Estimating the Geographic Distribution of a Species Using Presence-only Records

Robert M. Dorazio

1Southeast Ecological Science Center, U.S. Geological Survey, Gainesville Florida

USGS Webinar hosted by Southeast Ecological Science Center
Gainesville, Florida
08 September 2011
**Definition:** A SDM expresses a quantitative association/relationship between the occurrence probability 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 – say, habitat restoration (or degradation!) – on a species’ distribution
- etc.
SDMs For Presence-Absence Data

Presence-absence data

<table>
<thead>
<tr>
<th>Site</th>
<th>Apparent species occurrence</th>
<th>Covariate(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>$x_1$</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>$x_2$</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>$x_3$</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>n</td>
<td>0</td>
<td>$x_n$</td>
</tr>
</tbody>
</table>

Logistic regression and related models

- Generalized linear models (McCullagh and Nelder, 1989)
- Generalized additive models (Hastie and Tibshirani, 1990)
- Boosted regression trees (Friedman, 2001)
- Multivariate adaptive regression splines (Friedman, 1991)
Suppose $i = 1, \ldots, n$ locations are selected independently.

Observations:

- Apparent species occurrence: $y_i = \begin{cases} 1, & \text{if present;} \\ 0, & \text{if absent.} \end{cases}$
- $x_i = \text{landscape or habitat covariate}$

Model:

- $y_i \sim \text{Bernoulli}(\pi(x_i))$
- $\pi(x_i) = 1/[1 + \exp(-(\alpha + \beta x_i))]$

So what’s wrong with using $\hat{\pi}(x_i) = \Pr(y_i = 1|x_i, \hat{\alpha}, \hat{\beta})$ as an estimator of occurrence probability?
1. Repeated observations typically don’t look like this!

<table>
<thead>
<tr>
<th>Site</th>
<th>Replicate</th>
<th>Observed no. detections</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1 1 1 1 1</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>0 0 0 0 0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0 0 0 0 0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1 1 1 1 1</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>0 0 0 0 0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>0 0 0 0 0</td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
2. Inclusion of detection covariates is problematic

- e.g., sampling effort, sampling method, observer, weather
- How would we interpret a distribution map based on $\pi_i$?

$$\logit(\pi_i) = \alpha + \beta \ast \text{elevation}_i + \gamma \ast \text{effort}_i$$

- Such covariates affect detection *given species is present* – not species presence
- Values of such covariates may vary among within-site replicates

*We need a sampling protocol and statistical model that addresses both issues!*
Imperfect Detectability

In practice, *observed zeros are ambiguous*:
- sampling zeros
- “structural” (fixed) zeros

We have to make a formal distinction between the *true occupancy state* $z$ and the *observed occupancy state* $y$:
- detection probability: $p = \Pr(y = 1|z = 1)$
- occurrence probability: $\psi = \Pr(z = 1)$
Suppose we have $J = 4$ independent, binary observations per site:

<table>
<thead>
<tr>
<th>Site $i$</th>
<th>Replicate</th>
<th>Observed no. detections $y_i$</th>
<th>Partially observed indicator of presence $z_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0 1 0 1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>0 1 1 1</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>0 0 0 0</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>1 0 0 0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>0 0 0 0</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>6</td>
<td>0 0 0 0</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

By assuming closure (i.e., true occupancy state $z_i$ is fixed), replication provides direct information about the detection probability $p$. 
Hierarchical site-occupancy model (MacKenzie et al., 2002; Tyre et al., 2003):

**latent process model:** \( z_i | \psi \sim \text{Bernoulli}(\psi) \)

**observation model:**

- If \( z_i = 1 \), \( y_i \sim \text{Bin}(J, p) \)
- If \( z_i = 0 \), \( y_i = 0 \)
Hierarchical site-occupancy model (MacKenzie et al., 2002; Tyre et al., 2003):

