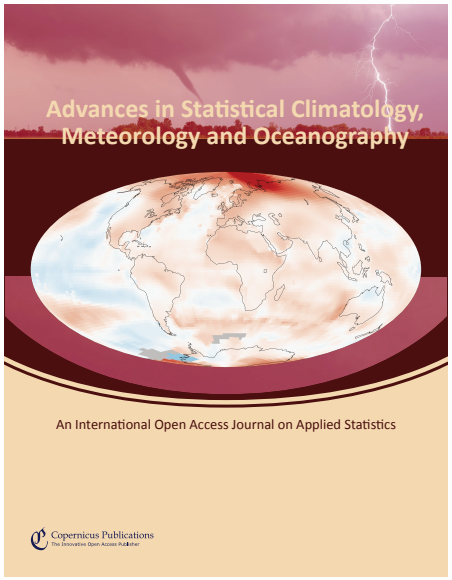
Evaluating Wetland Health: Multilevel Latent Gaussian Process Model for Mixed Discrete and Continuous Multivariate Response Data

Erin M. Schliep
Duke University

Jennifer A. Hoeting
Colorado State University

June 2014



ASCMO

Open Access

NEW JOURNAL

Advances in Statistical Climatology, Meteorology, and Oceanography (ASCMO)

- ASCMO is an interdisciplinary journal that publishes cutting-edge scientific advances and statistical methods.
- ASCMO serves at the interface between statistics and the atmospheric and oceanic sciences.
- Journal Website: ascmo.net

A few tips on fitting spatial statistics models

Sample size

You need more observations for spatially correlated data than for iid data.

- $n = 30$ is the rule of thumb for independent data
- Sample size needs to be much larger when the data are dependent ($n > 220$?)

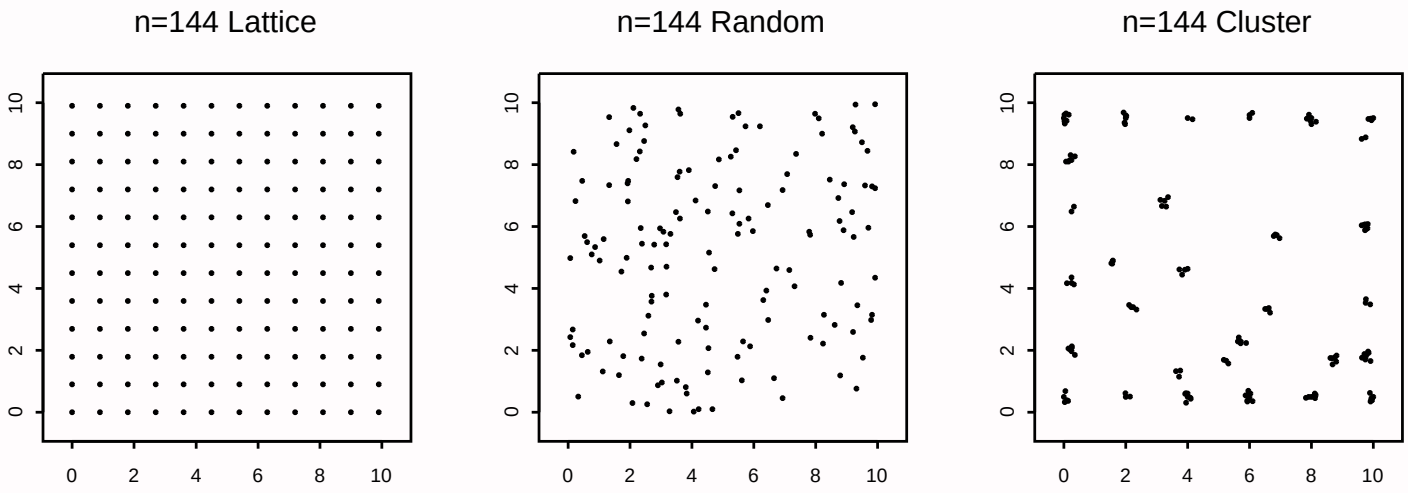
More on this topic: Irvine et al. (2007) “Spatial Designs and Properties of Spatial Correlation: Effects on Covariance Estimation”, *JABES*

Sampling Design

- For spatial data, want a design with some sites with small distances and some with large distances.
- Avoid simple grids (usually)

More on this topic:

- Effect of design on estimation (Irvine et al. (2007) *JABES* and references therein)
- GRTS design (US NPS website)



Model selection for spatial models

Consider the spatial regression model

$$Z(\mathbf{s}) = \beta_0 + X_1(\mathbf{s})\beta_1 + \cdots + X_p(\mathbf{s})\beta_p + \epsilon(\mathbf{s})$$

where $\epsilon \sim N(\mathbf{0}, \mathbf{\Sigma}(\rho))$.

