Accounting for Imperfect Detection and Survey Bias in Statistical Analysis of Presence-only Data

Robert M. Dorazio[1]

[1]Southeast Ecological Science Center, U.S. Geological Survey, Gainesville, FL 32653

2014 Graybill/ENVR Conference
Fort Collins, Colorado
08 Sep 2014

Species Distribution Models (SDMs)

Definition: A SDM expresses a functional relationship between the occurrence or abundance of a species and one or more aspects of its environment

Uses: Many!

- Predicting the geographic distribution of a species over its potential range
- Predicting consequences of management actions (e.g., habitat restoration) on a species' distribution

Limitations: Many!

- No dynamics (animal movements, plant dispersals)
- No interactions within or among species

Estimating SDMs From Planned Surveys

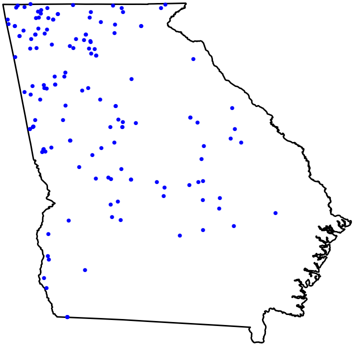
Presence-absence surveys

- Binary-regression modeling
- Occupancy modeling

Abundance surveys

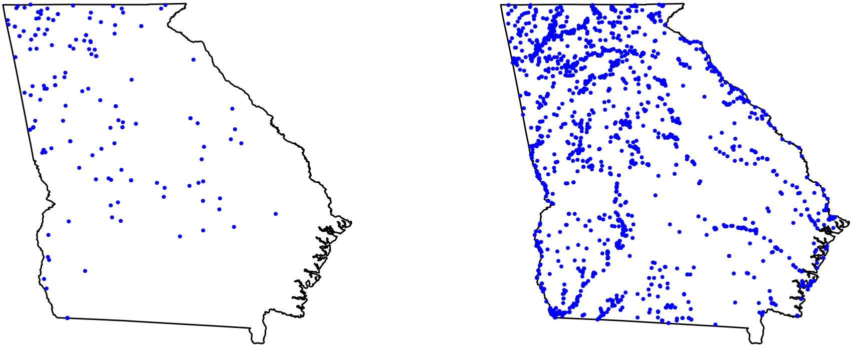
- Poisson-regression modeling
- N-mixture modeling

Georgia, USA

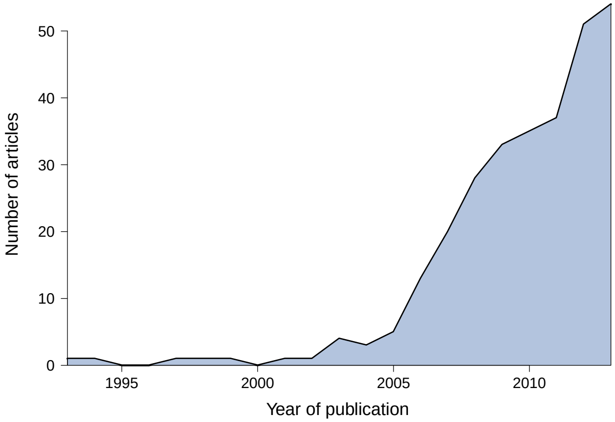


Planned vs. Opportunistic Surveys

Nonindigenous aquatic species in Georgia, USA



Estimating SDMs From Opportunistic Surveys

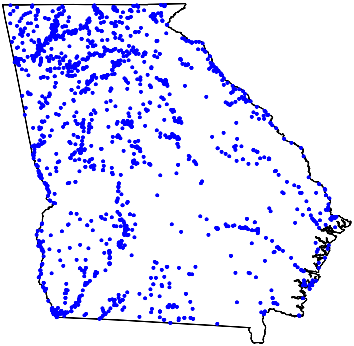


Source: Web of Science using "presence-only data"