**latent process model:** \( z_i | \psi \sim \text{Bernoulli}(\psi) \)

**observation model:**
- If \( z_i = 1 \), \( y_i \sim \text{Bin}(J, p) \)
- If \( z_i = 0 \), \( y_i = 0 \)

more compactly:

**latent process model:** \( z_i | \psi \sim \text{Bernoulli}(\psi) \)

**observation model:** \( y_i | z_i, p \sim \text{Bin}(J, pz_i) \)

\[ [y_i | \psi, p] = \psi \text{ Bin}(y_i | J, p) + (1 - \psi)I(y_i = 0) \]

*Site-occupancy model may be viewed as an extension of logistic regression for within-site replicates.*
Example: Estimate geographic distribution of Swiss breeding birds

Swiss survey of common breeding birds:

- 237 1-km² quadrats
- each route surveyed 2 or 3 times
- volunteer observers
- covariates: elevation, forest cover, route length, duration of sampling

Elevation (m)
- (0,500]
- (500,1000]
- (1000,1500]
- (1500,2000]
- (2000,2500]
- (2500,4341]

Forest cover (%)
- (0,10]
- (10,25]
- (25,50]
- (50,75]
- (75,90]
- (90,100]
### Willow tit data

- territory counts are quantized to presence/absence

<table>
<thead>
<tr>
<th>y.1</th>
<th>y.2</th>
<th>y.3</th>
<th>elev</th>
<th>forest</th>
<th>dur.1</th>
<th>dur.2</th>
<th>dur.3</th>
<th>day.1</th>
<th>day.2</th>
<th>day.3</th>
<th>length</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>420</td>
<td>3</td>
<td>240</td>
<td>240</td>
<td>240</td>
<td>29</td>
<td>58</td>
<td>73</td>
<td>6.2</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>450</td>
<td>21</td>
<td>160</td>
<td>155</td>
<td>140</td>
<td>13</td>
<td>39</td>
<td>62</td>
<td>5.1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1050</td>
<td>32</td>
<td>120</td>
<td>105</td>
<td>85</td>
<td>30</td>
<td>47</td>
<td>74</td>
<td>4.3</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1110</td>
<td>35</td>
<td>180</td>
<td>170</td>
<td>145</td>
<td>23</td>
<td>44</td>
<td>71</td>
<td>5.4</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>510</td>
<td>2</td>
<td>210</td>
<td>225</td>
<td>235</td>
<td>28</td>
<td>56</td>
<td>73</td>
<td>3.6</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>630</td>
<td>60</td>
<td>150</td>
<td>180</td>
<td>195</td>
<td>17</td>
<td>56</td>
<td>73</td>
<td>6.1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>590</td>
<td>5</td>
<td>115</td>
<td>105</td>
<td>95</td>
<td>16</td>
<td>37</td>
<td>76</td>
<td>5.1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>530</td>
<td>13</td>
<td>155</td>
<td>140</td>
<td>135</td>
<td>24</td>
<td>47</td>
<td>74</td>
<td>3.7</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1140</td>
<td>50</td>
<td>165</td>
<td>165</td>
<td>180</td>
<td>25</td>
<td>46</td>
<td>70</td>
<td>3.8</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>770</td>
<td>57</td>
<td>220</td>
<td>230</td>
<td>255</td>
<td>21</td>
<td>38</td>
<td>50</td>
<td>7.7</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1220</td>
<td>84</td>
<td>135</td>
<td>150</td>
<td>135</td>
<td>30</td>
<td>60</td>
<td>74</td>
<td>3.1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>460</td>
<td>17</td>
<td>240</td>
<td>230</td>
<td>250</td>
<td>28</td>
<td>49</td>
<td>81</td>
<td>4.4</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1010</td>
<td>58</td>
<td>145</td>
<td>180</td>
<td>165</td>
<td>41</td>
<td>69</td>
<td>71</td>
<td>6.0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>760</td>
<td>26</td>
<td>175</td>
<td>270</td>
<td>210</td>
<td>18</td>
<td>50</td>
<td>72</td>
<td>6.2</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1300</td>
<td>32</td>
<td>260</td>
<td>270</td>
<td>290</td>
<td>35</td>
<td>58</td>
<td>75</td>
<td>6.1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1270</td>
<td>66</td>
<td>176</td>
<td>145</td>
<td>150</td>
<td>28</td>
<td>39</td>
<td>61</td>
<td>4.5</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>380</td>
<td>45</td>
<td>180</td>
<td>160</td>
<td>165</td>
<td>21</td>
<td>40</td>
<td>86</td>
<td>6.2</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>550</td>
<td>31</td>
<td>195</td>
<td>240</td>
<td>270</td>
<td>20</td>
<td>42</td>
<td>63</td>
<td>6.9</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>390</td>
<td>8</td>
<td>150</td>
<td>130</td>
<td>140</td>
<td>35</td>
<td>50</td>
<td>75</td>
<td>4.6</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1380</td>
<td>78</td>
<td>160</td>
<td>160</td>
<td>150</td>
<td>54</td>
<td>75</td>
<td>99</td>
<td>3.7</td>
</tr>
</tbody>
</table>
Willow tit analysis

