Statistical Modeling to Analyze Species Distributions
Introduction

Learning about species distributions is a long-standing issue in ecology with an enormous literature.

Objectives include:
- predicting the response of species to climate change (e.g. Midgley et al. 2002),
- identifying and managing conservation areas (e.g. Austin and Meyers 1996),
- finding additional populations of known species or related sibling species (e.g. Raxworthy et al. 2003),
- seeking evidence of competition among species (e.g. Leathwick 2002).

Data on where a species occurs (and where it does not) with environmental information to predict presence or absence in unsampled locations.
Challenges in Biogeography

• Predicting or explaining species distributions
  – Ecological explanations
  – Evolutionary explanations, -- phylogeography or historical biogeography

• Predicting or explaining spatial patterns in Biodiversity
  -- Measures of biodiversity
  -- Species richness
  -- Species turnover
Explaining Patterns in Biodiversity

The usual approach with response variables:

• **Species richness**, i.e. the number of species per unit area
• **Index** of biodiversity, e.g. species richness incorporating proportional representation of species.

What about explanatory variables?
**TABLE 22.1** Factors hypothesized to influence biodiversity.

<table>
<thead>
<tr>
<th>Factor</th>
<th>Rationale</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. History</td>
<td>More time permits more complete colonization and the evolution of new species</td>
</tr>
<tr>
<td>2. Habitat heterogeneity</td>
<td>Physically or biologically complex habitats provide more niches</td>
</tr>
<tr>
<td>3. Competition</td>
<td>Competition affects niche partitioning</td>
</tr>
<tr>
<td>4. Predation</td>
<td>Predation retards competitive exclusion</td>
</tr>
<tr>
<td>5. Climate</td>
<td>Fewer species can tolerate climatically unfavorable conditions</td>
</tr>
<tr>
<td>6. Climatic variability</td>
<td>Fewer species are adapted to tolerate variable environments</td>
</tr>
<tr>
<td>7. Productivity</td>
<td>Richness is limited by the partitioning of production or energy among species</td>
</tr>
<tr>
<td>8. Disturbance</td>
<td>Moderate disturbance retards competitive exclusion</td>
</tr>
</tbody>
</table>

*Source: Modified after Currie (1991), and Fraser and Currie (1996).*
A comment on the status of the field:

“...in many ways the science of biodiversity [and biogeography] is not much further along than medicine was in the Middle Ages.”

From the Preface of *The Unified Neutral Theory of Biodiversity and Biogeography* by Stephen Hubbell, Princeton University Press (2001)
How to specify species distributions and ranges?

Convex hulls of observed presences are poor proxies.
Cape Floristic (Biogeographic) Region

90,000 km$^2$

9000+ plant species

70% found nowhere else

1400+ threatened or endangered species
Cape of Good Hope Peninsula

Area: <500km²

Plant species: 2600+ total (160+ found nowhere else)
    130+ threatened or endangered
    39 extinctions over last 100 years
Protea Atlas sample sites across the CFR

~60,000 sites, ~250,000 species-site records,
globally one of the most intensively sampled group of species for any biogeographic region.
Kogelberg-Hawequas sub-region

Sample locations, with a 1x1 minute grid overlain (\(\sim 41 \times 107 \text{ km} = 1554 \text{ grid cells}\))

[versus \(\sim 37,000 \text{ cells for full CFR}\)]
The areal grid

A set of i (1x1 minute) grid cells

\[i = 1554 \text{ for K-H sub-region}\]

Associated with each cell: a suite of site attributes.

For grid cell \(i\), \(n_i\) sample sites are located. At each \(j\), p/a of all protea species are inventoried.

\[
Y_{ij}^{(k)} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}
\]

species \(k\) observed at sample site \(j\) in grid cell \(i\).
Fynbos
Environmental Explanatory Variables (24)

• Topographic (elevation, topographic heterogeneity)
• Climate related (temperature, rainfall, solar radiation, seasonality, greenness, soil moisture availability, etc.)
• Soils [Geology driven] (soil fertility, texture, pH)
Species Attribute Variables:

- Plant height
- Local population size (1-50, 50-1000, >1000)
- Local regeneration (resprouting) of adults (yes, no)
- Dispersal mode (wind, animal)
- Pollination mode (wind, bird, other)
How to characterize species patterns in a human altered landscape?
Specify *potential* species patterns, in the absence of human activities, and *adjusted* patterns, given human transformed landscapes.
Specifying transformed landscapes:

Kogelberg-Hawequas sub-region

CFR

Red=transformed pixels
Specifying transformed landscapes:

$U_i =$ proportion of cell $i$ transformed, which \textit{adjusts} potential p/a patterns of species.

Sample site $j$ in cell $i$
Descriptive Summaries

- Site (rows) by species (columns) table
- Entries are either p/a or abundance
- Group sites containing similar species (communities)
- Group species at similar sites (assemblages)
- Row marginals – site level richness
- Column marginals – species prevalence
- (With or without aggregation)
A Binomial Regression Model

Let $Y(s)$ be the p/a (1/0) of the species at location $s$.

Summing $Y(s)$ over the $n_i$ sites in cell $i$ yields counts $Y_{i+}$

Assuming independence for the trials, $Y_{i+} \sim Bi(n_i, p_i)$.

$p_i$ is a function of vector of environmental variables, $w_i$.

Logit link to relate $p_i$ to linear predictor $w_i^T \beta$, i.e.,

$$\log\left(\frac{p_i}{1 - p_i}\right) = w_i^T \beta$$

Only sampled pixels contribute to the likelihood
A site-level binary regression

Here $Y(s)$ would be taken as $Y(s) \sim \text{Bernoulli}(p(s))$.

Probability $p(s)$ that the species occurs in site $s$ again relates to the set of environmental variables as

$$\log\left(\frac{p(s)}{1 - p(s)}\right) = w(s)^T \beta$$

Such modeling requires that we have vector $w(s)$ for each site.

If we set $w(s) = w_i$ when $s$ is within grid cell $i$, we return to the same model as in previous slide.
A spatially explicit model

- Add spatial random effects to the model
- At grid cell level, a spatial term $\rho_i$ associated with cell $i$. Now, $\log(p_i/(1-p_i)) = w_i^T \beta + \rho_i$
- $\rho_i$ is cell level adjustment to the probability of presence.
- Conditional auto-regressive (CAR) model for \{\rho_i\}.
- Conditional distribution of $\rho_i$ depends only on the effects of the neighboring pixels.
- The Gaussian CAR model looks like $\rho_i | \rho_j, j \neq i \sim N(\sum_{j \in \partial_i} w_{ij} \rho_j / \sum_{j \in \partial_i} w_{ij}, \tau^2 / \sum_{j \in \partial_i} w_{ij})$
- Limitations of CAR model - don’t model dependence directly, only a smoother, no interpolation
A point level spatial model

- Model spatial dependence at point level through a spatial process model
- Model pairwise association using a parametric covariance function, say, \( \text{cov}(s_i, s_j) = C(d_{ij}; \phi) \), \( d_{ij} \) is the distance between locations \( s_i \) and \( s_j \).
- For example, exponential correlation \( \exp(-\phi d_{ij}) \). More generally, the Matern class
- Overcomes the disadvantages of the CAR model but computational limitation. The MCMC sampling algorithm is \( O(N^3) \).
- Big N problem - likelihood evaluation. Much recent discussion - dimension reduction.
A hierarchical spatial model

- A multi-level hierarchical model - spatial dependence, irregular sampling, land transformation

- For each $i$, latent variable $X_i$ to represent potential p/a state, and $V_i$ to represent the transformed p/a state,

- Model the latent $X_i$, i.e., $p_i = P(X_i = 1)$ with environmental variables expected to affect the p/a in unit $i$.

