

Zero-inflated Bayesian Spatial Models with Repeated Measurements

Jing Zhang

Miami University

*IMS New Researchers Conference
07-28-2009*

Acknowledgement

- Dr. Chong He, University of Missouri-Columbia;
- Dr. Xiaoqian Sun, Clemson University;
- Dr. John Kabrick, USDA Forest Service, North Research Station;
- Missouri Department of Conservation (MDC).

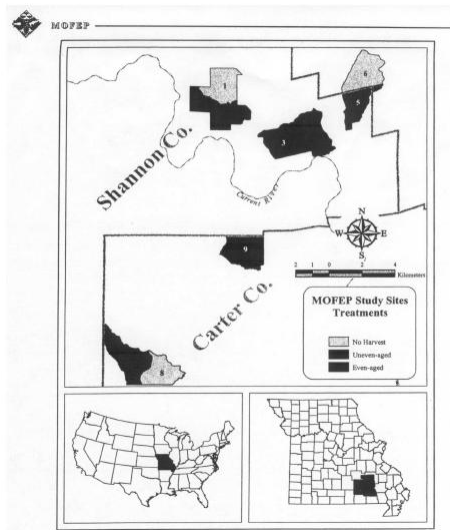
Outline

1. Missouri Ozark Forest Ecosystem Project (MOFEP)
2. Background Knowledge
2. Proposed Model and Inference
5. Data Analysis
6. Future Work and Summary

Missouri Ozark Forest Ecosystem Project

- Ozarks Highlands in Southeastern Missouri is an environmentally heterogeneous area;
- MOFEP was initiated in 1989 by the MDC;
- Monitor and assess the short and long-term effects of common management practices on Ozark ecosystems;
- Management practices: no-harvest, even-aged, and uneven-aged management;
- Short and long-term: season, year, and 100 years;
- A wide array of ecosystem attributes: soil, insects, invertebrates, birds, mammals, trees, and ground covers, etc; (32 different studies)
- Data is collected from 9 study sites ranging in size from 265-530 ha.

Study Sites in MOFEP



Questions That Biologists Are Interested In:

- How to understand the biology of a (ground flora) species?
- What is the associations between (ground flora) species?
- What are the species-specific response to the environment?
- Can we predict the spatially explicit ecological processes?
- How to predict the species distribution in a certain area?

Translated Into Statistics:

Information: point-referenced data with covariates;

Objective 1: analyze the spatial dependence among observations;

Objective 2: prediction at unmeasured locations;

Objective 3: inference about covariates effect.

Feedback From The Biologists:

1. How to predict the spatial distribution of ground flora on large domains? (Hooten, Larsen and Wikle (2003))

Product: a black/white prediction map, mapping the area with ground flora.

2. How to predict the spatial distribution of ground flora on large domains? (Sun, He, Zhang and Kabrick (2008))

Product: a colored prediction map, mapping the area with predicted coverage of ground flora.

Feedback From The Biologists:

3. How to predict the spatial distribution of a certain species of ground flora on large domains? (Zhang and He (2008))

Product: a colored prediction map, mapping the area with predicted coverage of a certain species.

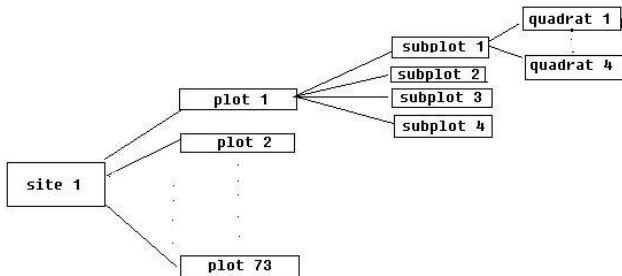
— We focus on this one in this presentation.