- Getting the covariance correct is an important part of model selection.
- If you are selecting covariates in a spatial regression model, do model selection for the spatial model.
- Don't select the covariates using an independent response model and then estimate the spatial covariates.

More on this topic: Hoeting et al. (2006) "Model selection for geostatistical models", *Ecological Applications*

Modeling a discrete response

- When the response is discrete it is even more challenging to estimate the parameters of the spatial covariance
- There is more info in the data from a continuous response about the spatial covariance parameters as compared to discrete data

A sampling of the literature: De Oliveira (2000, 2013), Diggle et al. (1997), many more!

Evaluating Wetland Health: Multilevel Latent Gaussian Process Model for Mixed Discrete and Continuous Multivariate Response Data

Erin M. Schliep
Duke University

Jennifer A. Hoeting
Colorado State University

June 2014

Thank you to

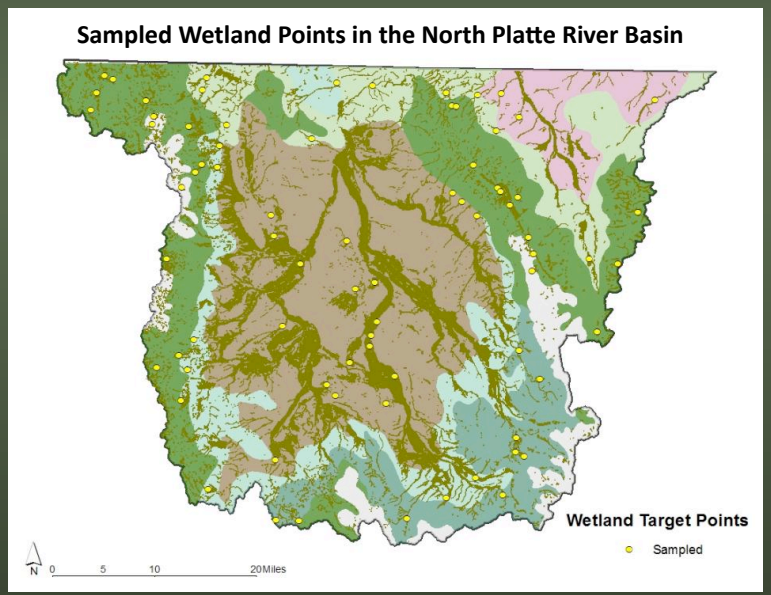
- U.S. Environmental Protection Agency for support of this work (EPA-1605-09).
- Joanna Lemly and Laurie Gilligan of Colorado Natural Heritage Program.
- Peter Guttorp and Alexandra Schmidt for organizing this conference

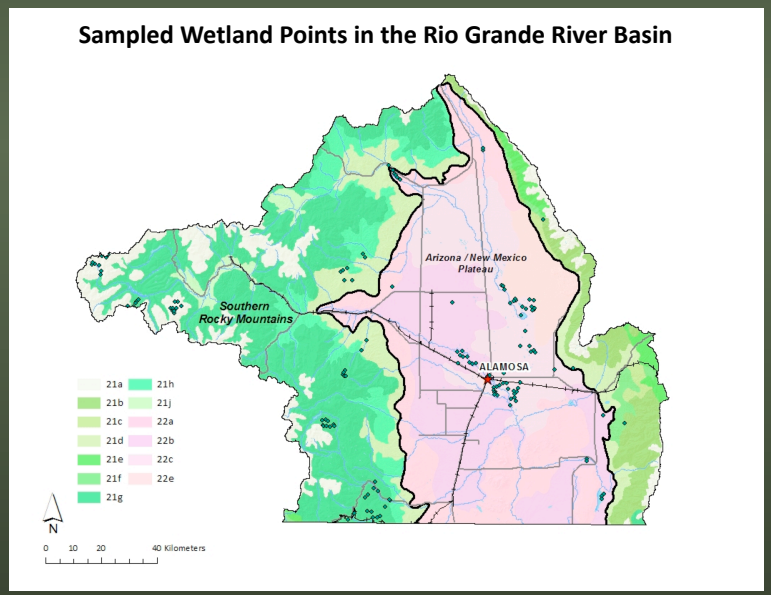
North Platte and Rio Grande Basinwide Wetland Profile and Condition Assessment