\[
\text{logit}(\psi_i) = \beta_0 + \beta_1 \ast \text{elev}_i + \beta_2 \ast \text{elev}^2_i + \beta_3 \ast \text{forest}_i
\]

\[
\text{logit}(p_i) = \alpha
\]
Willow tit analysis
- geographic distribution based on model of occurrence

Occurrence probability
- (0,0.1]
- (0.1,0.2]
- (0.2,0.4]
- (0.4,0.6]
- (0.6,0.8]
- (0.8,1]

(photo courtesy Wikipedia commons)
Cost of Replicated Presence-Absence Surveys

1. May be difficult or expensive
2. Rare species: large samples! alternative sampling protocol?

Motivates analysis of presence-only data

Presence-only sample + Background sample = SDM?

- Species-presence locations recorded in museums, herbaria, or online databases
- Potential covariates of species occurrence recorded in geographical databases (e.g., GIS)
- Similar to estimating resource selection probabilities from “use-availability” data
Envelope- or point-to-point environmental similarity models (Busby, 1991; Carpenter et al., 1993)
- do not use background sample

Ecological niche factor analysis, ENFA (Hirzel et al., 2002)

Genetic algorithm for rule-set prediction, GARP (Stockwell and Peters, 1999; Peterson and Kluza, 2003)

Logistic-regression type models
- assume species is absent in background sample
- do not assume species is absent in background sample
  - case-control adjustment (Ward et al., 2009; Di Lorenzo et al., 2011)

Maximum entropy model, MAXENT (Phillips et al., 2006; Elith et al., 2010)

Spatial point-process model (Warton and Shepherd, 2010)
Challenges in Modeling Presence-Only Data

1. Imperfect detection

2. Sample selection bias – presence-only locations may not be
   - randomly selected
   - representative
     - of study area
     - of species’ geographic range
Museum records of koalas found along northern coast of Australia (Margules and Austin, 1994)

Figure 1. Koala records (courtesy of New South Wales National Parks & Wildlife Service) and the road network on part of the New South Wales north coast.
Figure 1. A map of field records of tree species on the Yucatan Peninsula, Mexico. These records map the road network. The map is courtesy of Guillermo Ibarra Manríquez and Rafael Durán.
Another Challenge – Species May be Present at Locations in Background Sample

Species presence vs. background $\neq$ Species presence vs. absence

(Assumes presence-only and background samples are randomly selected.)
Some Models Acknowledge Species Presences and Absences in Background Sample

- Logistic regression with case-control adjustment (Ward et al., 2009; Di Lorenzo et al., 2011)
- MAXENT (Phillips et al., 2006; Elith et al., 2010)

Problem – Estimators of species occurrence probability ($\Pr(z = 1|x)$) based on these models are biased unless species prevalence ($\Pr(z = 1)$) is known.
Objectives

1. Develop a consistent estimator of species occurrence probability, $\Pr(z = 1|x)$, assuming presence-only and background samples are collected \textit{without} selection bias.