- Logit link given cell level environment variables and spatial random effects, $\log(p_i/(1-p_i)) = w_i^T \beta + \rho_i$. 
At the second level, model $P(V_i | X_i)$. Let $U_i$ denote the proportion of transformed area in the $i$th pixel.

Define $P(V_i = 0 | X_i = 1) = U_i$, i.e., probability of absence due to transformation given potential presence equals proportion transformed. $P(V_i = 1 | X_i = 0) = 0$.

Marginalizing over $X_i$, we have $P(V_i = 1) = (1 - U_i)p_i$.

At the third level the $Y_{ij}$, the p/a data.

To model $P(Y_{ij} | V_i)$, if $V_i = 1$, we again view the $Y_{ij}$ as i.i.d. Bernoulli($q_i$) and $Y_{ij} \sim Bi(n_i, q_i)$. Given $V_i = 0$, $Y_{ij} = 0$ with probability 1.

Again, a logistic regression model for $q_i$, reflecting species attributes.

We can accommodate “holes” arising from totally transformed cells or unsampled cells, through spatial modeling of the $X$'s, through the $\rho_i$'s.
Interpretation

- What does \( X_i = 1 \) mean?
- Not a binary process but a block average of a latent binary process
- \( P(X_i = 1) = p_i \) is a proportion, measures how suitable cell \( i \) is, i.e., \( p_i = \int_{cell_i} \lambda(s) ds / |A_i| \) where \( \lambda(s) \) is a binary process
- What does \( V_i = 1 \) mean?
- \( P(V_i = 1) = (1 - U_i)p_i \) is a proportion, measures how suitable and available, i.e., \( \int_{cell_i} T(s) \lambda(s) ds / |A_i| \) where \( T(s) = 0/1 \), transformed, not transformed
- Switches, indicators
- Again, conditionally, \( P(Y_{ij} = 1|V_i = 1) = q_i \). Marginally, \( P(Y_{ij} = 1) = q_i(1 - U_i)p_i \)
Figure 2: A graphical model for the hierarchical specification.
Comments

- Earlier Bayesian work
- Elementary approaches - Aspinall, Aspinall and Veitch, Brzeziecki
- Prior on presence/absence of a species at a site
- Convert to posterior probability of presence/absence given features of site using Bayes’ theorem
- Requires probabilities of features given presence/absence. SEEMS BACKWARDS!
- Hierarchical model - presence/absence at first stage, probabilities at second stage
- Where to be spatial? First stage - Hogmander et al., Hoeting et al., second stage - Gelfand et al.
Inference with regard to Biodiversity

- Range - observed; potential, $E(p_i|Y)$; transformed, $(1 - U_i)E(p_i|Y)$; soft or hard edges
- Prevalence - observed; potential, $[\sum_i p_i|Y]$
- Richness - observed, $\sum_k 1(Y_{i+}^{(k)} > 0)$ (only if $n_i > 0, U_i < 1$); potential, $[\sum_k p_i^{(k)}|Y]$ or $[\sum_k \logit(p_i^{(k)})|Y]$ (on scale of $w^T\beta^{(k)}$)
- Diversity measure (Shannon-Weiner)

$$\exp\left(-\sum_{k=1}^{L} \frac{p_i^{(k)}}{\sum_k p_i^{(k)}} \log \frac{p_i^{(k)}}{\sum_k p_i^{(k)}}\right)$$
cont.

- Max at \( \frac{p_i^{(k)}}{\sum_k p_i^{(k)}} = 1/L \), min if \( \frac{p_i^{(k)}}{\sum_k p_i^{(k)}} \to 1 \) for some \( k \).
- Many co-occur vs a few dominate
- Not a measure of richness; relative \( p \)'s vs. absolute \( p \)'s
- Turnover - \( p_i^{(k)} \to p_i \)

\[
TO_i = \exp\left(- \sum_{j \in \partial_i} \|E(p_i | Y) - E(p_j | Y)\|/m_i \right)
\]

- Average similarity; plots of \( TO_i \) vs components of \( w_i \)
TABLE 1. Posterior summary of significant/suggestive coefficients (β’s) for *Protea mundii.*

<table>
<thead>
<tr>
<th>Variable</th>
<th>Abbreviation</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Roughness</td>
<td>ROUGH</td>
<td>—</td>
<td>+</td>
<td>(+)</td>
<td>+</td>
</tr>
<tr>
<td>Elevation</td>
<td>ELEV</td>
<td>+</td>
<td>+</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>Potential evapotranspiration</td>
<td>POTEVT</td>
<td>+</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>Interannual CV precipitation</td>
<td>PPCTV</td>
<td>(-)</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>Frost season length</td>
<td>FROST</td>
<td>—</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>Heat units</td>
<td>HEATU</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>January maximum temperature</td>
<td>JANMAXT</td>
<td>+</td>
<td>+</td>
<td>(+)</td>
<td>+</td>
</tr>
<tr>
<td>July minimum temperature</td>
<td>JULMINT</td>
<td>NS</td>
<td>+</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>Seasonal concentration of precipitation</td>
<td>PPTCON</td>
<td>—</td>
<td>—</td>
<td>(-)</td>
<td>—</td>
</tr>
<tr>
<td>Summer soil moisture days</td>
<td>SUMSMD</td>
<td>+</td>
<td>NS</td>
<td>+</td>
<td>(+)</td>
</tr>
<tr>
<td>Winter soil moisture days</td>
<td>WINSMD</td>
<td>—</td>
<td>NS</td>
<td>—</td>
<td>NS</td>
</tr>
<tr>
<td>Enhanced vegetation index</td>
<td>EVI</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Low fertility</td>
<td>FERT1</td>
<td>+</td>
<td>NS</td>
<td>+</td>
<td>NS</td>
</tr>
<tr>
<td>Moderately low fertility</td>
<td>FERT2</td>
<td>NS</td>
<td>NS</td>
<td>(+)</td>
<td>NS</td>
</tr>
<tr>
<td>Moderately high fertility</td>
<td>FERT3</td>
<td>—</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>Fine texture</td>
<td>TEXT1</td>
<td>+</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>Acidic soil</td>
<td>PH1</td>
<td>—</td>
<td>NS</td>
<td>—</td>
<td>NS</td>
</tr>
<tr>
<td>Alkaline soil</td>
<td>PH3</td>
<td>—</td>
<td>NS</td>
<td>—</td>
<td>NS</td>
</tr>
</tbody>
</table>