Wetlands

Generalized random-tessellation stratified sample design (GRTS)

- 100 target points across North Platte Basin, 150 target points across Rio Grande Basin
- Stratification differed between the two river basins
- Represent all wetland areas in North Platte and Rio Grande Basins

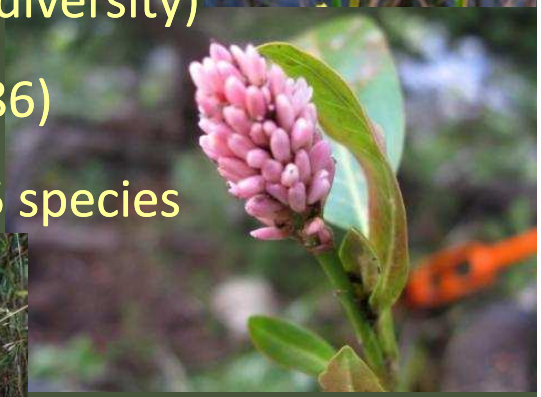




16 private points not shown

Results: Plant Diversity in North Platte

- 612 individual taxa encountered (high alpha diversity)
- 537 identified to species
- >200 species encountered only once (high beta diversity)
- 49 species per site on average (min = 11, max =86)
- Most diverse genera: *Carex:* 44 species, *Salix:* 15 species