4. How to predict the spatial distribution of multiple species of ground flora on large domains simultaneously? (Zhang and He, in preparation)

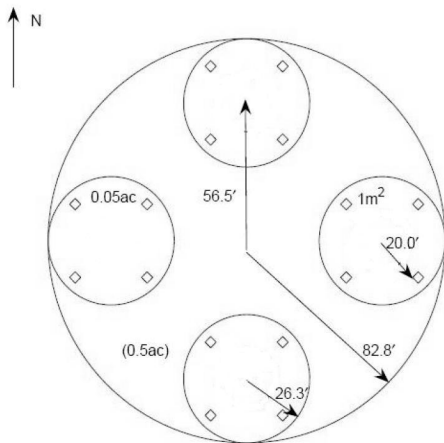
Product: a colored prediction map, mapping the area with predicted coverage of multiple species ground flora.

Vegetation Coverage Sampling

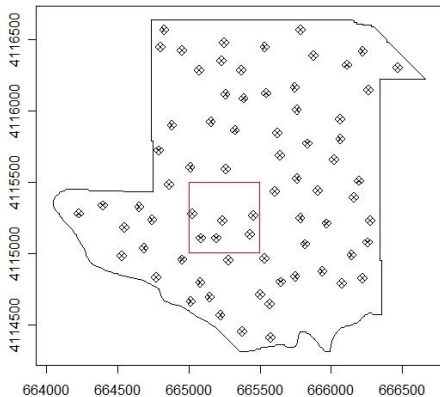
- MOFEP ground flora was sampled from the 70-76 0.5 ac circular plots located within each of the nine study sites;
- Four 0.05 ac subplots in each plot;
- Four $1 - m^2$ quadrats in each subplot;
- Vegetation proportions are collected in each $1 - m^2$ quadrat for selected subplots;



Design of Subplots



Sampled Locations In Site 1:



Motivation of Study:

- Quadrats are very close to each other; — Dangerous to treat quadrats as locations!
- Resolution of covariate information is $10m \times 10m$. — Not available for each quadrat!
- Coverage for individual species (one-species data): excess zeroes and 0.001s due to the limit of detection. — no suitable standard models!

Motivation of Study:

- Quadrats are very close to each other; — Dangerous to treat quadrats as locations!
- Resolution of covariate information is $10m \times 10m$. — Not available for each quadrat!
- Coverage for individual species (one-species data): excess zeroes and 0.001s due to the limit of detection. — no suitable standard models!

Solutions:

- Treat the centers of subplots as locations;
- Treat the measurements of four quadrats nested in the same subplot as repeated measurements;
- Design a Bayesian hierarchical model to account for variation from the spatial random effect, correlation among repeated measurements and zero-inflated observation.

Proposed Model and Inference

Notations:

- $y_1(s_1), \dots, y_r(s_n)$: r left-censored repeated measurements collected from n different locations: s_1, \dots, s_n ;
- $y_j^*(s_i)$: j th unobserved true value in location s_i ;
- $X = (\mathbf{x}(s_1), \dots, \mathbf{x}(s_n))'$: design matrix of covariates.

Censoring mechanism:

$$y_j(s_i) = \begin{cases} 0, & -\infty \leq y_j^*(s_i) < p_1, \\ \alpha, & p_1 \leq y_j^*(s_i) < p_2, \\ y_j^*(s_i), & p_2 \leq y_j^*(s_i) < +\infty. \end{cases}$$

Note:

- $0 < \alpha < \infty$;
- $y_j^*(s_i)$: proportional rates, or continuous values.

Proposed Model and Inference

Transformation of the unobserved true values: $U_j(s_i) = g(y_j^*(s_i))$.

$$y_j(s_i) = \begin{cases} 0, & U_j(s_i) < C_1, \\ \alpha, & C_1 \leq U_j(s_i) < C_2, \\ g^{-1}(U_j(s_i)), & U_j(s_i) \geq C_2. \end{cases}$$

where $C_1 = g(p_1)$, $C_2 = g(p_2)$.

Proposed Model and Inference

Transformation of the unobserved true values: $U_j(s_i) = g(y_j^*(s_i))$.

$$y_j(s_i) = \begin{cases} 0, & U_j(s_i) < C_1, \\ \alpha, & C_1 \leq U_j(s_i) < C_2, \\ g^{-1}(U_j(s_i)), & U_j(s_i) \geq C_2. \end{cases}$$

where $C_1 = g(p_1)$, $C_2 = g(p_2)$.