Estimating SDMs From Opportunistic Surveys

Presence-background models

- Binary regressions
- Case-augmented binary regressions (Lee et al., 2006; Lele and Keim, 2006)
- Maximum entropy (Phillips et al., 2006; Elith et al., 2010)
- Spatial point processes (Warton and Shepherd, 2010)



Poisson Process as a SDM

Conceptual unification

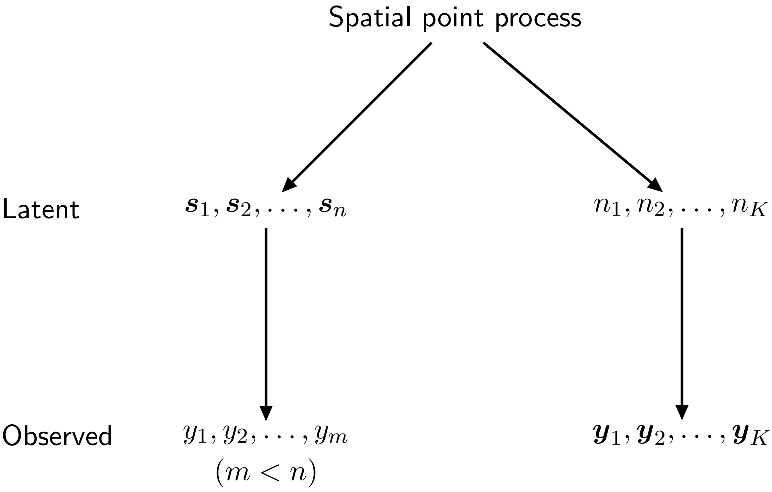
- asymptotic equivalence of estimators
  - CA-binary regressions modifed for spatial resolution (Dorazio, 2012)
  - Maxent models (Renner and Warton, 2013)
- parameters are invariant to spatial scale

Potential sources of bias

- imperfect detectability (Dorazio, 2012; Lahoz-Monfort et al., 2014)
- opportunistic sampling (Phillips et al., 2009; Yackulic et al., 2013)

  – location-dependent thinning of process helps in some cases (Chakraborty et al., 2011; Fithian and Hastie, 2013)

Hierarchical Modeling of Opportunistic and Planned Survey Data



Poisson Process as a SDM

Definitions

Spatial domain: $B \subset \mathbb{R}^2$

Individual activity center: $\boldsymbol{s} \in B$

First-order intensity function: $\lambda(\boldsymbol{s}) = \exp(\boldsymbol{\beta}'\boldsymbol{x}(\boldsymbol{s}))$

Assumptions

- $N(B) \sim \text{Poisson}(\mu(B))$, where $\mu(B) = \int_B \lambda(\boldsymbol{s})\, d\boldsymbol{s}$
- $f(\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_n | N(B) = n) = \prod_{i=1}^{n} \lambda(\boldsymbol{s}_i)/\mu(B)$

Latent state variables

- $g(\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_n, n) = \frac{\exp\{-\mu(B)\}}{n!} \prod_{i=1}^{n} \lambda(\boldsymbol{s}_i)$
- $N(C_k) \sim \text{Poisson}(\mu(C_k))$

  where $C_1 \cup \cdots \cup C_K = B$

Detections of Individuals in Opportunistic Surveys

Assumptions

- Each individual is detected independently with probability $p(\boldsymbol{s})$:

  $Y|\boldsymbol{s} \sim \text{Bernoulli}(p(\boldsymbol{s}))$

- $p(\boldsymbol{s})$ depends on an observer's detection ability and choice of survey location:
  $\text{logit}(p(\boldsymbol{s})) = \boldsymbol{\alpha}'\boldsymbol{w}(\boldsymbol{s})$

Observations

- $m =$ number of individuals detected in B
- $(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_m) =$ locations of detected individuals

$$L(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \frac{\exp\{-\nu(B)\}}{m!} \prod_{i=1}^{m} \lambda(\boldsymbol{s}_i)\, p(\boldsymbol{s}_i)$$

where $\nu(B) = \int_B \lambda(\boldsymbol{s}) p(\boldsymbol{s})\, d\boldsymbol{s} = \text{E}(M(B))$

