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

Robert M. Dorazio

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

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

Dorazio (USGS)  Presence-only data  08 Sep 2014
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 modified 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

Spatial point process

Latent

$s_1, s_2, \ldots, s_n$

$n_1, n_2, \ldots, n_K$

Observed

$y_1, y_2, \ldots, y_m$

$m < n$

$y_1, y_2, \ldots, y_K$
Poisson Process as a SDM

**Definitions**

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

Individual activity center: $s \in B$

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

**Assumptions**

- $N(B) \sim \text{Poisson}(\mu(B))$, where $\mu(B) = \int_B \lambda(s) \, ds$
- $f(s_1, s_2, \ldots, s_n | N(B) = n) = \prod_{i=1}^n \lambda(s_i) / \mu(B)$

**Latent state variables**

- $g(s_1, s_2, \ldots, s_n, n) = \frac{\exp\{-\mu(B)\}}{n!} \prod_{i=1}^n \lambda(s_i)$
- $N(C_k) \sim \text{Poisson}(\mu(C_k))$

where $C_1 \cup \cdots \cup C_K = B$
Detections of Individuals in Opportunistic Surveys

Assumptions

- Each individual is detected independently with probability $p(s)$:
  \[ Y|s \sim \text{Bernoulli}(p(s)) \]
- $p(s)$ depends on an observer’s detection ability and choice of survey location:
  \[ \logit(p(s)) = \alpha' w(s) \]

Observations

- $m =$ number of individuals detected in $B$
- $(s_1, \ldots, s_m) =$ locations of detected individuals

\[
L(\beta, \alpha) = \frac{\exp\{-\nu(B)\}}{m!} \prod_{i=1}^{m} \lambda(s_i) p(s_i)
\]

where \( \nu(B) = \int_{B} \lambda(s)p(s) \, ds = E(M(B)) \)
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 $j$th survey of unit $C_k$:
  \[
  \text{logit}(p_{kj}) = \gamma' \nu(C_k)
  \]

Observations (e.g., $J_k$ replicate counts)

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

\[
L(\beta, \gamma) = \prod_{k=1}^{K} \sum_{n_k = \max(y_k)}^{\infty} \frac{\exp\{-\mu(C_k)\}\mu(C_k)^{n_k}}{n_k!} \prod_{j=1}^{J_k} \binom{n_k}{y_{kj}} p_{kj}^{y_{kj}} (1 - p_{kj})^{n_k - y_{kj}}
\]
Information in Opportunistic Surveys Can Be Limited

\[
\log\{L(\beta, \alpha)\} = - \int_B \lambda(s)p(s) \, ds + \sum_{i=1}^{m} \log\{\lambda(s_i) p(s_i)\}
\]

where

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

Identifiability problems:

1. If \( p(s) = p, \beta_0 \) and \( \alpha_0 \) are not identified.
2. If \( p(s) \) is low \( \forall s \), \( \lambda(s)p(s) = \exp\{\beta' x(s) + \alpha' w(s)\} \)
   - \( \beta_0 \) and \( \alpha_0 \) are not identified
   - other elements of \( \beta \) and \( \alpha \) are not identified if \( x \) and \( w \) are linearly dependent
3. If Fisher information matrix \( I(\theta) \) is less than full rank, the parameters in \( \theta = (\beta', \alpha')' \) are not identified (Bowden, 1973).
Two models

\[
\log(\lambda(s)) = \\
\log(8000) + 0.5x(s)
\]

1. \[
\text{logit}(p(s)) = \\
\alpha_0 - 1.0w(s)
\]

2. \[
\text{logit}(p(s)) = \\
\alpha_0 - 1.0x(s)
\]
Using Planned Surveys To Overcome Limited Information in Opportunistic Surveys

\[ L(\beta, \alpha, \gamma) = L(\beta, \alpha) \times L(\beta, \gamma) \]

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

Two models

\[ \log(\lambda(s)) = \log(8000) + 0.5 \cdot x(s) \]

1. \[ \logit(p(s)) = -1.0 - 1.0 \cdot w(s) \]
   \[ \logit(p_{kj}) = 0.0 - 1.0 \cdot v(C_k) \]
   where \( v(C) = \int_C w(s) ds \)

2. \[ \logit(p(s)) = -1.0 - 1.0 \cdot x(s) \]
   \[ \logit(p_{kj}) = 0.0 - 1.0 \cdot v(C_k) \]
   where \( v(C) = \int_C x(s) ds \)
Abundance (covariate $x$)

Detections (covariate $w$)

Detections (covariate $x$)
Simulation Results: Detection covariate $w$

Bias

$\hat{\beta}_0$

$\hat{\beta}_1$

Std. Deviation

Number of sample units

Dorazio (USGS)
Simulation Results: Detection covariate $x$

<table>
<thead>
<tr>
<th>Number of sample units</th>
<th>Std. Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.5</td>
</tr>
<tr>
<td>200</td>
<td>0.45</td>
</tr>
<tr>
<td>400</td>
<td>0.4</td>
</tr>
<tr>
<td>600</td>
<td>0.35</td>
</tr>
<tr>
<td>800</td>
<td>0.3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Number of sample units</th>
<th>Bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-0.3</td>
</tr>
<tr>
<td>200</td>
<td>-0.25</td>
</tr>
<tr>
<td>400</td>
<td>-0.2</td>
</tr>
<tr>
<td>600</td>
<td>-0.15</td>
</tr>
<tr>
<td>800</td>
<td>-0.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Number of sample units</th>
<th>$\beta_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>200</td>
<td>-0.05</td>
</tr>
<tr>
<td>400</td>
<td>0.0</td>
</tr>
<tr>
<td>600</td>
<td>0.05</td>
</tr>
<tr>
<td>800</td>
<td>0.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Number of sample units</th>
<th>$\beta_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>200</td>
<td>-0.1</td>
</tr>
<tr>
<td>400</td>
<td>0.0</td>
</tr>
<tr>
<td>600</td>
<td>0.1</td>
</tr>
<tr>
<td>800</td>
<td>0.2</td>
</tr>
</tbody>
</table>

Dorazio (USGS)

Presence-only data

08 Sep 2014
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

Fisher Information Matrix

\[ I(\beta, \alpha) = \begin{pmatrix} I(\beta, \beta) & I(\beta, \alpha) \\ I(\beta, \alpha)' & I(\alpha, \alpha) \end{pmatrix} \]

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

\[
I(\beta_p, \beta_q) = \int_B x_p(s) x_q(s) \lambda(s) p(s) \, ds
\]

\[
I(\beta_p, \alpha_q) = \int_B x_p(s) w_q(s) \lambda(s) p(s) \{1 - p(s)\} \, ds
\]

\[
I(\alpha_p, \alpha_q) = \int_B w_p(s) w_q(s) \lambda(s) p(s) \{1 - p(s)\}^3 \left[1 - \exp\{2\eta(s)\}\right] \, ds
\]

\[
+ \int_B w_p(s) w_q(s) \lambda(s) p(s)^2 \{1 - p(s)\} \, ds
\]

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