*Notes:* For Tables 1 and 2, + and − denote positive and negative coefficients with 95% credible intervals that do not overlap 0; (+) and (−) denote positive and negative coefficients with 90% credible intervals that do not overlap 0; NS, not statistically significant.
<table>
<thead>
<tr>
<th>Variable</th>
<th>Abbreviation</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Roughness</td>
<td>ROUGH</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Elevation</td>
<td>ELEV</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Potential evapotranspiration</td>
<td>POTEVT</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Interannual CV precipitation</td>
<td>PPCTV</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Frost season length</td>
<td>FROST</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Heat units</td>
<td>HEATU</td>
<td>-</td>
<td>NS</td>
<td>-</td>
<td>NS</td>
</tr>
<tr>
<td>January maximum temperature</td>
<td>JANMAXT</td>
<td>-</td>
<td>NS</td>
<td>-</td>
<td>NS</td>
</tr>
<tr>
<td>July minimum temperature</td>
<td>JULMIN</td>
<td>+</td>
<td>NS</td>
<td>NS</td>
<td>(-)</td>
</tr>
<tr>
<td>Mean annual precipitation</td>
<td>MAP</td>
<td>+</td>
<td>NS</td>
<td>(+)</td>
<td>NS</td>
</tr>
<tr>
<td>Seasonal concentration of precipitation</td>
<td>PPTCON</td>
<td>-</td>
<td>NS</td>
<td>-</td>
<td>NS</td>
</tr>
<tr>
<td>Winter soil moisture days</td>
<td>WINSMD</td>
<td>-</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>Enhanced vegetation index</td>
<td>EVI</td>
<td>-</td>
<td>NS</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>Moderately high fertility</td>
<td>FERT3</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Fine texture</td>
<td>TEXT1</td>
<td>(+)</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
</tr>
<tr>
<td>Medium coarse texture</td>
<td>TEXT3</td>
<td>-</td>
<td>NS</td>
<td>-</td>
<td>NS</td>
</tr>
<tr>
<td>Coarse texture</td>
<td>TEXT4</td>
<td>-</td>
<td>NS</td>
<td>-</td>
<td>NS</td>
</tr>
<tr>
<td>Alkaline soil</td>
<td>PH3</td>
<td>-</td>
<td>NS</td>
<td>NS</td>
<td>NS</td>
</tr>
</tbody>
</table>

**Notes:** For Tables 1 and 2, + and − denote positive and negative coefficients with 95% credible intervals that do not overlap 0; (+) and (−) denote positive and negative coefficients with 90% credible intervals that do not overlap 0; NS, not statistically significant.
Model Choice

- Only with regard to specification for the $p_i$’s
- G - geological variables
- T - topographic variables
- C - climate variables
- A - species attributes
- $\rho$ - spatial random effects
- DIC (Spiegelhalter et al.) for model comparison
Model choice for the hierarchical model

- DIC (Spiegelhalter et al., 2002):

  \[ \text{DIC}(\Theta) = \bar{D}(\Theta) + p_D(\Theta) \]

  \( \bar{D}(\Theta) \) is posterior expectation of model deviance
  \( p_D(\Theta) \) is effective model dimension

  - \( p_D \) depends on parametrization

  - We consider two parametrizations, \( \Theta \) is the set of all model parameters, \( \Theta = \{ P(Y_{ij}^{(k)} = 1) \} \)

  - Fortunately model rankings agree for both parametrizations
Table 1: Model Comparison Using DIC

\[ \Theta = \{ P(Y_{ij}^{(k)} = 1) \} \quad \Theta = \{ \text{model parameters} \} \]

<table>
<thead>
<tr>
<th>models</th>
<th>( p_D )</th>
<th>DIC</th>
<th>( p_D )</th>
<th>DIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>G(Geology)</td>
<td>490.79</td>
<td>101614.73</td>
<td>718.30</td>
<td>101842.24</td>
</tr>
<tr>
<td>C(Clima., Precip.)</td>
<td>624.04</td>
<td>91369.39</td>
<td>994.58</td>
<td>91739.93</td>
</tr>
<tr>
<td>T(Topography)</td>
<td>268.77</td>
<td>104564.86</td>
<td>399.16</td>
<td>104695.25</td>
</tr>
<tr>
<td>( \rho )(Spatial effect)</td>
<td>934.61</td>
<td>109070.14</td>
<td>1128.69</td>
<td>109264.22</td>
</tr>
<tr>
<td>G,( \rho )</td>
<td>1160.62</td>
<td>100620.46</td>
<td>1724.01</td>
<td>101183.85</td>
</tr>
<tr>
<td>C,( \rho )</td>
<td>1167.91</td>
<td>90302.25</td>
<td>1904.65</td>
<td>91038.98</td>
</tr>
<tr>
<td>T,( \rho )</td>
<td>1021.68</td>
<td>103679.11</td>
<td>1392.69</td>
<td>104050.12</td>
</tr>
<tr>
<td>G,C,( \rho )</td>
<td>1271.82</td>
<td>88195.33</td>
<td>2166.15</td>
<td>89089.66</td>
</tr>
<tr>
<td>G,T,( \rho )</td>
<td>1216.52</td>
<td>96903.28</td>
<td>1917.91</td>
<td>97604.67</td>
</tr>
<tr>
<td>C,T,( \rho )</td>
<td>1213.75</td>
<td>89532.33</td>
<td>2004.36</td>
<td>90322.93</td>
</tr>
<tr>
<td>G,C,T</td>
<td>943.14</td>
<td>88739.56</td>
<td>1487.35</td>
<td>89283.78</td>
</tr>
<tr>
<td>G,C,T,( \rho )</td>
<td>1327.15</td>
<td>87788.84</td>
<td>2263.40</td>
<td>88725.09</td>
</tr>
<tr>
<td>G,C,T,A(Attributes)</td>
<td>950.77</td>
<td>88861.54</td>
<td>1500.34</td>
<td>89411.10</td>
</tr>
<tr>
<td>C,T,( \rho ),A</td>
<td>1217.49</td>
<td>89586.47</td>
<td>2013.98</td>
<td>90382.96</td>
</tr>
<tr>
<td>G,T,( \rho ),A</td>
<td>1214.63</td>
<td>96882.00</td>
<td>1908.65</td>
<td>97576.02</td>
</tr>
<tr>
<td>G,C,( \rho ),A</td>
<td>1269.62</td>
<td>88147.81</td>
<td>2156.59</td>
<td>89034.78</td>
</tr>
<tr>
<td>G,C,T,( \rho ),A(Full)</td>
<td>1323.41</td>
<td>87738.79</td>
<td>2243.82</td>
<td>88659.19</td>
</tr>
</tbody>
</table>
Table 2: Posterior Summary of Species Level Environmental Coefficients ($\beta$’s).