Assumption: $Y_j(s_i)$ are independent given \mathbf{U} ;

Likelihood:

$$\begin{aligned} [\mathbf{Y}|\mathbf{U}] &= \prod_{i=1}^n \prod_{j=1}^r [I(Y_j(s_i) = 0, U_j(s_i) < C_1) \\ &\quad + I(Y_j(s_i) = \alpha, C_1 \leq U_j(s_i) < C_2) \\ &\quad + I(Y_j(s_i) > \alpha, U_j(s_i) \geq C_2)]. \end{aligned} \quad (1)$$

Proposed Model and Inference

Distribution on Latent Variable \mathbf{U} :

$$U_j(s_i) = \mathbf{x}(s_i)' \boldsymbol{\beta} + w(s_i) + \varepsilon_j(s_i), \quad (2)$$

$X = (\mathbf{x}(s_1), \dots, \mathbf{x}(s_n))'$: design matrix of covariate information;

$\boldsymbol{\beta}$: coefficients of covariates;

$\mathbf{w} = (w(s_1), \dots, w(s_n))'$: spatial process;

$\boldsymbol{\varepsilon}(s_i) = (\varepsilon_1(s_i), \dots, \varepsilon_r(s_i))'$: error terms accounting for variation among the repeated measurements.

Proposed Model and Inference

Distribution on Latent Variable \mathbf{U} :

$$U_j(s_i) = \mathbf{x}(s_i)' \boldsymbol{\beta} + w(s_i) + \varepsilon_j(s_i), \quad (2)$$

$X = (\mathbf{x}(s_1), \dots, \mathbf{x}(s_n))'$: design matrix of covariate information;

$\boldsymbol{\beta}$: coefficients of covariates;

$\mathbf{w} = (w(s_1), \dots, w(s_n))'$: spatial process;

$\boldsymbol{\varepsilon}(s_i) = (\varepsilon_1(s_i), \dots, \varepsilon_r(s_i))'$: error terms accounting for variation among the repeated measurements.

Distribution on Latent Variable \mathbf{w} :

$$\mathbf{w} \sim N_n(\mathbf{0}, \sigma^2 K(\theta)), \quad (3)$$

where $K(\theta) = (K(\|s_i - s_j\|))_{n \times n}$.

$K(\theta, s_i, s_j) = e^{-d_{ij}/\theta}$, where d_{ij} is the distance between s_i and s_j .

Proposed Model and Inference

Priors:

$$\varepsilon(s_i) \sim N_r(\mathbf{0}, \tau^2 H(\rho)), \quad H(\rho)_{r \times r} = \rho J_r + (1 - \rho) I_r. \quad (4)$$

Proposed Model and Inference

Priors:

$$\varepsilon(s_i) \sim N_r(\mathbf{0}, \tau^2 H(\rho)), \quad H(\rho)_{r \times r} = \rho J_r + (1 - \rho) I_r. \quad (4)$$

$$\mathbf{U} | \mathbf{w}, \boldsymbol{\beta}, \tau^2, \rho \sim N_{rn}(\boldsymbol{\mu}, \mathbf{G}),$$

where $\boldsymbol{\mu} = (\mathbf{1}_r \otimes \mathbf{X})\boldsymbol{\beta} + \mathbf{1}_r \otimes \mathbf{w}$, $\mathbf{G} = \tau^2 H(\rho) \otimes I_n$.

Proposed Model and Inference

Priors:

$$\varepsilon(s_i) \sim N_r(\mathbf{0}, \tau^2 H(\rho)), \quad H(\rho)_{r \times r} = \rho J_r + (1 - \rho) I_r. \quad (4)$$

$$\mathbf{U} | \mathbf{w}, \beta, \tau^2, \rho \sim N_{rn}(\boldsymbol{\mu}, G),$$

where $\boldsymbol{\mu} = (\mathbf{1}_r \otimes \mathbf{X})\beta + \mathbf{1}_r \otimes \mathbf{w}$, $G = \tau^2 H(\rho) \otimes I_n$.

$$\beta \sim N_p(\mathbf{0}, dI_p); \quad (5)$$

$$\sigma^2 \sim IG(a_1, b_1); \quad (6)$$

$$\tau^2 \sim IG(a_2, b_2); \quad (7)$$

$$\theta \sim IG(a_3, b_3); \quad (8)$$

$$\rho \sim U(-1/(r-1), 1). \quad (9)$$

Proposed Model and Inference

Numerical Simulation Methods:

- Slice sampler and generalized ratio-of-uniform algorithm nested in Gibbs Sampler is used.