Biotic Condition Index

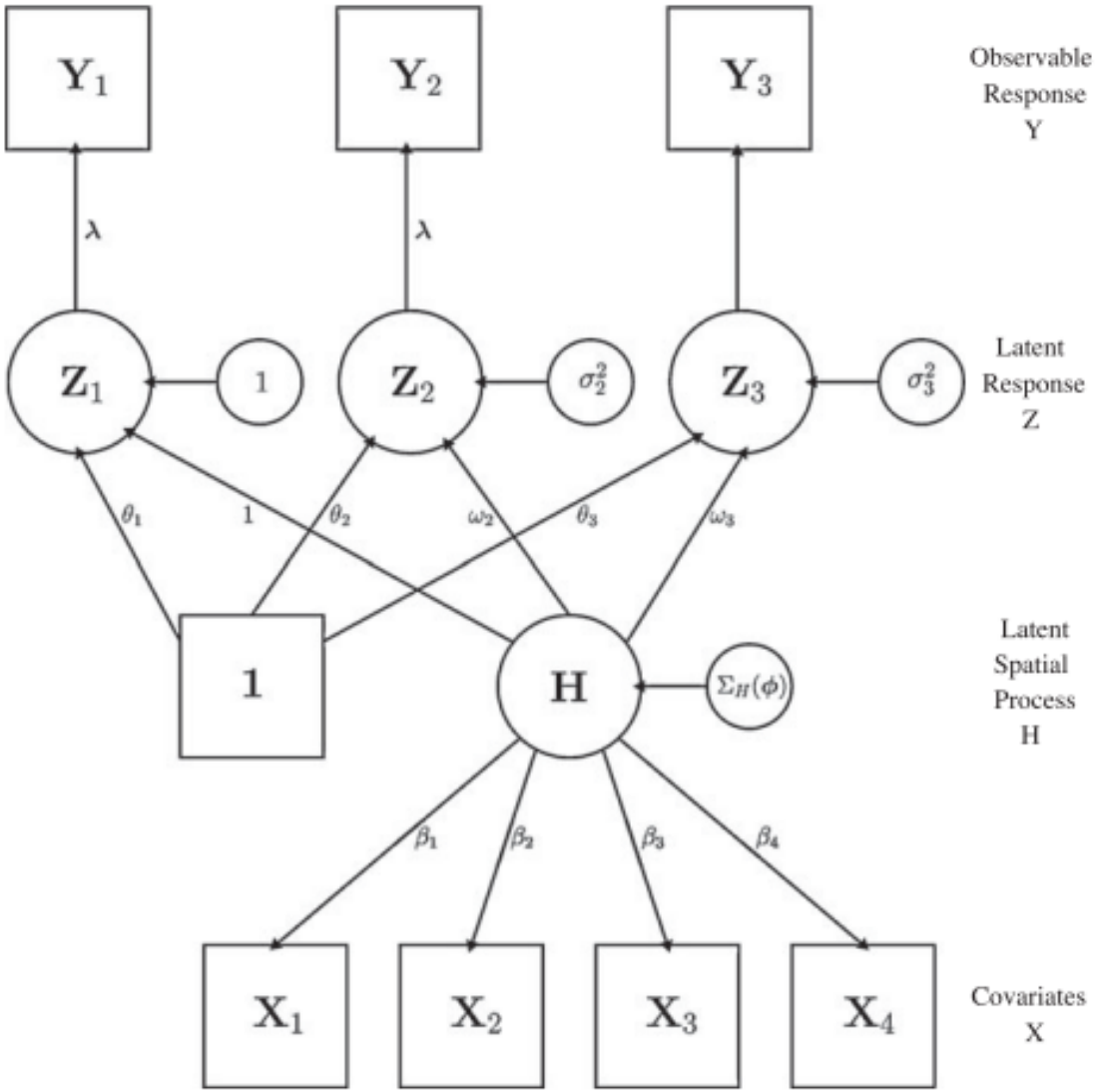
Calculations for a representative site

Metric	A	B	C	D/E		Weight	Score (weight x rating)
Native Plant Cover	5	4	**3**	2	1	0.20	**0.60**
Noxious Weed Cover	5	**4**	3	1		0.20 (lowest value used)	**0.80**
Aggressive Native Cover	**5**	4	3	1			
Structural Complexity	5	**4**	3	1		0.20	**0.80**
Floristic Quality Assessment (Mean C)	5	4	**3**	2	1	0.40	**1.20**
Biotic Condition Rating A = >4.5 - 5.0 **B = >3.5 – 4.5** **C = >2.5 – 3.5** D = 1.0 – 2.5							**Total** (sum) **= 3.40**

Floristic Quality Assessment (Mean C) is also reported as a continuous response.

Goals of our model

- Improve upon best professional judgment-based use of metric data
- Predict unobserved biotic condition (health) at unobserved sites
- Produce maps of predictions and prediction uncertainty
- Rank the sites so decision makers can easily use the results.


Y_1
Y_2
Y_3
Observable Response Y
λ
λ
Z_1
1
Z_2
σ_2^2
Z_3
σ_3^2
Latent Response Z
θ_1
1
θ_2
ω_2
θ_3
ω_3
1
H
$\Sigma_H(\phi)$
Latent Spatial Process H
β_1
β_2
β_3
β_4
X_1
X_2
X_3
X_4
Covariates X

Multivariate multilevel latent variable model

$i = 1, \ldots, n$ sites
$j = 1, \ldots, J$ metrics, $J = J_o + J_c$

Y_{ij} = Observed response for site i and metric j.

Z_{ij} = Latent continuous variable for site i and metric j.

H_i = Latent continuous variable for site i.
Use to assess biotic condition at site i.

X_i = Site specific covariates.

Link function

Let F_j = function that relates observed Y_{ij} to latent Z_{ij}.

For continuous Y_{ij}, identity link: $F_j(Y_{ij}) = Z_{ij}$.