| SPECIES (#pixel obs.) | ROUGH | EV | TE | VT | PO | PP | OS | T | FR | AT | U | JM | AN | LM | MD | IN | MAP | M | PP TC | ON | MS | MD | EV | I | TE XT | TE XT | TE XT | P | H | PH | TOTAL |
|-----------------------|-------|----|----|----|----|----|----|---|----|----|---|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| LDSGNM(629)           | +     | -  | -  | -  | -  | +  | (+)| - | +  | (+)| - | +  | (+)| -  | -  | +  | (+)| -  | +  | (+)| -  | -  | +  | (+)| -  | -  | +  | +  | 6(3)| 2(2)| 4(3)| 10(0) |
| PRREPE(541)           | +     |    |   |   | +  | -  | +  | (+)| -  | - | +  | (+)| -  | +  | (+)| -  | - | +  | (+)| -  | +  | (+)| -  | -  | +  | +  | 5(1)| 4(2)| 11(1)|
| HASBERI(525)          | +     | (+)| -  | -  | -  | +  | (+)| - | +  | (+)| - | +  | (+)| -  | -  | +  | (+)| -  | +  | (+)| -  | -  | +  | +  | 9(1)| 7(3)| 5(1)| 4(1) |
| PRCCYNA(371)          | +     |    |    |    | +  | -  | -  | - | +  | (+)| - | +  | (+)| -  | -  | +  | (+)| -  | +  | (+)| -  | -  | +  | +  | 4(2)| 4(1)| 5(5)| 3(2) |
| LDSPISS(324)          | +     | (+)|   |   | -  | -  | -  | - | +  | (+)| - | +  | (+)| -  | -  | +  | (+)| -  | +  | (+)| -  | -  | +  | +  | 7(2)| 4(1)| 5(5)| 3(2) |
| MICUCU(289)           | +     |    |    |    | +  | -  | -  | - | +  | (+)| - | +  | (+)| -  | -  | +  | (+)| -  | +  | (+)| -  | -  | +  | +  | 11(1)| 7(3)| 5(1)| 4(1) |
| LDSFLM(271)           | +     |    |    |    | +  | -  | -  | - | +  | (+)| - | +  | (+)| -  | -  | +  | (+)| -  | +  | (+)| -  | -  | +  | +  | 9(1)| 7(3)| 5(1)| 4(1) |
| LDRUBR(202)           | -     |    |    |    | +  | -  | -  | - | +  | (+)| - | +  | (+)| -  | -  | +  | (+)| -  | +  | (+)| -  | -  | +  | +  | 4(2)| 4(1)| 5(5)| 3(2) |
| LDMICR(135)           | -     |    |    |    | +  | -  | -  | - | +  | (+)| - | +  | (+)| -  | -  | +  | (+)| -  | +  | (+)| -  | -  | +  | +  | 3(2)| 4(1)| 4(2)| 3(2) |
| PRNERI(124)           | +     |    |    |    | +  | -  | -  | - | +  | (+)| - | +  | (+)| -  | -  | +  | (+)| -  | +  | (+)| -  | -  | +  | +  | 5(5)| 7(3)| 5(1)| 4(1) |
| SEELON(114)           | +     |    |    |    | +  | -  | -  | - | +  | (+)| - | +  | (+)| -  | -  | +  | (+)| -  | +  | (+)| -  | -  | +  | +  | 4(2)| 4(1)| 5(5)| 3(2) |
| LSOLEI(110)           | +     |    |    |    | +  | -  | -  | - | +  | (+)| - | +  | (+)| -  | -  | +  | (+)| -  | +  | (+)| -  | -  | +  | +  | 7(3)| 7(3)| 5(1)| 4(1) |
| AUUMBE(106)           | +     |    |    |    | +  | -  | -  | - | +  | (+)| - | +  | (+)| -  | -  | +  | (+)| -  | +  | (+)| -  | -  | +  | +  | 3(2)| 7(3)| 5(1)| 4(1) |
| SEFASC(103)           | +     |    |    |    | +  | -  | -  | - | +  | (+)| - | +  | (+)| -  | -  | +  | (+)| -  | +  | (+)| -  | -  | +  | +  | 5(5)| 7(3)| 5(1)| 4(1) |
| AUPALL(66)            | +     |    |    |    | +  | -  | -  | - | +  | (+)| - | +  | (+)| -  | -  | +  | (+)| -  | +  | (+)| -  | -  | +  | +  | 7(3)| 7(3)| 5(1)| 4(1) |
| LDCORY(65)            | +     |    |    |    | +  | -  | -  | - | +  | (+)| - | +  | (+)| -  | -  | +  | (+)| -  | +  | (+)| -  | -  | +  | +  | 3(2)| 7(3)| 5(1)| 4(1) |
| AUCANC(53)            | +     |    |    |    | +  | -  | -  | - | +  | (+)| - | +  | (+)| -  | -  | +  | (+)| -  | +  | (+)| -  | -  | +  | +  | 2(2)| 2(2)| 3(1)| 3(2) |
| PRGRAN(52)            | +     |    |    |    | +  | -  | -  | - | +  | (+)| - | +  | (+)| -  | -  | +  | (+)| -  | +  | (+)| -  | -  | +  | +  | 2(2)| 2(2)| 3(1)| 3(2) |
| LDDAPH(48)            | +     |    |    |    | +  | -  | -  | - | +  | (+)| - | +  | (+)| -  | -  | +  | (+)| -  | +  | (+)| -  | -  | +  | +  | 2(2)| 2(2)| 3(1)| 3(2) |
| DSESS(42)             | +     |    |    |    | +  | -  | -  | - | +  | (+)| - | +  | (+)| -  | -  | +  | (+)| -  | +  | (+)| -  | -  | +  | +  | 2(2)| 2(2)| 3(1)| 3(2) |
| PRNANA(40)            | +     |    |    |    | +  | -  | -  | - | +  | (+)| - | +  | (+)| -  | -  | +  | (+)| -  | +  | (+)| -  | -  | +  | +  | 2(2)| 2(2)| 3(1)| 3(2) |
| SPCURV(25)            | +     |    |    |    | +  | -  | -  | - | +  | (+)| - | +  | (+)| -  | -  | +  | (+)| -  | +  | (+)| -  | -  | +  | +  | 2(2)| 2(2)| 3(1)| 3(2) |
| SPECIES       | ROUGH (## pixel obs.) | EL | EV | TE | VT | PP | TC | OS | T | HE | AT | U | JA | NM | LM | IN | M | A | PP | TC | MS | MD | WI | E | V | I | TE | TEXT | TEXT | TEXT | p | p | TOTAL |
|--------------|-----------------------|----|----|----|----|----|----|----|---|----|----|---|---|----|----|----|---|---|----|----|----|----|----|----|---|----|    |   |      | 11 | 3  | 2 | 0 | 0 | 3 | 1 | 2 | 11 | 0 | 7 | 0 | 4 | 8 | 3 | 3 | 3 | 0 | 2 | 3 |
| TOTAL(+)     | 1                     | 3  | 1  | 0  | 0  | 2  | 0  | 3  | 0 | 1  | 2  | 1 | 5 | 1 | 4 | 1 | 0 | 2 |   |   |   |
| TOTAL(-)     | 2                     | 2  | 2  | 0  | 3  | 2  | 3  | 2  | 7 | 2  | 12 | 4 | 1 | 1 | 1 | 5 | 1 | 4 | 3 | 3 | 1 | 2 |
| TOTAL(-)(+)  | 1                     | 1  | 2  | 0  | 2  | 0  | 1  | 2  | 11 | 0 | 7 | 0 | 4 | 8 | 3 | 3 | 3 | 0 | 2 | 3 |
| ALL +        | 12                    | 6  | 3  | 0  | 3  | 3  | 3  | 2  | 14 | 0 | 8  | 2 | 5 | 13 | 4 | 7 | 1 | 2 | 5 |   |
| ALL -        | 3                     | 4  | 0  | 5  | 2  | 4  | 4  | 7  | 4 | 13 | 5 | 2 | 2 | 6 | 3 | 5 | 2 | 3 |   |   |
Figure 2: Example of Boxplots of Species Level Environmental Coefficients (Boxes are arranged by the number of pixels where the species is observed from high to low.)
Table 3: Posterior Summary of the Coefficients of the Species Level Attributes ($\gamma$'s).

