density function, and will likely lead to lower uncertainty about P, and to improved applicability of the approach.

Also, what remains to be thought through is: Can we estimate the uncertainty about the estimate of uncertainty of P? Or: can we estimate the uncertainty about the estimate of uncertainty about the estimate of uncertainty of P? The approach presented here can only be relied on if the 'nested' uncertainties converge to a point.

Acknowledgements

I am grateful to Arnošt L. Šizling, Bob O'Hara, Walter Jetz, Xiao Xiao and Adam M. Wilson for suggestions, comments and discussions on the subject.

References

[1] Whereof one cannot speak: When input distributions are unknown, http://goo.gl/a8NstO.

[2] M. B. Araujo and M. New. Ensemble forecasting of species distributions. *Trends in Ecology & Evolution*, 22:42–47, 2007.

[3] B. M. Bolker, M. E. Brooks, C. J. Clark, S. W. Geange, J. R. Poulsen, M. H. H. Stevens, and J.-S. S. White. Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution*, 24:127–135, 2009.

[4] J. S. Clark. *Models for ecological data: An Introduction.* Princeton University Press, Princeton, 2007.

[5] K. Conrad. Probability distributions and maximum entropy, http://goo.gl/sSAZH.

[6] J. Elith, M. A. Burgman, and H. M. Regan. Mapping epistemic uncertainties and vague concepts in predictions of species distribution. *Ecological Modelling*, 157:313–329, 2002.

[7] N. Golding. *Mapping and understanding the distributions of potential vector mosquitoes in the UK: New methods and applications.* PhD thesis, University of Oxford, Oxford, UK, 2013.

[8] J. Harte. *Maximum entropy and ecology: A theory of abundance, distribution, and energetics.* Oxford University Press, Oxford, 2011.

[9] S. Hartley, R. Harris, and P. J. Lester. Quantifying uncertainty in the potential distribution of an invasive species: Climate and the argentine ant. *Ecology Letters*, 9:1068–1079, 2006.

[10] S. I. Hay, C. A. Guerra, P. W. Gething, A. P. Patil, A. J. Tatem, A. M. Noor, C. W. Kabaria, B. H. Manh, I. R. F. Elyazar, S. Brooker, D. L. Smith, R. A. Moyeed, and R. W. Snow. A world malaria map: Plasmodium falciparum endemicity in 2007. *PLoS Med*, 6:e1000048, 2009.

[11] R. J. Hijmans, S. E. Cameron, J. L. Parra, P. G. Jones, and A. Jarvis. Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, 25:1965–1978, 2005.

[12] I. Ibanez, J. A. Silander, A. M. Wilson, N. LaFleur, N. Tanaka, and I. Tsuyama. Multivariate forecasts of potential distributions of invasive plant species. *Ecological Applications*, 19:359–375, 2009.

[13] N. L. Johnson. *Continuous Univariate Distributions.* Wiley & Sons, New York, 1994.

[14] C. Liu, P. M. Berry, T. P. Dawson, and R. G. Pearson. Selecting thresholds of occurrence in the prediction of species distributions. *Ecography*, 28:385393, 2005.

[15] M. Luoto, M. Marmion, and J. Hjort. Assessing spatial uncertainty in predictive geomorphological mapping: A multi-modelling approach. *Computers & Geosciences*, 36:355–361, 2010.

[16] B. G. Marcot. Metrics for evaluating performance and uncertainty of bayesian network models. *Ecological Modelling*, 230:50–62, 2012.

[17] S. Menard. *Applied logistic regression.* Sage Publications, Thousand Oaks, 2002.

[18] R. G. Newcombe. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine*, 17(8):857–872, 1998.

[19] M. Plummer. Jags: A program for analysis of bayesian graphical models using gibbs sampling. In *roceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*, Vienna, 2003.

[20] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2014.

[21] J. A. Royle and R. M. Dorazio. *Hierarchical modeling and inference in ecology: the analysis of data from populations, metapopulations and communities.* Academic Press, Amsterdam, 2008.

[22] H. Rue, S. Martino, and N. Chopin. Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71:319–392, 2009.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.446v1 | CC-BY 4.0 Open Access | received: 21 Jul 2014, published: 21 Jul