2. Establish technical conditions for estimator’s consistency when species are detected imperfectly.

3. Analyze avian survey data
   - compare site-occupancy estimates vs. presence-only estimates.

4. Summarize implications of results for alternative approaches (MAXENT; Spatial point-process models).
Definitions

Species occurrence: $z = 1$ (present), $z = 0$ (absent)

Covariate(s) of occurrence: $x$

Pdf for distribution of covariates: $f(x)$

Conditional probability of species presence: $\Pr(z = 1|x) = \psi(x, \beta)$

e.g., $\psi(x, \beta) = \frac{\exp(\beta_0 + \beta'x)}{1 + \exp(\beta_0 + \beta'x)}$
Definitions

Species occurrence: \( z = 1 \) (present), \( z = 0 \) (absent)

Covariate(s) of occurrence: \( x \)

Pdf for distribution of covariates: \( f(x) \)

Conditional probability of species presence: \( \Pr(z = 1|x) = \psi(x, \beta) \)

\[ \text{e.g., } \psi(x, \beta) = \frac{\exp(\beta_0 + \beta'x)}{1 + \exp(\beta_0 + \beta'x)} \]

Using Bayes’ rule

\[ g(x|z = 1) = \frac{\Pr(z = 1|x) f(x)}{\Pr(z = 1)} \]

\[ = \frac{\psi(x, \beta) f(x)}{\int \psi(w, \beta) f(w) \, dw} \]

\[ = \frac{\psi(x, \beta) f(x)}{\mathbb{E}_{f}[\psi(w, \beta)]} \]
area of shaded region = $\Pr(z = 1) = \mathbb{E}_f[\psi(x, \beta)]$
Maximum Likelihood Estimation of $\theta = (\beta_0, \beta)$

**Observations – assumed representative and independent**

Presence-only sample: $x = (x_1, \ldots, x_n)$

Background sample: $w = (w_1, \ldots, w_m)$

**Log-likelihood function**

$$l(\beta_0, \beta | x, w) = -n \log \left\{ \int \psi(w, \beta) f(w) \, dw \right\} + \sum_{i=1}^{n} \log \{ \psi(x_i, \beta) \}$$

$$= -n \log \{ E_f[\psi(w, \beta)] \} + \sum_{i=1}^{n} \log \{ \psi(x_i, \beta) \}$$

**Monte Carlo integration:**

$E_f[\psi(w, \beta)] \approx \frac{1}{m} \sum_{j=1}^{m} \psi(w_j, \beta)$

(Lele and Keim, 2006)

**Numerical quadrature:**

$E_f[\psi(w, \beta)] \approx \int \psi(w, \beta) \hat{f}(w) \, dw$

where $\hat{f}(w)$ is a parametric or semiparametric approximation of $f(w)$
Simulation Study – Effect of Background Sample Size, $m$

- $f(x) = 0.6 \text{N}(x|2, 1) + 0.4 \text{N}(x|6, 4)$
- presence-absence sample of 4000 $\Rightarrow \bar{n} = 1614$
“Presence-only methods attempt to circumvent the detectability issue by only using data from locations where a species was known to occur.” (Rota et al., 2011)

Log-likelihood function $l(\beta_0, \beta | x, w)$ does not depend on observed zeros; so do detection errors influence the MLE?
Definitions

Species detection: $y = 1$ (detected), $y = 0$ (not detected)

Conditional probability of detecting species: $\Pr(y = 1|z = 1) = p$

Covariate(s) of detection: $u$

therefore, $p = p(u, \alpha) = \frac{\exp(\alpha_0 + \alpha'u)}{1 + \exp(\alpha_0 + \alpha'u)}$
Influence of Errors in Species Detection