<table>
<thead>
<tr>
<th>Covariate</th>
<th>Mean</th>
<th>2.5%</th>
<th>50.0%</th>
<th>97.5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Height</td>
<td>-0.61</td>
<td>-2.14</td>
<td>-0.6</td>
<td>0.99</td>
</tr>
<tr>
<td>Locpop1 (1-50)</td>
<td>-4.83</td>
<td>-7.58</td>
<td>-4.84</td>
<td>-1.95</td>
</tr>
<tr>
<td>Locpop2 (50-1000)</td>
<td>-3.02</td>
<td>-5.67</td>
<td>-3.09</td>
<td>-0.05</td>
</tr>
<tr>
<td>Resprout (yes)</td>
<td>3</td>
<td>-0.04</td>
<td>3.01</td>
<td>5.93</td>
</tr>
<tr>
<td>Pollen1 (bird)</td>
<td>-0.79</td>
<td>-3.39</td>
<td>-0.77</td>
<td>1.71</td>
</tr>
<tr>
<td>Pollen2 (wind)</td>
<td>1.08</td>
<td>-2.45</td>
<td>1.06</td>
<td>4.75</td>
</tr>
<tr>
<td>Disp (wind)</td>
<td>1.95</td>
<td>-0.69</td>
<td>1.88</td>
<td>4.63</td>
</tr>
</tbody>
</table>
Table 4: Posterior Summary of the Area Level Attributes in terms of 
*Potential Richness* ($\sum_k \beta_{l}^{(k)}$).

<table>
<thead>
<tr>
<th>Covariate</th>
<th>Mean</th>
<th>2.5%</th>
<th>50%</th>
<th>97.5%</th>
<th>#:+::#-</th>
</tr>
</thead>
<tbody>
<tr>
<td>ROUGH</td>
<td>8.01</td>
<td>-0.45</td>
<td>8.02</td>
<td>16.51</td>
<td>12:3</td>
</tr>
<tr>
<td>ELEV</td>
<td>0.21</td>
<td>-12.37</td>
<td>0.29</td>
<td>12.35</td>
<td>6:4</td>
</tr>
<tr>
<td>POTEVT</td>
<td>5.03</td>
<td>-10.84</td>
<td>5.13</td>
<td>21.15</td>
<td>3:0</td>
</tr>
<tr>
<td>PPTCV</td>
<td>-11.12</td>
<td>-23.88</td>
<td>-11.03</td>
<td>1.46</td>
<td>(-)</td>
</tr>
<tr>
<td>FROST</td>
<td>11.45</td>
<td>-22.37</td>
<td>-11.63</td>
<td>0.52</td>
<td>(-)</td>
</tr>
<tr>
<td>HEATU</td>
<td>-14.19</td>
<td>-27.17</td>
<td>-14.11</td>
<td>1.88</td>
<td>(-)</td>
</tr>
<tr>
<td>JANMAXT</td>
<td>-2.16</td>
<td>-15.13</td>
<td>-2.27</td>
<td>11.63</td>
<td>2:4</td>
</tr>
<tr>
<td>JULMINT</td>
<td>16.74</td>
<td>1.78</td>
<td>16.86</td>
<td>31.7</td>
<td>+</td>
</tr>
<tr>
<td>MAP</td>
<td>-19.61</td>
<td>-32.91</td>
<td>-19.69</td>
<td>-5.47</td>
<td>-</td>
</tr>
<tr>
<td>PPTCON</td>
<td>-12.26</td>
<td>-25.33</td>
<td>-12.29</td>
<td>1.18</td>
<td>(-)</td>
</tr>
<tr>
<td>WINSMD</td>
<td>14.3</td>
<td>1.17</td>
<td>14.31</td>
<td>27.63</td>
<td>+</td>
</tr>
<tr>
<td>EVI</td>
<td>7.97</td>
<td>-0.49</td>
<td>7.88</td>
<td>16.45</td>
<td>(+)</td>
</tr>
<tr>
<td>FERT1</td>
<td>13.73</td>
<td>-0.9</td>
<td>13.72</td>
<td>27.29</td>
<td>(+)</td>
</tr>
<tr>
<td>FERT2</td>
<td>3.62</td>
<td>-7.56</td>
<td>3.49</td>
<td>15.6</td>
<td>3:2</td>
</tr>
<tr>
<td>FERT3</td>
<td>-4.44</td>
<td>-19.46</td>
<td>-4.11</td>
<td>8.79</td>
<td>1:2</td>
</tr>
<tr>
<td>TEXT1</td>
<td>-10.56</td>
<td>-24.42</td>
<td>-10.55</td>
<td>2.76</td>
<td>4:6</td>
</tr>
<tr>
<td>TEXT3</td>
<td>0.54</td>
<td>-10.45</td>
<td>0.49</td>
<td>11.87</td>
<td>7:3</td>
</tr>
<tr>
<td>TEXT4</td>
<td>-7.4</td>
<td>-19.28</td>
<td>-7.5</td>
<td>4.83</td>
<td>1:5</td>
</tr>
<tr>
<td>PH1</td>
<td>3.78</td>
<td>-10.62</td>
<td>3.42</td>
<td>17.85</td>
<td>2:2</td>
</tr>
<tr>
<td>PH3</td>
<td>-10.36</td>
<td>-21.18</td>
<td>-10.2</td>
<td>-0.18</td>
<td>-</td>
</tr>
</tbody>
</table>
Modeled Subregion
Full Model: *Protea cynaroides*

10 significant environmental covariates (betas)
Full Model: *Sorocephalus imbricatus*

Potential

Adjusted

Variance

0 significant environmental covariates (betas)
Full Model: *Protea restionifolia*

