The following supplement accompanies the article

Importance of machine learning for enhancing ecological studies using information-rich imagery

Antoine M. Dujon*, Gail Schofield*

*Corresponding authors: antoine.dujon@yahoo.fr; gail.schofield@qmul.ac.uk

Endangered Species Research 39: 91–104 (2019)

Supplement 2.

Supplementary Methods: Details of the Bayesian statistical analyses.

To investigate whether the size of the regions of interest influenced the probability (P) that machine learning was used in a publication, we fitted a series of logistic regression models to the data using a linear logistic function of equation (Kruschke 2015):

$$Ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 * size$$

where size in metres is β_0 the intercept and β_1 is the slope of the logistic function. Once a model fitted the response probability P was calculated as:

$$P = \frac{e^{(\beta_0 + \beta_1 * size)}}{1 + e^{(\beta_0 + \beta_1 * size)}}$$

To investigate potential differences between individual body size, individual plant size, feature cluster size and nest size and the probability machine learning being used, we started by separately fitting a model to each grouping. Then, we fitted an additional three models by cumulatively adding each group to the explanatory variable of the model (i.e. the first model only included body size as the explanatory variable, the second model included body size and plant size, the third model included body size, plant size and feature size and the fourth model included all four groups). We used Bayesian statistics to calculate the confidence intervals for small sample sizes (n < 40) by specifying an appropriate prior distribution (Brown et al. 2001). We used Jeffreys prior (beta distribution with □ = 0.5 and ß = 0.5) to compute the confidence interval for proportion estimates. This type of prior is used to obtain more accurate confidence intervals than frequentist methods, when the estimated proportion is close to zero or one and when the sample size is small (n < 40), but also without losing accuracy for intermediate values (see Agresti & Min 2005).

We used uninformative priors to estimate the coefficients of the logistic regression models (normal distribution of mean = 0 and standard deviation = 10) and to estimate the 95% confidence intervals of mean impact factors (normal distribution of mean = 0 and standard deviation = 2). Uninformative priors do not make any assumption on the underlying distribution of the estimated parameter (see Kruschke 2015).

Two proportions or two means were considered significantly different if their 95 % credible interval did not overlap. The coefficients of the logistic regression models were considered to be statistically significant if their 95 % credible interval did not include 0. Throughout this manuscript, we report all of the estimated parameters as 'parameters estimates' [95 % confidence interval].

All Bayesian statistical analyses were performed using the RStan package (Stan Development Team 2018) within the R software version 3.3.2. (R Development Core Team 2013).

References

Agresti A, Min Y (2005) Frequentist performance of Bayesian confidence intervals for comparing proportions in 2 × 2 contingency tables. Biometrics 61:515–523

Brown LD, Cai TT, DasGupta A (2001) Interval estimation for a binomial proportion. Stat Sci 16:101–133

Kruschke JK (2015) Doing bayesian data analysis: a tutorial with R, JAGS, and Stan, Second Edi. Academic Press, London, United Kingdom

Stan Development Team. 2018. RStan: the R interface to Stan. Available from http://mc-stan.org/

Supplementary Results: Results of the Bayesian logistic regression models investigating potential relationships between the probability that machine learning was used in a publication and the size of the four regions of interest.

Table S1. Estimates of the intercept (β_0) and the slope (β_1) parameters for the seven logistic regression models investigating a potential relationship between the size of the region of interests and the probability machine learning being used to detect them. Statistically significant intercept and slope estimates are indicated by an *. A slope close to zero indicates that the size of the region of interest had a minimal effect on the probability machine learning being used to detect it.

Size variable	**Intercept $\boldsymbol{\beta_0}$**	**Slope $\boldsymbol{\beta_1}$**	**Number of size measurements**
Body size	-1.65 [-2.2,-1.08]*	-0.15 [-0.43,0.04]	165
Plant size	0.41 [-0.17,0.10]	0.01 [0.00,0.02]	81
Feature size	0.77 [-0.18,1.75]	0.03 [-0.07,0.15]	34
Nest size	2.32 [-1.11,6.51]	-8.01 [-18.71,-1.41]*	10
Body size and plant size	-1.19 [-1.52,-0.88]*	0.02 [0.01,0.04]*	246
Body size, plant size and feature size	-0.89 [-1.18,-0.61]*	0.02 [0.01,0.03]*	280
Body size, plant size, feature size and nest size	-0.91 [-1.19, -0.63]*	0.02 [0.01,0.03]*	290