Detections of Individuals in Planned Surveys

Assumptions

- Only individuals whose activity centers lie within sample unit C_k are available to be detected
- Each individual is detected with probability p_{kj} during the jth survey of unit C_k:
  $\text{logit}(p_{kj}) = \boldsymbol{\gamma}'\boldsymbol{v}(C_k)$

Observations (e.g., J_k replicate counts)

- $Y_{kj} | N(C_k) = n_k \sim \text{Binomial}(n_k, p_{kj})$

$$L(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \prod_{k=1}^{K} \sum_{n_k=\max(\boldsymbol{y}_k)}^{\infty} \frac{\exp\{-\mu(C_k)\}\{\mu(C_k)\}^{n_k}}{n_k!} \times \prod_{j=1}^{J_k} \binom{n_k}{y_{kj}} p_{kj}^{y_{kj}} (1 - p_{kj})^{n_k - y_{kj}}$$

Information in Opportunistic Surveys Can Be Limited

$$\log\{L(\boldsymbol{\beta},\boldsymbol{\alpha})\} = -\int_B \lambda(\boldsymbol{s})p(\boldsymbol{s})\,d\boldsymbol{s} + \sum_{i=1}^{m}\log\{\lambda(\boldsymbol{s}_i)\,p(\boldsymbol{s}_i)\}$$

where

$$\lambda(\boldsymbol{s})p(\boldsymbol{s}) = \frac{\exp\{\boldsymbol{\beta}'\boldsymbol{x}(\boldsymbol{s}) + \boldsymbol{\alpha}'\boldsymbol{w}(\boldsymbol{s})\}}{1+\exp\{\boldsymbol{\alpha}'\boldsymbol{w}(\boldsymbol{s})\}}$$

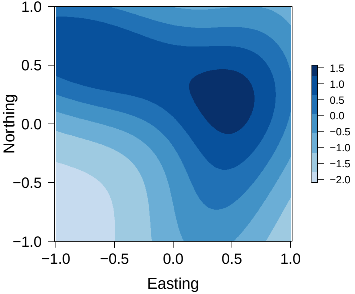
Identifiability problems:

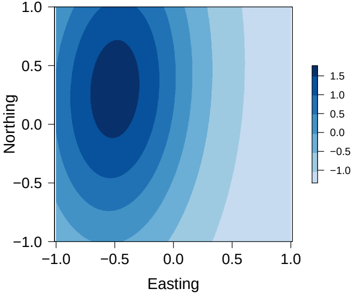
1. If $p(\boldsymbol{s}) = p$, β_0 and α_0 are not identified.
2. If $p(\boldsymbol{s})$ is low $\forall \boldsymbol{s}$, $\lambda(\boldsymbol{s})p(\boldsymbol{s}) \doteq \exp\{\boldsymbol{\beta}'\boldsymbol{x}(\boldsymbol{s}) + \boldsymbol{\alpha}'\boldsymbol{w}(\boldsymbol{s})\}$
   - β_0 and α_0 are not identified
   - other elements of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are not identified if $\boldsymbol{x}$ and $\boldsymbol{w}$ are linearly dependent
3. If Fisher information matrix $\boldsymbol{I}(\boldsymbol{\theta})$ is less than full rank, the parameters in $\boldsymbol{\theta} = (\boldsymbol{\beta}', \boldsymbol{\alpha}')'$ are not identified (Bowden, 1973).