0 significant environmental covariates (betas)
PREVALENCE: MEAN PROBABILITY OF OCCURRENCE

\[ \sum_{i} [E(p_{i}^{(k)} | Y)] \]
Diversity Plot

Shannon-Weiner index

\[ e^{\exp\left\{ - \sum_{k=1}^{L} \frac{p_{i}^{(k)}}{\sum_{k} p_{i}^{(k)}} \log \frac{p_{i}^{(k)}}{\sum_{k} p_{i}^{(k)}} \right\}} \]
Similarity in Species Composition

\[
\exp \left( - \sum_{j \in \delta_i} \frac{||E(p_i|Y) - E(p_j|Y)||}{\text{number of neighbors of } i} \right)
\]
Spatial Random Effects $\rho_i$
Data thinning

• How does full model perform with less than a very rich data set?
• Data thinned randomly by sample point
Data Thinning: *Protea cynaroides*

![Graph showing data thinning at different percentages: 100%, 50%, 25%, 10%](image)
Data Thinning: *Aulax cancellata*

![Data Thinning: Aulax cancellata](image)

100%  50%  25%  10%

**Probability**
- 0 - 0.1
- 0.1 - 0.2
- 0.2 - 0.3
- 0.3 - 0.4
- 0.4 - 0.5
- 0.5 - 0.6
- 0.6 - 0.7
- 0.7 - 0.8
- 0.8 - 0.9
- 0.9 - 1
Data Thinning: Potential Richness

Pot. richness
0 - 3
3 - 6
6 - 9
9 - 12
12 - 15
15 - 18
18 - 21
21 - 24

100%
50%
25%
10%
Modeling of Presence-Only Data over Large Regions
Presence-only data

- We focus on presence-only setting
- Arguably more common, from museum databases and other non-systematic surveys
- One approach - convert presence-only problem to a presence/absence problem
- “background samples” - a random sample of known locations in the region with known environmental features where the species has not been observed
- Early on called “pseudo-absences” (work of Guisan, of Ferrier)
- How many???????
Recent work acknowledges that presence/absence is unknown for these background samples and attempts to adjust the binary regression accordingly.

Requires knowledge of overall population prevalence of species; can not be estimated from presence-only data.

Moreover, still, how many???

Also, non-spatial in the modeling.

The most illuminating discussion of all of this is in Ward et al. (2009, Biometrics)
A bit more detail

- Assume observed presences are a random sample of all presences
- Assume background samples provide a good representation of the landscape
- Assume know overall prevalence, say \( \pi \), and choose \( n_u \) background samples with \( n_p \) observed presences
- Naive model assumes all background samples are absences, then a (nonspatial) logistic regression. Underestimates the probability of presence given environmental features
- With \( \pi \) and \( n_u \) can “correct” the logistic regression
- With covariates \( x \), logit of probability of presence given \( x \) is revised from \( \eta(x) \) to \( \eta(x) + \log\left(\frac{n_p + \pi n_u}{\pi n_u}\right) \)
Other approaches

- Algorithmic strategies
  - The two most common are GARP (genetic algorithm for rule set prediction) and Maxent
  - GARP is based upon an AI framework - a set of positive and negative rules that give a binary prediction. Rules are favored according to their effectiveness, relative to random prediction, based upon a sample of background and presence data.
  - Maxent is a constrained optimization which we describe in more detail on next slide
  - GARP was most widely used due to available software. But, Maxent predictions typically show higher predictive accuracy
  - With recent attractive software (Schapire at Princeton), Maxent is becoming standard for such analysis.
Some Maxent details

- Maxent is a constrained optimization solution
- Suppose locations \( \{s_i\} \) with associated environmental variables \( x(s_i) \)
- Obtain maximum entropy distribution (closest to uniform), \( p(s) \), i.e. \( \max \int p(s) \log p(s) ds \) subject to \( E_{p(s)}(x(s)) = \bar{x} \)
- \( p(s) \) is a relative density; nothing regarding probability of presence at \( s \)
- Regularization - \( |E_{p(s)}(x_j(s)) - \bar{x}_j| \leq \epsilon_j \) with weights \( \lambda_j \)
- Duality in terms of Gibbs distributions
- \( \lambda \)'s are not regression coefficients
- No uncertainty
- Sampling effort correction - \( p(s) \pi(s) / \int p(s) \pi(s) ds \)
The “right” perspective

- Presence-only data should be viewed as a point pattern (Also, Warton and Shepherd, 2010)
- So, goal is to model the intensity as a function of the available environments across the region
- We do this through typical regression modeling, enabling natural interpretation for the coefficients (quite different from the regularization coefficients that are employed in the constrained Maxent optimization).
- Use a hierarchical model to enable uncertainty in the inference, especially with regard to the intensity surface.
- Can normalize this intensity to a density over the region
- Maxent is limited to such a normalized probability density but can not be recover the intensity.
Introduce spatial structure into our modeling for the intensity surface through spatial random effects.

Do not assume any background samples; rather, we assume that the covariates we employ are available as surfaces over the region so that we can interpolate an intensity over the entire region.

Observed point pattern is biased through anthropogenic processes, e.g., human intervention to transform the landscape and non-uniform (in fact, often very irregular) sampling effort.

Adjust the potential species intensity to a realized intensity, a degradation to the intensity.
Conceptualizing the model

- Again, view the observed presence-only data as a point pattern subject to degradation.

- Express realization of point process over $D$ as $Y_D = \{Y(s)|s \in D\}$, where $Y(s)$ is 1/0 - observed or not at $s$.

- For any bounded region $D$, $Y_D$ consists of finitely many 1's and uncountably many 0's. Hence, inferring about $p(Y(s) = 1|s)$ is not possible with this type of data.

- Background samples methods “cheat”, choose a finite number from this uncountable “absence” set.

- Conditioning in wrong direction, i.e., not mark(presence/absence) given location but location given mark. With presence-only data, can only model location given presence.

- So, we can infer about the intensity for presences, hence density $p(s|y = 1)$ from $Y_D$. 
More conceptualization

- We assume a nonhomogeneous Poisson process model for the set of presence locations.

- We introduce degradation caused by sampling bias as well as by land transformation.

- As a result, we conceptualize a *potential* intensity, i.e., the intensity in the absence of degradation, as well as a *realized* (or effective) intensity that operates in the presence of degradation.

- We work at grid cell level (minute × minute) due to the resolution of our environmental covariates

- So, we tile the intensity to reflect our inability to explain it at spatial resolution finer than our grid cells.
More conceptualization

- We begin by imagining three surfaces over $D$.
- Let $\lambda(s)$ be the potential intensity surface, i.e., a positive function which is integrable over $D$.  
- Let $\int_D \lambda(s) \, ds = \lambda(D)$. Then, $g(s) = \lambda(s) / \lambda(D)$ gives the potential density over $D$.
- Modeling for $\lambda(s)$ will be provided below.