Ordinal multivariate response

For a metric with K categories,

$$Y_{ij} = \begin{cases} 1 & \lambda_{j,0} < Z_{ij} < \lambda_{j,1} \\ 2 & \lambda_{j,1} < Z_{ij} < \lambda_{j,2} \\ \vdots & \vdots \\ K-1 & \lambda_{j,K-2} < Z_{ij} < \lambda_{j,K-1} \\ K & \lambda_{j,K-1} < Z_{ij} < \lambda_{j,K} \end{cases}$$

where

$$-\infty = \lambda_{j,0} \leq \lambda_{j,1} \leq \ldots \lambda_{j,K} = \infty$$

A related univariate ordinal model: Higgs and Hoeting (2011)

Ordinal link function

The conditional distribution of $\mathbf{Y}$ given $\mathbf{Z}$ and the parameter $\boldsymbol{\lambda}$ can be written as

$$F_j(Y_{ij}|\lambda_j) = \sum_{k=1}^{K} k I_{\{\lambda_{j,k-1} < Z_{ij} \leq \lambda_{j,k}\}}, \quad \text{for } i = 1, \ldots, n, \;\; j = 1, \ldots, J_o$$

The values of $\boldsymbol{\lambda}$ are dependent on j, meaning they are specific to the different variables of the multivariate response.

Latent **Z**

For each metric j in $1, ..., J$,

$$\mathbf{Z}_j \sim GP(\theta_j \mathbf{1} + \omega_j \mathbf{H}, \sigma_j^2 \mathbf{I})$$

- For $j \neq l$, $\mathbf{Z}_j$ and $\mathbf{Z}_l$ are assumed independent *apriori*.
- The correlation is established through the random variables θ, ω, and $\mathbf{H}$.

Latent **H**

$$\mathbf{H} \sim GP(\mathbf{X}\beta, \Sigma_{\mathbf{H}}(\phi))$$

where $\mathbf{X}$ has dimension $l \times p$ and β is a vector of coefficients of dimension $p \times 1$.

- The latent variable **H** makes this model a multilevel latent model.
- **H** is assumed to be driving the observed multivariate response, **Y**.
- Inference of the spatially referenced random variable **H** is of particular interest as it will be used in comparing locations within the region of interest.

Model Summary

For $i = 1, \ldots, n$

$$F_j(Y_{ij}) = \begin{cases} Z_{ij} & \text{for } J_c \text{ continuous metrics} \\ \sum_{k=1}^{K} k I_{\{\lambda_{k-1} < Z_{ij} \leq \lambda_K\}} & \text{for } J_o \text{ ordinal metrics} \end{cases}$$

where

$$-\infty = \lambda_{j,0} \leq \lambda_{j,1} \leq \ldots \lambda_{j,K} = \infty$$

For each j in 1, ..., J,

$$\mathbf{Z}_j \sim GP(\theta_j 1 + \omega_j \mathbf{H}, \sigma_j^2 \mathbf{I})$$

where

$$\mathbf{H} \sim GP(\mathbf{X}\beta, \Sigma_{\mathbf{H}}(\phi))$$

Priors

1. Transform cut-points, $\lambda_0, \ldots, \lambda_K$, and use a MVN prior
2. Conjugate priors for the metric-specific parameters
   - $\sigma_j^2 \sim$ Inverse-Gamma(a,b)
   - $\boldsymbol{\theta} \sim MVN(\mathbf{0}, \sigma_\theta^2 \mathbf{I})$
   - $\boldsymbol{\omega} \sim MVN(\mathbf{0}, \sigma_\omega^2 \mathbf{I})$
3. Conjugate prior for $p \times 1$ vector of coefficients
   - $p(\beta) \sim MVN(\mathbf{0}, \sigma_\beta^2 \mathbf{I})$
4. Some care needs to be taken to maintain identifiability of the parameters.

Spatial correlation: Gaussian Field

We assume

$$\Sigma_{\mathbf{H}}(\phi) = \rho(\mathbf{s}_i - \mathbf{s}_k; \phi) = \phi_1 \exp^{-d_{ik}\phi_2}$$

where d_{ik} = Euclidean distance between locations i and k.

- $\phi_1 \sim$ Inv. Gamma(a_{ϕ_1}, b_{ϕ_1})
- $\phi_2 \sim \Gamma(a_{\phi_2}, b_{\phi_2})$

Bayesian estimation for GF model

Estimation using MCMC (Gibbs + extra finesse)

Useful MCMC tricks to improve convergence:

- Data augmentation
- Parameter expansion

More details in Schliep and Hoeting (2014) “Data augmentation and parameter expansion for spatially correlated ordinal data”, available on arXiv

Assessing Wetlands in Colorado

95 locations within the North Platte region and 137 locations within the Rio Grande region

5 metrics:

1. Native Plant Cover
2. Noxious Weed Cover
3. Aggressive Native Cover
4. Structural Complexity
5. Floristic Quality Assessment (Mean C)

For each metric an integer is recorded from 1 to 5,
1 = worst and 5=best

Posterior estimates & 95% credible interval for discrete-only model

Parameter	Estimate	(95 % Cred. Int.)
β_1 Elevation	0.54	(0.22, 0.89
β_2 % Closed tree canopy	0.40	(0.20, 0.62)
β_3 Saline	0.62	(0.16, 1.10
β_4 Marsh	0.60	(0.26, 0.97)
β_5 Wet Meadow	-0.03	(-0.28 0.21)
β_6 Fen	1.00	(0.55, 1.53)

Ecological system-specific as compared to riparian

Effective Range

The effective range is the distance at which the covariance function does not exceed 0.05 times the variance.

Estimated Effective Range: 88 km (45, 184)

- Max distance within the North Platte River Basin =93 km
- Max distance within the Rio Grande River Basin = 202 km
- Min distance between the two river basins = 240 km

Posterior estimates & 95% credible int. for discrete-only model

Parameter	Estimate	(95 % Cred. Int.)
ω_1 Native Plant Cover	1.00	
ω_2 Noxious Weed Cover	1.37	(1.00, 1.90
ω_3 Aggressive Native Cover	2.54	(0.89, 6.01)
ω_4 Structural Complexity	0.21	(0.11, 0.33)
ω_5 Floristic Quality Assess.	1.59	(1.33, 1.91)
σ_1^2	1.00	
σ_2^2	1.34	(0.86, 2.16)
σ_3^2	20.46	(8.00, 67.64)
σ_4^2	0.89	(0.67,1.18)
σ_5^2	0.36	(0.22, 0.57

Comparing the metrics

Estimate the relationship between the multivariate response **Z** and the latent process **H** by computing multiple correlation.

Partition the covariance matrix of the matrix **Z** and vector **H** as

$$\mathbf{S} = \begin{bmatrix} \mathbf{S}_{ZZ} & \mathbf{S}_{ZH} \\ \mathbf{S}_{HZ} & \mathbf{S}_{HH} \end{bmatrix}$$

- $\mathbf{S}_{ZZ}$ is the $J \times J$ sample covariance matrix of **Z**
- $\mathbf{S}_{ZH}$ is the $J \times 1$ vector of sample covariances between **Z** and **H**
- $\mathbf{S}_{HH}$ is the sample variance of **H**.

Comparing the metrics

Evaluate the multiple correlation for each metric using the posterior simulations.
The correlation of $\mathbf{Z}_j$ and $\mathbf{H}$ is the square root of

$$R^2_{Z_j|H} = \frac{\mathbf{S}^{-1}_{HH}(\mathbf{S}_{ZH})_j}{\text{diag}(\mathbf{S}_{ZZ})_j}$$

where $(\mathbf{S}_{ZH})_j$ is the j^{th} element of the $J \times 1$ vector $\mathbf{S}_{ZH}$.

Large $R_{Z_j|H}$ values indicate that metric j is useful for estimating the unobserved latent spatial process.

Comparing the metrics

Comparing the metrics to one another.