Prediction Simulation Algorithm:

- simulate $\mathbf{U}, \beta, \sigma^2, \tau^2, \theta, \rho$ from $[\mathbf{U}, \beta, \sigma^2, \tau^2, \theta, \rho | \mathbf{Y}]$;
- sample u_0 from $u_0 | \mathbf{U}, \beta, \sigma^2, \tau^2, \theta, \rho \sim N(\boldsymbol{\mu}_0 + (\mathbf{1}_r \otimes \mathbf{k}_\theta)' G_l^{-1}(\mathbf{U} - \boldsymbol{\mu}_l), \sigma^2(1 + \frac{\tau^2}{\sigma^2} - (\mathbf{1}_r \otimes \mathbf{k}_\theta)' G_l^{-1}(\mathbf{1}_r \otimes \mathbf{k}_\theta)))$.
- compute $y_0^* = g^{-1}(u_0)$.

Proposed Model and Inference

Hypotheses Testing About Covariate Effect

Prior probability: $Pr(H_i)$, $i = 0, 1$.

Posterior probability:

$$Pr(H_i|\mathbf{Y}) = \frac{P(\mathbf{Y}|H_i)Pr(H_i)}{P(\mathbf{Y}|H_0)Pr(H_0) + P(\mathbf{Y}|H_1)Pr(H_1)}.$$

Posterior odds ratio:

$$\frac{Pr(H_0|\mathbf{Y})}{Pr(H_1|\mathbf{Y})} = \frac{Pr(H_0)}{Pr(H_1)} \cdot \frac{Pr(\mathbf{Y}|H_0)}{Pr(\mathbf{Y}|H_1)}.$$

Bayes' factor:

$$BF_{01} = \frac{Pr(H_0|\mathbf{Y})}{Pr(H_1|\mathbf{Y})} / \left(\frac{Pr(H_0)}{Pr(H_1)} \right) = \frac{Pr(\mathbf{Y}|H_0)}{Pr(\mathbf{Y}|H_1)}.$$

Data Analysis

Objective of Analysis:

- Studying the spatial dependence among the coverage of some specific species of ground flora at different locations in MOFEP sites;
- predicting the vegetation coverage at unsampled locations;
- testing covariate effects.

Data:

- Locations: 292 subplots in site 1 ($n = 292$);
- Repeated Measurements: coverage of *desmodium nudiflorum* of the 4 quadrats within each subplot ($r = 4$);
- Covariates: soil depth and aspect class ($p = 3$ by adding the intercept).

Data Analysis

Transformation:

$$U_j(s_i) = \log \left(\frac{Y_j^*(s_i) + 0.0005}{1 - (Y_j^*(s_i) + 0.0005)} \right).$$

Relationship between **Y** and **U**:

$$Y_j(s_i) = \begin{cases} 0, & U_j(s_i) < C_1, \\ 0.001, & C_1 \leq U_j(s_i) < C_2, \\ \frac{e^{U_j(s_i)} * 0.9995 - 0.0005}{1 + e^{U_j(s_i)}}, & U_j(s_i) \geq C_2. \end{cases}$$

where $C_1 = -6.500789$ and $C_2 = -4.545825$.