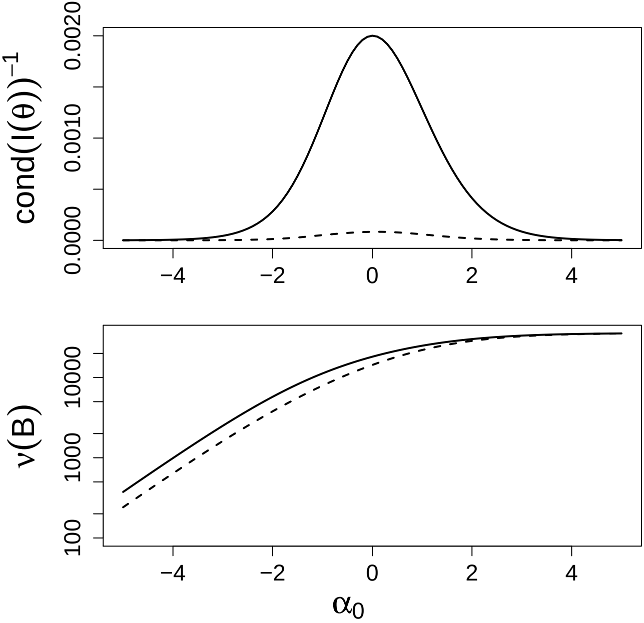
Two models

$\log(\lambda(\boldsymbol{s})) = \log(8000) + 0.5\, x(\boldsymbol{s})$

1. $\text{logit}(p(\boldsymbol{s})) = \alpha_0 - 1.0\, w(\boldsymbol{s})$
2. $\text{logit}(p(\boldsymbol{s})) = \alpha_0 - 1.0\, x(\boldsymbol{s})$






cond(I(θ))⁻¹
0.0000
0.0010
0.0020
−4
−2
0
2
4
ν(B)
100
1000
10000
α₀

Using Planned Surveys To Overcome Limited Information in Opportunistic Surveys

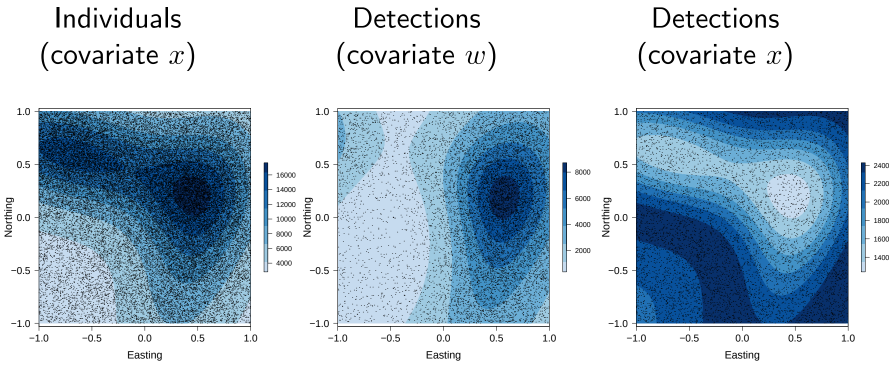
$$L(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) = L(\boldsymbol{\beta}, \boldsymbol{\alpha}) \times L(\boldsymbol{\beta}, \boldsymbol{\gamma})$$

- partition B into sample units
- select K units randomly
- conduct $J > 1$ replicate surveys in each unit

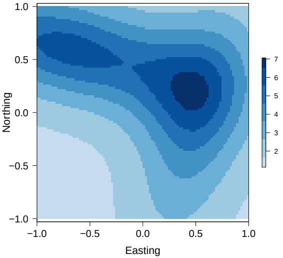
Two models

$\log(\lambda(\boldsymbol{s})) = \log(8000) + 0.5\, x(\boldsymbol{s})$

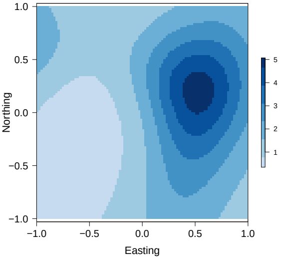
1. $\text{logit}(p(\boldsymbol{s})) = -1.0 - 1.0\, w(\boldsymbol{s})$
   $\text{logit}(p_{kj}) = 0.0 - 1.0\, v(C_k)$ where $v(C) = \int_C w(\boldsymbol{s}) d\boldsymbol{s}$
2. $\text{logit}(p(\boldsymbol{s})) = -1.0 - 1.0\, x(\boldsymbol{s})$
   $\text{logit}(p_{kj}) = 0.0 - 1.0\, v(C_k)$ where $v(C) = \int_C x(\boldsymbol{s}) d\boldsymbol{s}$