- Next, an availability surface, $U(s)$, a binary surface over $D$ such that $U(s) = 1$ or $0$ according to whether location $s$ is untransformed or not.
- Assuming no sampling bias, $\lambda(s)U(s)$ can only be $\lambda(s)$ or $0$ according whether $s$ is available or not.
- If we average $U(s)$ over grid cell $A_i$, we obtain $\int_{A_i} U(s) \, ds / |A_i| = u_i$, the proportion of cell $A_i$ available.
- $u_i$ is known for all grid cells.
Untransformed Proportions
Similarly, a sampling effort surface over $D$, $T(s)$.

$T(s)$ is also a binary surface and $T(s)U(s) = 1$ indicates that location $s$ is both available and sampled.

Set $q_i = \int_{A_i} T(s)U(s) ds / |A_i|$. $q_i$ is probability that a randomly selected location in $A_i$ was available and sampled.

Model availability and sampling effort at areal unit scale.

Hence, $\lambda(s)U(s)T(s)$ becomes the degradation in intensity at location $s$ - can never be greater than $\lambda(s)$.

In regions where no locations were sampled, the operating intensity for the species is 0.
Sampled Cells
cont.

- **Note**: We do *not* envision a probability density surface for sampling effort as in Maxent literature. With such a surface, say \( p(s) \), we could have \( \lambda(s)p(s) > \lambda(s) \) which does not seem sensible.

- Sensible to multiply an intensity by an intensity or a density by a density (renormalizing the product) but not an intensity by a density.

- In different terms, \( \int_{A_i} T(s)ds / |A_i| \) can be viewed as the sampling probability associated with cell \( A_i \).

- If \( T(s) \) is viewed as random, so is the integral.

- Expectation of the integral would yield \( \int_{A_i} p(s)ds / |A_i| \) where, now, \( p(s) = P(T(s) = 1) \in [0, 1] \).

- \( p(s) \) gives the local probabilities of sampling, not a probability density over \( D \).
A remark

- It is tempting to try to model $T(\cdot)$ using a logistic regression with variables such as distance to nearest road, distance to nearest city, distance to national park, etc. to learn about $p(s)$.
- Would provide a possible prior for $T(s)$.
- We have examined this approach for the CFR but the resulting regression performed poorly.
- Furthermore, while we have a conceptual point-level model, for the CFR, we work at grid cell level, i.e., with $q_i$. Very difficult to induce a prior for $\{q_i\}$ from a specification for $T(s)$. 
Still conceptualizing

Assume $\lambda(s)$ independent of $T(s)U(s)$, i.e., potential intensity for a species is independent of degradation process.

Then, $\int_{A_i} \lambda(s)T(S)U(s)ds = \lambda_i q_i$ where $\lambda_i = \int_{A_i} \lambda(s)ds$ is the cumulative intensity associated with cell $A_i$

Again, $q_i = \int_{A_i} T(s)U(s)ds/|A_i|$. 

Not sensible that sampling effort independent of land transformation. If $U(s) = 0$ then $T(s) = 0$.

If $q_i = u_i p_i$, then $p_i = \frac{\int_{A_i} T(s)U(s)ds}{\int_{A_i} U(s)ds}$

$p_i$ is the conditional probability that a randomly selected location in cell $A_i$ is sampled given it is available.

We set $p_i$ equal to 1 or 0 which we interpret as $T(s) = U(s) \forall s \in A_i$ or $T(s) = 0 \forall s \in A_i$, respectively.
Likelihood and posterior

- We model the potential intensity surface $\lambda(\cdot)$ using a log GP prior, which results in a log Gaussian Cox process model for the observed data.

- We expect the environmental covariates to influence the species abundance, thus a regression model for the mean of the GP in the $x'$s, i.e., $s \in D$, we have

  $$\log \lambda(s) = x(s)\beta + w(s)$$

- $w(\cdot)$, a zero-mean stationary, isotropic GP over $D$, accounts for spatial association in $\lambda$ surface across grid cells beyond the similarity of $x$ values.

- We use the exponential covariance function for $w(\cdot)$. 
Suppose we have $n_i$ presence locations $(s_{i,1}, s_{i,2}, \ldots, s_{i,n_i})$ within cell $i$ for $i = 1, 2, \ldots, I$.

From above, $U(s_{i,j})T(s_{i,j}) \equiv 1$, $0 \leq j \leq n_{i,j}, 1 \leq i \leq I$.

Then the likelihood function corresponding to NHPP($\lambda(\cdot)$) becomes

$$L(\lambda(\cdot); \{s_{i,j}\}) \propto e^{-\int_D \lambda(s)U(s)T(s)ds} \prod_{i=1}^I \prod_{j=1}^{n_i} \lambda(s_{i,j})$$

We have finitely many presence locations but the integral term in $L$ involves the uncountable random field $\{\lambda(s) : s \in D\}$. 
A grid cell approximation

- The dataset consists of observed presence locations at the scale of grid cells in the CFR.
- $D$ is divided into $I = 36,907$ grid cells of equal area.
- For each cell $i = 1, 2, 3, ..., I$, we are given information on $p$ covariates as $x_i = (x_{i1}, x_{i2}, ..., x_{ip})$.
- Cell-level information about land availability across $D$, as a proportion of the area of the cell, $u_i$.
- Within cell $i$, $n_i$ sites where the species was observed.
- For many cells $n_i = 0$ primarily because a large fraction (72%) were actually unsampled.
cont.

Don’t work at the site level. Geo-coded locations for the observed sites, with covariate information at grid cell level. So, explain the point pattern at grid cell level. Can not learn about covariate effects at finer resolution.

A computational advantage - work with a product Poisson likelihood approximation.

Let us denote the center and area of each grid cell by \((s_i, \Delta)\) respectively for \(i = 1, 2, ..., I\)

Approximate \(\lambda(s)\) with \(\lambda(s_i)\) for \(s \in A_i\).

Rewrite the integral \(\int_D \lambda(s)U(s)T(s) \, ds\) as \(\sum_{i=1}^I \int_{s \in A_i} \lambda(s)U(s)T(s) \, ds\).
Under the above approximation, with \( \lambda(s) \) still as above, the likelihood becomes

\[
\tilde{L}(\lambda(\cdot); n) \propto e^{-\sum_{i=1}^{I} \lambda(s_i) \Delta q_i} \prod_{i=1}^{I} \lambda^{n_i}(s_i)
\]

If \( q_i = 0 \) (i.e., either \( p_i = 0 \) or \( u_i = 0 \)) there is no contribution from \( A_i \) in \( \tilde{L} \).

If only \( m \) of \( I \) cells sampled, \( \tilde{L} \) involves only \( \lambda(s_{1:m}) \).

Inference on \( \lambda(s_{m+1:I}) \) can be drawn from their respective posterior predictive distributions.
cont.

Since, \( \log \lambda(s) \) follows a GP, the posterior distribution corresponding to \( \tilde{L} \) takes the form

\[
\pi(\lambda(s_{1:m}), \beta, \theta | n, x, u, p) \propto e^{-\sum_{i=1}^{m} \lambda(s_i)q_i} \Delta \prod_{i=1}^{m} \lambda^{n_i}(s_i)
\times \phi_m(ln \lambda(s_{1:m}) | \beta, x, \theta) \pi(\beta) \pi(\theta)
\]

\( \phi_m \) is \( m \) dimensional Gaussian density, \( \theta \) parameters in covariance function of \( w(\cdot) \)
Richness with presence-only data

- For a set $A$ within the study region, the richness for $A$, $R(A)$ is the expected number of distinct species in $A$ - we expect more species as $A$ grows larger and no species as the area of $A$ goes to 0.