Data Analysis

Table: Posterior quantities of β , σ^2 , τ^2 , θ and ρ

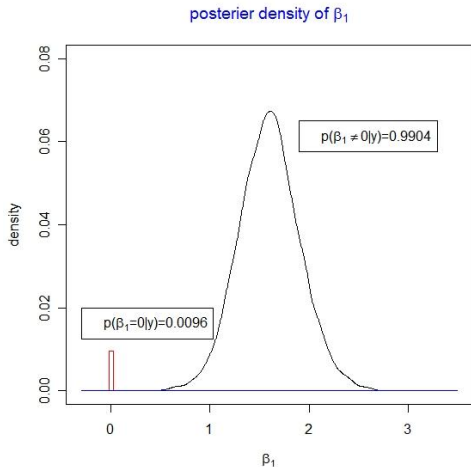
parameter	mean	median	st. d.	95% Bayesian CI
β_0	-6.6152	-6.5903	0.4055	[-7.4779, -5.7783]
β_1	1.6003	1.6019	0.3053	[0.9974, 2.2098]
β_2	1.2159	1.2244	0.3429	[0.5267, 1.8728]
σ^2	2.1654	2.1191	0.5406	[1.2682, 3.3746]
τ^2	5.4251	5.4317	0.5223	[4.4165, 6.4185]
θ	90.3325	81.3004	36.9200	[43.1562, 187.9171]
ρ	-0.0062	-0.0024	0.0669	[-0.1528, 0.1094]

Data Analysis

Hypothesis Testing:

- For $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$, $BF_{21} = 0.0097$.

Posterior probabilities for H_0 and H_1 :

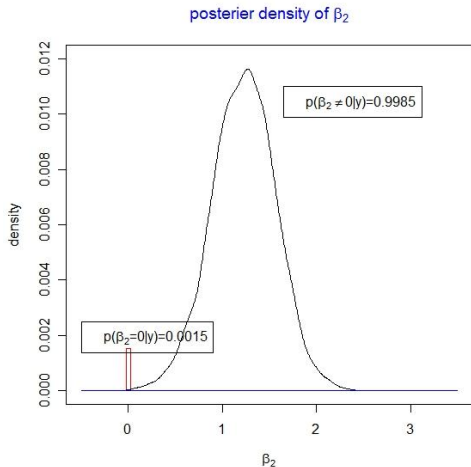


Data Analysis

Hypothesis Testing:

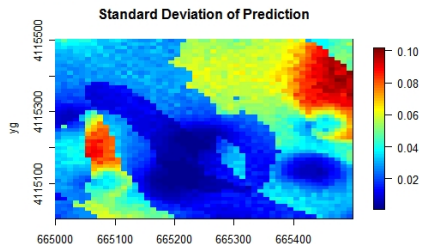
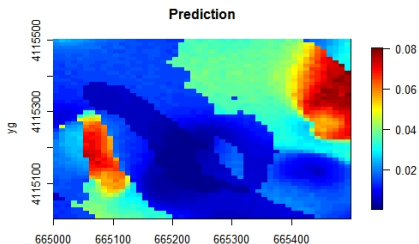
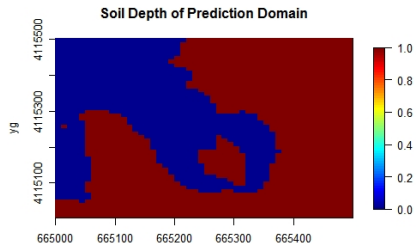
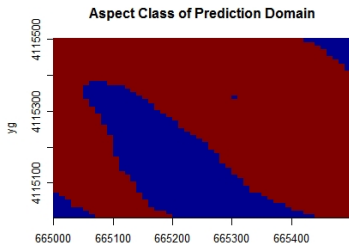
- For $H_0 : \beta_2 = 0$ vs. $H_1 : \beta_2 \neq 0$, $BF_{31} = 0.0015$.

Posterior probabilities for H_0 and H_1 :



Data Analysis

Predictions and Covariates:



Future Work and Summary

Proposed Multivariate Model:

- $\mathbf{z}_{k1}(s_1), \dots, \mathbf{z}_{kr}(s_n)$: r left-censored repeated measurements for k th variables collected from locations s_1, \dots, s_n ;
- $z_{kj}^*(s_i)$: j th unobserved true values for k th species at s_i ;
- $X = (\mathbf{x}(s_1), \dots, \mathbf{x}(s_n))'$: design matrix of covariate information.