Individuals
(covariate x)
Detections
(covariate w)
Detections
(covariate x)
Northing
Easting
1.0
0.5
0.0
−0.5
−1.0
16000
14000
12000
10000
8000
6000
4000
2000
2400
2200
2000
1800
1600
1400

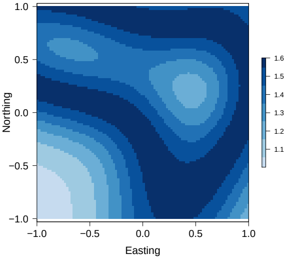
Abundance (covariate x)



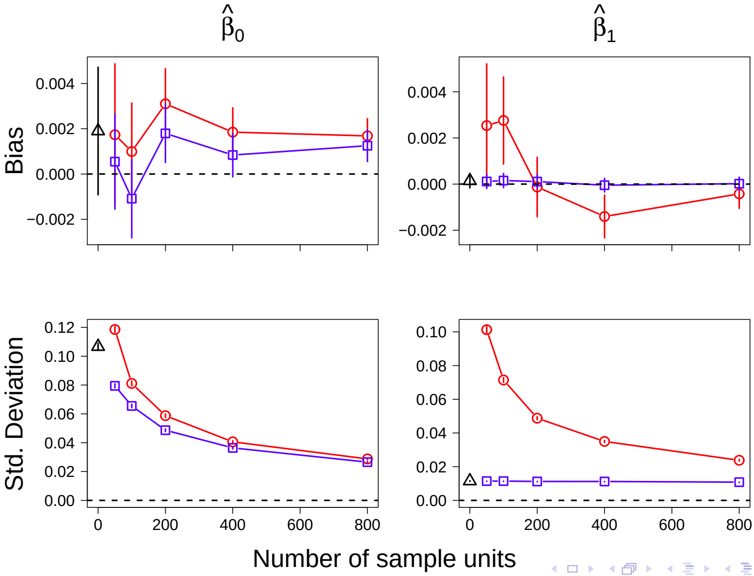
Detections (covariate w)



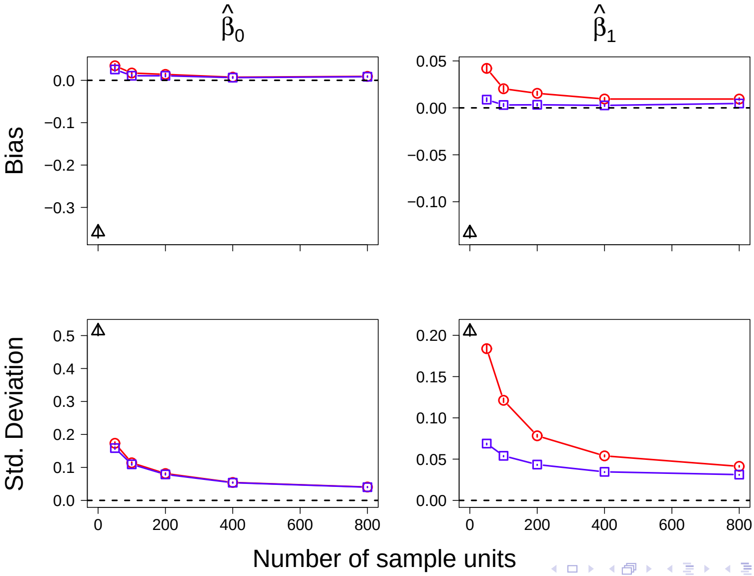
Detections (covariate x)



Simulation Results: Detection covariate w



Simulation Results: Detection covariate x



Summary

1. Bias in estimates of SDMs induced by detection errors or survey bias can be reduced or eliminated using a joint analysis of data collected in opportunistic and planned surveys.
2. This approach is widely applicable because a variety of sampling protocols can be used in planned surveys.
   - double observers
   - removals
   - capture-recapture
   - occupancy (presence-absence sampling with replicates)
3. Spatial point processes are formulated at the level of an individual; therefore, extensions of the Poisson process can be developed to
   - specify effects of biological interactions between individuals
   - predict changes in spatial distribution driven by changes in climate, habitat, non-indigenous species, etc.