Definitions

Species detection: \( y = 1 \) (detected), \( y = 0 \) (not detected)

Conditional probability of detecting species: \( \Pr(y = 1|z = 1) = p \)

Covariate(s) of detection: \( u \)

therefore, \( p = p(u, \alpha) = \frac{\exp(\alpha_0 + \alpha' u)}{1 + \exp(\alpha_0 + \alpha' u)} \)

Assuming no false-positive errors in detection

\[
\Pr(y = 1|x, u) = \Pr(z = 1|x) \Pr(y = 1|z = 1) + \Pr(z = 0|x) \Pr(y = 1|z = 0)
\]
\[
= \psi(x, \beta) p(u, \alpha) + \{1 - \psi(x, \beta)\} \cdot 0
\]
\[
= \psi(x, \beta) p(u, \alpha)
\]
In analysis of presence-only data, covariates are observed only at locations where the species is detected \((y = 1)\).

Using Bayes’ rule

\[
p(x, u | y = 1) = \frac{\Pr(y = 1 | x, u) p(x, u)}{\Pr(y = 1)}
\]
In analysis of presence-only data, covariates are observed only at locations where the species is detected \((y = 1)\).

Using Bayes’ rule:

\[
p(x, u|y = 1) = \frac{\Pr(y = 1|x, u) p(x, u)}{\Pr(y = 1)}
\]

Consider 3 cases:

- Distinct and independent covariates: \(p(x, u) = f(x) h(u)\)
- Identical covariates: \(p(x, u) = f(x)\)
- Distinct and dependent covariates: \(p(x, u) = f(x) h(u|x)\)
Influence of Errors in Species Detection

Case 1: Distinct and independent covariates

\[ p(x, u|y = 1) = \frac{\Pr(y = 1|x, u) f(x)h(u)}{\Pr(y = 1)} \]

\[ = \frac{\psi(x, \beta) p(u, \alpha) f(x)h(u)}{\int \int \psi(w, \beta) p(v, \alpha) f(w)h(v) \, dw \, dv} \]

\[ = \frac{\psi(x, \beta) f(x) p(u, \alpha)h(u)}{\int \psi(w, \beta) f(w) \, dw \int p(v, \alpha)h(v) \, dv} \]

\[ = g(x|z = 1) \frac{p(u, \alpha) h(u)}{\int p(v, \alpha) h(v) \, dv}. \]

Integrating both sides w.r.t. \( u \) yields

\[ p(x|y = 1) = g(x|z = 1) \]
Case 2: Identical covariates

\[
p(x|y = 1) = \frac{\Pr(y = 1|x) f(x)}{\Pr(y = 1)} = \frac{\psi(x, \beta) p(x, \alpha) f(x)}{\int \psi(w, \beta) p(w, \alpha) f(w) dw}
\]

Thus,

\[p(x|y = 1) \neq g(x|z = 1)\]

In fact, \(p(x|y = 1)\) cannot even be used as a likelihood function for \(\alpha\) and \(\beta\) because they are unidentified.
Case 3: Distinct and dependent covariates

\[
p(x, u|y = 1) = \frac{\Pr(y = 1|x, u) f(x) h(u|x)}{\Pr(y = 1)} = \frac{\psi(x, \beta) p(u, \alpha) f(x) h(u|x)}{\int \int \psi(w, \beta) p(v, \alpha) f(w) h(v|w) \, dw \, dv}.
\]

Thus,

\[p(x|y = 1) \neq g(x|z = 1)\]

**Conclusion:** MLE is consistent for the parameters of \(\psi(x, \beta)\) provided the covariates of species occurrence probability are distinct and independently distributed from those of species detection probability.
Simulation Study – Site-occupancy vs. Presence-only models