- With Maxent, sum over the individual species densities, integrate over $A$. Interpretation of this as a richness?

- No uncertainty can be attached to this sum.

- Our modeling provides a natural parametric function for inferring about richness, with associated uncertainty.

- Suppose $(s_i, l(s_i)), i = 1, 2, ..., n$, i.e., a random location and a species label associated with that location.

- Use above modeling to create a species intensity, $\lambda_l(s)$, for species $l = 1, 2, ..., L$. 
Richness cont.

- \( n(A) \) is total number of observations in \( A \), i.e., total number of locations in \( A \) where a “presence-only” observation of any species was recorded.

- Let \( n_l(A) \) be the number of locations in \( A \) where species \( l \) was observed.

- Let \( r(A) = \sum_l 1(n_l(A) > 0) \). Then \( r(A) \) is the “observed” richness in \( A \) where \( r(A) \) is random for each \( A \).

- We seek to infer about \( R(A) = E(r(A)) \).

- \( E(1(n_l(A) > 0)) = 1 - \exp(-\lambda_l(A)) \) since \( n_l(A) \sim Po(\lambda_l(A)) \).

- Hence, \( R(A) = E(r(A)) = \sum_l (1 - \exp(-\lambda_l(A))) \).

- Richness is not additive, i.e.,
  \( E(r(A_1 \cup r(A_2)) \neq E(r(A_1)) + E(r(A_2)) \).
Richness cont.

- With model fitting for each $\lambda_l(s)$, we can obtain posterior samples of $E(r(A))$ for any $A$.
- If we work with the collection of grid cells $A_i$, we can supply a richness surface for the entire CFR.
- Adjusting for transformation and sampling intensity allows a potential and realized richness surface.
Model fitting

- First, a computational remark
- With 37,000 grid cells we have a “big n” problem
- Can handle with a Gaussian process with locations at centers of grid cells. We handle this with the predictive process approximation (Banerjee et al, 2008, Finley et al, 2009) with bias adjustment.
- Approach: Replace $w(s)$ by a process $\tilde{w}(s)$ that will project the model into a smaller subspace
- An alternative - use a CAR model implemented with parallelization (Chakraborty et al., 2010)
The actual model

We have the following hierarchical model

\[ n_i | \lambda_i \overset{\text{ind}}{\sim} \text{Poi}(\lambda_i q_i), \ i = 1, 2, \ldots, I \]
\[ \log \lambda(s_i) = x_i \beta + \tilde{w}(s_i) + \epsilon_i^* \]
\[ \tilde{w}(s_i) = R_{i,m}^T(\phi) R_{m}^{-1}(\phi) w(s_m^0) \]
\[ w(s_m^0) \sim N_m(s_m^0, R_m(\phi)) \]
\[ \epsilon_i^* \overset{\text{ind}}{\sim} N(0, \sigma^2 (1 - (R_{i,m}^T(\phi) R_{m}^{-1}(\phi) R_{i,m}(\phi)))) \]
\[ \pi(\beta, \phi, \sigma^2) = \pi(\beta) \pi(\phi) \pi(\sigma^2) \]
The data analysis

- Six different plant species over the whole CFR.
- Centered and scaled all the $x$’s
- Assume a 0/1 sampling effort at resolution of grid cells.
- The datasets were handled efficiently using C++ with Intel math kernel library
- Run 15,000 iterations of MCMC, 5,000 burn-in and thinning the rest at every fifth sample.
- Table for the posterior mean and 95% credible interval for covariate effects for all species.
- Figure for the mean posterior mean of the spatial effects for all 6 species
- Figures for the posterior intensities (potential and transformed) for the 6 species
Covariate effects

<table>
<thead>
<tr>
<th>Species</th>
<th>APAN.MEAN</th>
<th>MAX01</th>
<th>MIN07</th>
<th>MEAN.AN.PR</th>
<th>SMDSUM</th>
<th>FERT1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protea</td>
<td>-4.909</td>
<td>2.702</td>
<td>-0.3014</td>
<td>-1.222</td>
<td>-0.04872</td>
<td>0.5005</td>
</tr>
<tr>
<td>Aure</td>
<td>(3.449)</td>
<td>(2.104)</td>
<td>(1.392)</td>
<td>(1.552)</td>
<td>(1.527)</td>
<td>(0.9327)</td>
</tr>
<tr>
<td>Protea</td>
<td>-2.447</td>
<td>1.268</td>
<td>-1.032</td>
<td>-0.8328</td>
<td>0.5524</td>
<td>0.802</td>
</tr>
<tr>
<td>Cyna</td>
<td>(1.122)</td>
<td>(0.7655)</td>
<td>(0.845)</td>
<td>(0.3948)</td>
<td>(0.5752)</td>
<td>(0.3145)</td>
</tr>
<tr>
<td>Protea</td>
<td>0.721</td>
<td>-0.4204</td>
<td>0.1368</td>
<td>-0.3759</td>
<td>0.4878</td>
<td>0.09871</td>
</tr>
<tr>
<td>Ldsg</td>
<td>(0.7112)</td>
<td>(0.4768)</td>
<td>(0.3052)</td>
<td>(0.2759)</td>
<td>(0.369)</td>
<td>(0.1069)</td>
</tr>
<tr>
<td>Protea</td>
<td>-0.2194</td>
<td>0.0277</td>
<td>-0.6087</td>
<td>-0.1986</td>
<td>1.082</td>
<td>0.5066</td>
</tr>
<tr>
<td>Mund</td>
<td>(4.154)</td>
<td>(2.865)</td>
<td>(0.9844)</td>
<td>(1.534)</td>
<td>(1.531)</td>
<td>(0.8282)</td>
</tr>
<tr>
<td>Protea</td>
<td>2.076</td>
<td>-1.59</td>
<td>-1.722</td>
<td>0.3625</td>
<td>0.5354</td>
<td>0.1864</td>
</tr>
<tr>
<td>Punc</td>
<td>(2.065)</td>
<td>(1.37)</td>
<td>(0.6391)</td>
<td>(0.5804)</td>
<td>(1.027)</td>
<td>(0.395)</td>
</tr>
<tr>
<td>Protea</td>
<td>1.69</td>
<td>-1.206</td>
<td>-0.2754</td>
<td>0.124</td>
<td>0.09401</td>
<td>0.2241</td>
</tr>
<tr>
<td>Repe</td>
<td>(0.881)</td>
<td>(0.5913)</td>
<td>(0.3215)</td>
<td>(0.2896)</td>
<td>(0.4319)</td>
<td>(0.1428)</td>
</tr>
</tbody>
</table>
Spatial effects
Intensities – potential, degraded
Intensities – potential, degraded