Acknowledgments

Funding:

- South Atlantic Landscape Conservation Cooperative (Agreement No. 4500038932)
- USGS, Southeast Ecological Science Center

Reviewers:

- N. Zimmermann
- C. Yackulic

Dorazio, R.M. 2014. Accounting for imperfect detection and survey bias in statistical analysis of presence-only data. Global Ecology and Biogeography (DOI: 10.1111/geb.12216).

Fisher Information Matrix

$$\boldsymbol{I}(\boldsymbol{\beta},\boldsymbol{\alpha}) = \begin{pmatrix} \boldsymbol{I}(\boldsymbol{\beta},\boldsymbol{\beta}) & \boldsymbol{I}(\boldsymbol{\beta},\boldsymbol{\alpha}) \\ \boldsymbol{I}(\boldsymbol{\beta},\boldsymbol{\alpha})' & \boldsymbol{I}(\boldsymbol{\alpha},\boldsymbol{\alpha}) \end{pmatrix}$$

The p, q-th element for each of these submatrices is:

$$\begin{aligned}
\boldsymbol{I}(\beta_p,\beta_q) &= \int_B x_p(\boldsymbol{s})\, x_q(\boldsymbol{s})\, \lambda(\boldsymbol{s})\, p(\boldsymbol{s})\, d\boldsymbol{s} \\
\boldsymbol{I}(\beta_p,\alpha_q) &= \int_B x_p(\boldsymbol{s})\, w_q(\boldsymbol{s})\, \lambda(\boldsymbol{s})\, p(\boldsymbol{s})\{1-p(\boldsymbol{s})\} d\boldsymbol{s} \\
\boldsymbol{I}(\alpha_p,\alpha_q) &= \int_B w_p(\boldsymbol{s})\, w_q(\boldsymbol{s})\, \lambda(\boldsymbol{s})\, p(\boldsymbol{s})\{1-p(\boldsymbol{s})\}^3\, [1-\exp\{2\eta(\boldsymbol{s})\}]\, d\boldsymbol{s} \\
&+ \int_B w_p(\boldsymbol{s})\, w_q(\boldsymbol{s})\, \lambda(\boldsymbol{s})\, p(\boldsymbol{s})^2\{1-p(\boldsymbol{s})\} d\boldsymbol{s}
\end{aligned}$$

where $\eta(\boldsymbol{s}) = \mathrm{logit}\{p(\boldsymbol{s})\}$.

References I

Bowden, R. (1973). The theory of parametric identification. *Econometrica*, 41:1069–1074.

Chakraborty, A., Gelfand, A. E., Wilson, A. M., Latimer, A. M., and Silander, J. A. (2011). Point pattern modelling for degraded presence-only data over large regions. *Applied Statistics*, 60:757–776.

Dorazio, R. (2012). Predicting the geographic distribution of a species from presence-only data subject to detection errors. *Biometrics*, 68:1303–1312.

Elith, J., Phillips, S. J., Hastie, T., Dudik, M., Chee, Y. E., and Yates, C. J. (2010). A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions*, 17:43–57.

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

Lahoz-Monfort, J. J., Guillera-Arroita, G., and Wintle, B. A. (2014). Imperfect detection impacts the performance of species distribution models. *Global Ecology and Biogeography*, 23:504–515.

Lee, A. J., Scott, A. J., and Wild, C. J. (2006). Fitting binary regression models with case-augmented samples. *Biometrika*, 93:385–397.

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

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

References II

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

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.

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

Yackulic, C. B., Chandler, R., Zipkin, E. F., Royle, J. A., Nichols, J. D., Grant, E. H. C., and Veran, S. (2013). Presence-only modelling using MAXENT: when can we trust the inferences? *Methods in Ecology and Evolution*, 4:236–243.