- \( f(x) = 0.6 \, N(x|2, 1) + 0.4 \, N(x|6, 4) \)
- \( \psi(x, \beta) = \frac{\exp(-4+1x)}{1+\exp(-4+1x)} \)
- \( p(u, \alpha) = \frac{\exp(-1.5+0.4u)}{1+\exp(-1.5+0.4u)} \)
- detection-nondetection sample of 4000 locations
  
  \[ z_i \sim \text{Bernoulli}(\psi(x_i, \beta)) \]
  \[ y_i | z_i \sim \text{Binomial}(5, z_i \, p(u_i, \alpha)) \]

- presence-only sample \((i : y_i > 0)\); background sample \((m = 10000)\)

Two cases:

Independent covariates: \( h(u) = N(u|1, 4) \)

Identical covariates: \( u = x \)
Simulation Study – Site-occupancy (OCC) vs. Presence-only (PO) models

Independent covariates

\[
\hat{\beta}_0
\]

\[
\begin{array}{c}
\text{OCC} \\
\text{PO}
\end{array}
\]

Identical covariates

\[
\hat{\beta}_1
\]

\[
\begin{array}{c}
\text{OCC} \\
\text{PO}
\end{array}
\]
482 permanently marked transects during breeding seasons of 1994–2008 (Rota et al., 2011)

Figure 1. Location of permanently marked Northern Region Landbird Monitoring Program transects used to estimate the distribution of forest birds. The black dots represent transects selected for building species distribution models (SDMs), and the open rectangles represent transects selected for validating SDMs.
Case Study: Geographic Distribution of Avian Species

Sampling protocol

- 10 sites spaced 300 m apart along each transect
- standard, 10-min point count survey of birds detected within 100 m
  - 10-min survey divided into two 5-minute intervals
  - observer recorded sampling interval of first detection

Criteria for comparison

- sites surveyed using two 5-minute intervals
- sites with reliable GIS data
- 2004 (year with highest number of transects (268))
Data

- detection-nondetection data at 2635 sites
- identities of 13 trained observers
- 3 species:
  - Townsend’s warbler (TW) detected at 51% (1340) of sites
  - Golden-crowned kinglet (GK) detected at 20% (529) of sites
  - Pileated woodpecker (PW) detected at 7% (184) of sites
- background covariates for 3,506,162 pixels
  - based on a 200-m resolution (40,000 m$^2$), which roughly corresponds to the size of a sample unit (31,416 m$^2$)
  - random subset of background data (25%: $m = 876,540$ observations) used in analysis
- potential covariates of occurrence differed by species
## Case Study: Occurrence Parameter Estimates

<table>
<thead>
<tr>
<th>Species</th>
<th>Parameter</th>
<th>Site-occupancy model</th>
<th>Presence-only model</th>
</tr>
</thead>
<tbody>
<tr>
<td>TW</td>
<td>Intercept</td>
<td>-4.51 0.64</td>
<td>-8.81 2.77</td>
</tr>
<tr>
<td>TW</td>
<td>CC2</td>
<td>1.00 0.15</td>
<td>0.60 0.12</td>
</tr>
<tr>
<td>TW</td>
<td>DBH1</td>
<td>0.45 0.12</td>
<td>0.31 0.10</td>
</tr>
<tr>
<td>TW</td>
<td>ELEV</td>
<td>8.91 1.02</td>
<td>10.03 1.12</td>
</tr>
<tr>
<td>TW</td>
<td>ELEV²</td>
<td>-3.79 0.38</td>
<td>-4.22 0.46</td>
</tr>
<tr>
<td>GK</td>
<td>Intercept</td>
<td>-5.60 0.87</td>
<td>-8.83 3.99</td>
</tr>
<tr>
<td>GK</td>
<td>DBH1</td>
<td>0.92 0.15</td>
<td>0.60 0.19</td>
</tr>
<tr>
<td>GK</td>
<td>STREAM</td>
<td>-0.89 0.24</td>
<td>-1.38 0.27</td>
</tr>
<tr>
<td>GK</td>
<td>ELEV</td>
<td>7.49 1.37</td>
<td>9.60 1.72</td>
</tr>
<tr>
<td>GK</td>
<td>ELEV²</td>
<td>-2.52 0.50</td>
<td>-3.64 0.63</td>
</tr>
<tr>
<td>PW</td>
<td>Intercept</td>
<td>-2.03 1.20</td>
<td>-13.05 45.87</td>
</tr>
<tr>
<td>PW</td>
<td>DBH1K1</td>
<td>1.05 0.40</td>
<td>1.90 0.30</td>
</tr>
<tr>
<td>PW</td>
<td>DBH1K2</td>
<td>1.28 0.51</td>
<td>1.25 0.36</td>
</tr>
<tr>
<td>PW</td>
<td>ELEV</td>
<td>1.50 1.77</td>
<td>5.75 1.10</td>
</tr>
<tr>
<td>PW</td>
<td>ELEV²</td>
<td>-0.82 0.65</td>
<td>-2.45 0.43</td>
</tr>
</tbody>
</table>
Case Study: Occurrence Parameter Estimates

