LATENT SPATIAL MODELING FOR SPECIES ABUNDANCE

Avishek Chakraborty    Alan E. Gelfand
DSS, Duke University
Joint Work with J. A. Silander, A. Latimer & A. Wilson

March 27, 2009
Study region:: Cape Floristic Region (CFR) of South Africa

- Encompasses an area of around 90000 km²
- Rich in biodiversity; existence of \( \sim 9000 \) plant species
Objective

- Construction of map of species abundance over whole CFR
- Understanding the effect of climate and soil-type factors on species existence
- Analysis of the impact of land transformation on species abundance
- Determining possible areas and strategies for species conservation
Analysis of species distribution in Gelfand et al. (2003)

On a smaller domain (Kogelberg Hawequas subregion of ~1600 cells)

Worked with a 0/1 presence-absence type dataset

Fitted a two-stage hierarchical suitability/presence spatial model at areal level
To work with (ordinal) categorical abundance information rather than binary presence/absence
- Taking into account possible uncertainty in data reporting
- Dealing with the effect of land transformation statistics
- To study the species distribution on entire CFR
CFR consists of \( \sim 37000 \) 1min rectangular gridcells of approx. \( \sim 1.55 \times 1.85 \) km\(^2\).

Within cell \( i \), categorical abundance information collected at \( n_i \) many sampling sites.

Each site has a reported abundance category \( y_{ij} \in \{0, 1, 2, 3\} \), \( j = 1, 2, \ldots, n_i \).

\( y \) is an ordinal categorization of actual counts, which may range from 0 to thousands.
Want to assess the effect of environmental factors on the abundance pattern

Important covariates related to temperature, precipitation, soil-type

Covariate choice based on plant physiology and previous modelling

All covariate information \((x)\) given also at the resolution of 1min cells
Introduction

Data Description

Model

Computation

Analysis

Key feature :: large number of cells with $n_i = 0$ sampled sites.

Sites located within 10158 cells; around 28% of whole CFR

Sampling density varies with factors like distance from city/road, presence of a reserve etc.

Figure: (L) sampled cells (R) sampling distribution
For each cell, also given proportion of land transformed $(1 - u)$ within.

For a transformed location $y$ must be 0

Including $u$ in the model essential to explain high number of 0 reporting.

**Figure:** Transformation Map
Objective

- To construct *Potential Abundance* (PA) distribution \( p_i = (p_i(0), p_i(1), p_i(2), p_i(3)) \) for all cells within CFR
- PA conceptualized as abundance distribution in absence of land transformation, reporting inaccuracies and any other source of error
  - should be governed by environmental conditions and have spatial dependence
- Want to use latent layers ("abundance score") rather than discrete categorical distributions
Assumptions

- Within each cell PA distribution is same for all locations, \( p_i = p_s \) if \( s \in \text{cell} \ i \)
  - Can be viewed as block average \( p_i(k) = \frac{1}{|\text{cell} \ i|} \int_{\text{cell} \ i} p_s(k) \, ds \)
  - Can’t expect to learn point level distributions with areal level covariates

- Transformation is independent of species occurrence and reporting inaccuracies
At site \( j \) within cell \( i \), \( P(y_{ij} = k) = q_i(k) \) for \( 0 \leq k \leq 3 \)

In absence of measurement error,
\[
q_i(k) = (1 - u_i)1_{k=0} + u_ip_i(k),
\]
where \( 100 \times (1 - u_i)\% \) land is transformed within cell \( i \)

Model transition from \( p \) to \( q \) to be affected by both transformation and noise

Introduce the transition in latent variable scale rather than probability scale
Introduce $z^p$ (*PA score*) as underlying continuous variables for PA distribution $p$

- $\{z^p(s), s \in D\}$ can be thought of potential abundance surface for a species

- $z^p$ distribution is discretized at cut points $\alpha$ to generate $p$, $p(k) = P(\alpha_{k-1} \leq z^p \leq \alpha_k)$
During data collection, a noisy version of $z^p$, say $z^q$ (centered at $z^p$) is recorded.

- The $z^q$ distribution is categorized again at $\alpha$’s to generate $q$.
- With zero noise, in absence of transformation, should have $p = q$.
- In practice, noise $\equiv$ under/over reporting of categories or missing an occurrence.
y non informative about how to separate $z^p$, $z^q$ surfaces

Ordinary MEMs like following not meaningful due to unidentifiability of scale

$$z^q \sim N(\cdot; z^p, \cdot)$$
$$z^p \sim N(\cdot; \mu, \cdot)$$

In above, one may marginalize and just work with $z^q$
We have a constraint on error ::

- Can't record a presence (at any category) if it was not there.
- \( z^p \leq \alpha_0 \Rightarrow z^q \leq \alpha_0 \), cant use a ordinary \( \phi(z^q|z^p) \)

Instead specify

\[
f(z^q|z^p, \sigma^2) = \phi(\cdot; z^p, \sigma^2)1_{z^p \geq \alpha_0} + \frac{\phi(\cdot; z^p, \sigma^2)1_{(-\infty, \alpha_0)}}{\Phi(\alpha_0; z^p, \sigma^2)}1_{z^p \leq \alpha_0} \tag{1}
\]

- Fits the error constraint
- In place of truncated normal can use any distribution on \( \mathcal{R}^- \), but like to
  - get identical \( z^q, z^p \) with \( \sigma^2 = 0 \)
  - interpretation of \( z^q \) surface to be actually centered around \( z^p \) surface
Can meaningfully talk about two layers
- $z^q$ is the non-gaussian surface corresponding to the data
- $z^p$ extracts the gaussian part of the surface
- Difference between the surfaces controlled by the parameters