Censored Data:

$$z_{kj}(s_i) = \begin{cases} 0, & -\infty \leq z_{kj}^*(s_i) < p_1, \\ \alpha, & p_1 \leq z_{kj}^*(s_i) < p_2, \\ z_{kj}^*(s_i), & p_2 \leq z_{kj}^*(s_i) < +\infty. \end{cases}$$

where $k = 1, 2, j = 1, \dots, r, i = 1, \dots, n$ and $p_1 < \alpha < p_2$.

Future Work and Summary

Transformation:

$$U_{kj}(s_i) = g(z_{kj}^*(s_i)).$$

Likelihood:

$$\begin{aligned} [\mathbf{z}|\mathbf{U}] &= \prod_{i=1}^n \prod_{j=1}^r \prod_{k=1}^2 [I(z_{kj}(s_i) = 0, U_{kj}(s_i) < C_1) \quad (10) \\ &+ I(z_{kj}(s_i) = \alpha, C_1 \leq U_j(s_i) < C_2) \\ &+ I(z_{kj}(s_i) > \alpha, U_j(s_i) \geq C_2)]. \end{aligned}$$

Future Work and Summary

Regression Model:

$$U_{kj}(s_i) = \mathbf{x}(s_i)' \boldsymbol{\beta} + \eta_k + w_k(s_i) + \varepsilon_{kj}(s_i), \quad (11)$$

- $\boldsymbol{\beta}$: regression coefficients for location-specific covariates;
- $\mathbf{x}(s_i)$: $p \times 1$ vector of covariates;
- $w_k(s_i)$: spatial effect on the k th variable of interest at location s_i ;
- η_k : variable-specific effect;
- $\varepsilon_{kj}(s_i)$: measurement error.

Future Work and Summary

$$\mathbf{U} = \mathbf{1}_{kr} \otimes \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\eta} \otimes \mathbf{1}_{nr} + \Omega(\mathbf{1}_r \otimes \mathbf{w}) + \boldsymbol{\varepsilon}, \quad (12)$$

When $r = 4$ and $k = 2$, one possible choice is that

$$\Omega_{8n \times 8n} = \begin{pmatrix} I_n & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & I_n & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & I_n & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & I_n & 0 \\ 0 & I_n & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & I_n & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & I_n & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & I_n \end{pmatrix}.$$

Future Work and Summary

Spatial Process: All the priors used in the univariate model could be used, except for the spatial process. Let

$$\mathbf{w}_1 = (w_1(s_1), \dots, w_1(s_n))', \quad \mathbf{w}_2 = (w_2(s_1), \dots, w_2(s_n))', \\ \mathbf{w} = (\mathbf{w}_1', \mathbf{w}_2')',$$

$$\mathbf{w} \sim N(\mathbf{0}, B). \quad (13)$$

Specification of B :

- Separable covariance matrix: $B_1 = \Sigma_v \otimes \sigma^2 K(\theta)$;
- Similar to the covariance proposed by Woodard (1999), let

$$B_2 = \begin{pmatrix} \sigma_1^2 K(\theta_1) & \phi \sigma_1 \sigma_2 K(\theta_1)^{\frac{1}{2}} K(\theta_2)^{\frac{1}{2}} \\ \phi \sigma_1 \sigma_2 K(\theta_2)^{\frac{1}{2}} K(\theta_1)^{\frac{1}{2}} & \sigma_2^2 K(\theta_2) \end{pmatrix};$$

Future Work and Summary

Comparison of Choices of B :

- The separable covariance matrix is easy to implement in computation, but the assumption is very strong;
- The second specification of B is guaranteed to be valid though relaxed the assumption in the first one, but hard to interpret, the prediction is also more complicated;
- How to construct the structure of B to achieve faster computation, better model fit, more reasonable physical meaning is still an open question.
- Application study is needed to compare the results. Different specifications might be preferred in different case studies.

Future Work and Summary

What we have seen:

- Zero-inflated model for continuous spatially correlated data;
- Inference;
- Possible generalization of the model.

What else?

- Questions!

Thank you!