![Graph showing Site-occupancy estimate vs Presence-only estimate for different species]

- **Species**:
  - TW
  - GK
  - PW

- **Legend**:
  - TW: Black
  - GK: Blue
  - PW: Red

- **Axes**:
  - Site-occupancy estimate
  - Presence-only estimate

- **Note**: Dorazio (USGS) - Presence-only Models - 08 Sep 2011
Case Study: Detection Probability Estimates

![Graph showing probability of detection vs. observer](image)

- Observer
- Probability of detection

Presence-only Models (courtesy Wikipedia commons)
Case Study: Profile Likelihood of $\beta_0$
(Townsend’s Warbler: $\hat{\beta}_0 = -2.70$ for centered and scaled $x$)
Case Study: Profile Likelihood of $\beta_0$

(Townsend’s Warbler: $\hat{\beta}_0 = -2.70$ for centered and scaled $x$)
\[
\Pr(z = 1|x) = \frac{\Pr(z = 1) g(x|z = 1)}{f(x)} = \Pr(z = 1) \exp(\alpha + \beta h(x))
\]

**Approach:** Estimate \(g(x|z = 1)/f(x)\) by minimizing the Kullback-Leibler divergence between distributions \(x|z = 1\) and \(x\) subject to constraints that keep the model-based mean \(x|z = 1\) close to the mean value of \(x\) observed in the presence-only sample.

**Condition:** Species prevalence \(\Pr(z = 1)\) is neither known nor estimated.

**Question:** By def: \(f(x) = \Pr(z = 1)g(x|z = 1) + \Pr(z = 0)g(x|z = 0)\); therefore, why minimize KL divergence?

"because \(f(x)\) is a null model for \(g(x|z = 1)\): without any occurrence data, we would have no reason to expect the species to prefer any particular environmental conditions over any others, so we could do no better than predict that the species occupies environmental conditions proportionally to their availability in the landscape." (Elith et al., 2010)
Logistic Regression With Case-control Adjustment
(Lancaster and Imbens, 1996; Ward et al., 2009)

Definitions

- Pr(“case”): $\gamma_1 = \Pr(s = 1|z = 1)$
- Pr(“control”): $\gamma_0 = \Pr(s = 1|z = 0)$
- Prevalence: $\pi = \Pr(z = 1)$

$logit(\Pr(z = 1|s = 1, x)) = \log(\gamma_1/\gamma_0) + logit(\Pr(z = 1|x))$

$= \log(1 + n/(m\pi)) + logit(\psi(x, \beta))$

$= \log(1 + n/(m\pi)) + \beta_0 + \beta'x$
Logistic Regression With Case-control Adjustment
(Ward et al., 2009)

\[ L(\beta_0, \beta | x, w) = \prod_{i=1}^{n} \frac{\frac{n}{m\pi} \psi(x_i, \beta)}{1 + (1 + \frac{n}{m\pi}) \psi(x_i, \beta)} \cdot \prod_{i=1}^{m} \frac{1 + \psi(w_i, \beta)}{1 + (1 + \frac{n}{m\pi}) \psi(w_i, \beta)} \]