In presence of transformation, modify the previous equation as

$$f(z^q | z^p_{\geq 0}, \sigma^2) = u \phi(\cdot; z^p, \sigma^2) + (1 - u) \frac{\phi(\cdot; z^p, \sigma^2)\mathbf{1}_{(-\infty, \alpha_0)}}{\Phi(\alpha_0; z^p, \sigma^2)}$$

Alternatively, separate transformation, then introduce $z^p, z^q$ and work with (1) (no change in model summaries)
This specification supports inflated number of zero abundances

- For \( z^p \) we work with \( f(z^p) = \phi(\cdot; x^T \beta + \theta, \tau^2) \)

- \( \theta \) is the set of spatial random effects

- A CAR prior structure is assumed for \( \theta \) with a prefixed neighborhood matrix \( W \).
Graph

\[ x \xrightarrow{\beta} z^p \xrightarrow{\sigma^2} z^q \xrightarrow{\alpha} q \xrightarrow{\alpha} y \]

\[ \theta \xrightarrow{\alpha} p \]

\[ \text{Joint Work with J. A. Silander, A. Latimer & A. Wilson} \]

Avishek Chakraborty, Alan E. Gelfand  DSS, Duke University

LATENT SPATIAL MODELING FOR SPECIES ABUNDANCE
Assume sampled sites are within first $m$ of total $s$ cells

Define $\Theta = (\alpha, \beta, \sigma^2, \tau^2)$

$L(y|z^q, \alpha, u) = \prod_{k=0}^{3} [1_{z^q \in (\alpha_{c-1}, \alpha_c)}]^{1(y=k)}$. 

\[
\pi(z^p, z^q, \Theta | \cdot) \propto \prod_{i=1}^{m} \prod_{j=1}^{n_i} L(y_{ij} | z^q_{ij}, \cdot) p(z^q_{ij} | z^p_{ij}, \cdot) p(z^p_{ij}) \pi(\theta_{1:s}) \pi(\Theta) \]

\[
\pi(\theta_{1:s} | \eta) = CAR(\eta_0, W) \]

\[
\pi(\Theta) = \pi(\alpha) \pi(\beta) \pi(\sigma^2) \pi(\tau^2) \]
Due to latent nature of $z^p, z^q$, not possible to simultaneously learn all elements of $\Theta$.

Mean will be interpretable only up to fixed scale.

As a convention, assume $\alpha_0 = 0$.

Scale parameters like $\sigma, \tau$ need to be fixed.

- Extent of nongaussianity in the observed data a function of $(\mu, \sigma, \tau)$.
- Relative to fixed $\sigma^2, \tau^2, \mu$ an indicator of the departure of $z^q$ surface from $z^p$.
- Estimation of mean surface not of interest in absolute scale, but for comparison.
- For any $\beta$ sign is important, not the magnitude.
Only nontrivial part is to sample from \( \pi(z^q, z^p|\cdot, \cdot) \)

For an observed \( y > 0 \) simulate \( \pi(z^q|z^p, \cdot), \pi(z^p|z^q, \cdot) \), truncated normals within appropriate intervals

A 0 value of \( y \) can occur in 3 ways

(i) the species was not there originally; prior prob. \( \pi_1 = P(z^p \leq 0) \)

(ii) untransformed, but the species could potentially be there, but the location is transformed; prior prob. \( \pi_2 = (1 - u)P(z^p \geq 0) \)

(iii) untransformed location, the species was there, but missed it during data collection; prior prob. \( \pi_3 = uP(z^p \geq 0, z^q \leq 0) \)
Conditional on \( y = 0 \), sample the possible event using Bayes rule.

For case (i), (ii) sample using \( \pi(z^p | \cdot), \phi(z^q | z^p, \cdot) \).

Case (iii) can be sampled

- either by an M-H step for \( z^p | \cdot \) followed by sampling \( \pi(z^q | z^p, \cdot) \)
- OR a Gibbs-within-Gibbs framework of drawing \( \pi(z^p | z^q, \cdot) \) and \( \pi(z^q | z^p, \cdot) \)
Computational Issues

- Have to introduce a \((z^q, z^p)\) pair for each site; \(\sim 50000\) are there
- Independence of \((z^p, z^q)\) across sites enables block updating
- Computationally demanding is sequential update of spatial random effects
- Parallel updating scheme of dividing the region into disjoint sets along with a set of "boundary cells" based on \(W\)
Worked so far with six species

Want to look at following diagnostics

- Covariate effects
- Potential abundance classification probabilities
- Transformation adjusted probabilities
- Uncertainty in estimating probability distribution
Covariate Effect

Avishek Chakraborty, Alan E. Gelfand  DSS, Duke University  LATENT SPATIAL MODELING FOR SPECIES ABUNDANCE
Covariate Effect
Covariate Effect

- Avishek Chakraborty, Alan E. Gelfand
- DSS, Duke University
- Joint Work with J. A. Silander, A. Latimer & A. Wilson
- LATENT SPATIAL MODELING FOR SPECIES ABUNDANCE
Abundance Distribution (prrepe)
Abundance Distribution (prcyna)
Abundance Distribution (prrepe)
Uncertainty
Future work on this includes ::

- Constructing similar abundance maps for broader family of species
- To come up with a unified measure of abundance rather than reporting categorical probabilities
- Extending the model to bivariate data to analyze competing or co-existent family of species

Reference ::