“Proposition 7: Identifiability of \( \pi \) can be summarized as follows:

1. \( \pi \) is not identifiable if we make no assumptions about the structure of \( \eta \).
2. \( \pi \) is identifiable only if we make unrealistic assumptions about the structure of \( \eta(x) \) such as in logistic regression where \( \eta(x) \) linear in \( x \) : \( \eta(x) = x^T \beta \).
3. Even when \( \pi \) is identifiable, the estimate is highly variable.”
Let $\lambda(s) = \lim_{|ds| \to 0} \frac{E[N(ds)]}{|ds|}$ for $s \in A(\subset \mathbb{R}^2)$

i.e., $\lambda(s)$ is the limiting expected density of animals (no. per area) at point location $s$

**Inhomogeneous Poisson point process**

1. $\Pr(N(A) = n) = \exp(-\mu(A)) \mu(A)^n / n!$

2. $[s_1, s_2, \ldots, s_n | n] = \prod_{i=1}^{n} \lambda(s_i) / \mu(A)$

where $\mu(A) = \int_A \lambda(s) ds$

$\therefore [s_1, s_2, \ldots, s_n, n] = \frac{\exp(-\mu(A))}{n!} \prod_{i=1}^{n} \lambda(s_i)$

$$l(\beta_0, \beta | n, s) = - \int_{A} \lambda(s) ds + \sum_{i=1}^{n} \log(\lambda(s_i))$$

$$= - \int_{A} \exp(\beta_0 + \beta' x(s)) ds + \sum_{i=1}^{n} \beta_0 + \beta' x(s_i)$$
In most analyses, \( A \) is discretized into \( M \) pixels

1. Size of \( i \)th pixel = \( a_i = |A|/M \)
2. Average value of \( x(s_i) = x_i \)

Poisson approximation is assumed for number of individuals \( N_i \) present in \( i \)th pixel

\[
N_i|x_i, a_i \sim \text{Poisson}(a_i \exp(\beta_0 + \beta' x_i))
\]

Note: \( z_i = 1 \iff N_i > 0 \)

\[
\Pr(z_i = 1|x_i) = \Pr(N_i > 0|x_i, a_i)
\]

\[
\psi(x_i, \beta) = 1 - \exp(-a_i \exp(\beta_0 + \beta' x_i))
\]

\[
\log(1 - \psi(x_i, \beta)) = -a_i \exp(\beta_0 + \beta' x_i)
\]

\[
\log(-\log(1 - \psi(x_i, \beta))) = \log(a_i) + \beta_0 + \beta' x_i
\]

Using cloglog link of \( \psi(x_i, \beta) \) allows estimation of parameters of spatial point-process models (also, see Baddeley et al. (2010)).
Conclusions

- Given enough presence-only data, a species’ geographic distribution can be predicted accurately provided:
  1. background and presence-only samples are selected without bias
  2. covariates of species occurrence probabilities are distinct and independently distributed from covariates of species detection probabilities
- Predictions of species occurrence do not require species prevalence to be known
  - in fact, $E_f[\psi(w, \hat{\beta})]$ can be computed

Remaining challenges

- Condition #1 is seldom met
- Analyst never knows whether condition #2 is satisfied
Acknowledgments

**Motivation:** Ed Connor (San Francisco State Univ) and Josh Ladau (Univ California, San Francisco)

**Avian data:** Rob Fletcher (Univ Florida), Chris Rota (Univ Missouri), Dick Hutto (Univ Montana)

**Ms. review:** Marc Kéry (Swiss Ornithological Institute)

**Funding:**
- Natl Inst for Mathematical and Biological Synthesis (NSF Award #EF-0832858)
- Univ Tennessee, Knoxville
- U.S. Geological Survey, Southeast Ecological Science Center