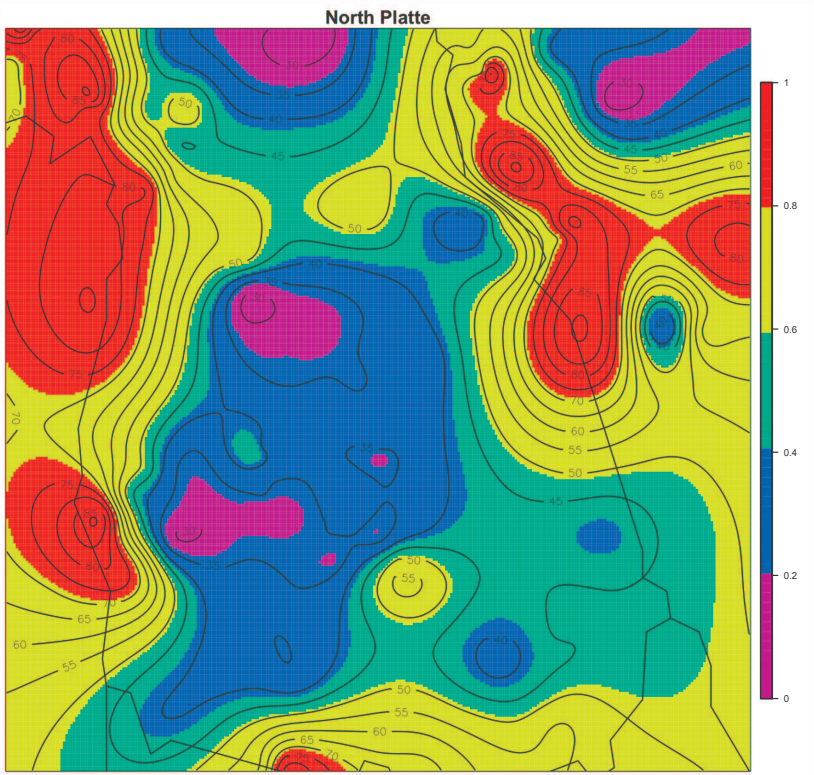
Metric	Corr.	Estimate	95 % CI
Native Plant Cover	$R_{Z_1\|H}$	0.80	(0.68, 0.88)
Noxious Weed Cover	$R_{Z_2\|H}$	0.84	(0.70, 0.92)
Aggressive Native Cover	$R_{Z_3\|H}$	0.58	(0.23, 0.83)
Structural Complexity	$R_{Z_4\|H}$	0.28	(0.09, 0.49)
Floristic Quality Assess.	$R_{Z_5\|H}$	0.96	(0.92, 0.98)

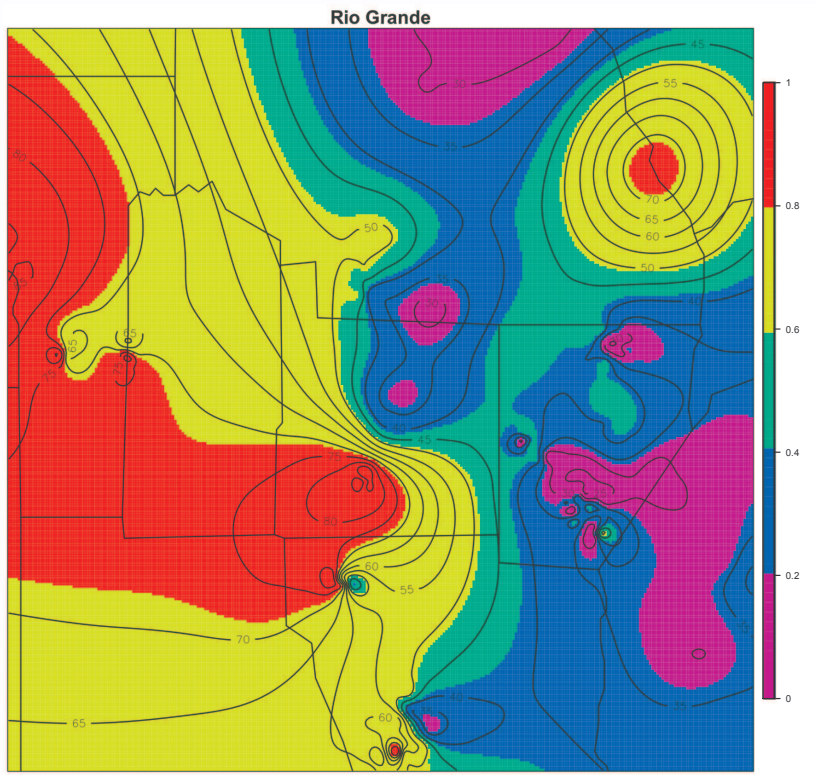
Comparing the metrics

Comparing the metrics to one another and comparing our results to the original index.

Metric	% Contrib.	% Contrib. CI	Index
Native Plant Cover	0.23	(0.20, 0.26)	20%
Noxious Weed Cover	0.24	(0.21, 0.28)	0 or 20%
Aggressive Native Cover	0.17	(0.08, 0.22)	0 or 20%
Structural Complexity	0.08	(0.03, 0.13)	20%
Floristic Quality Assess.	0.28	(0.25, 0.32)	40%

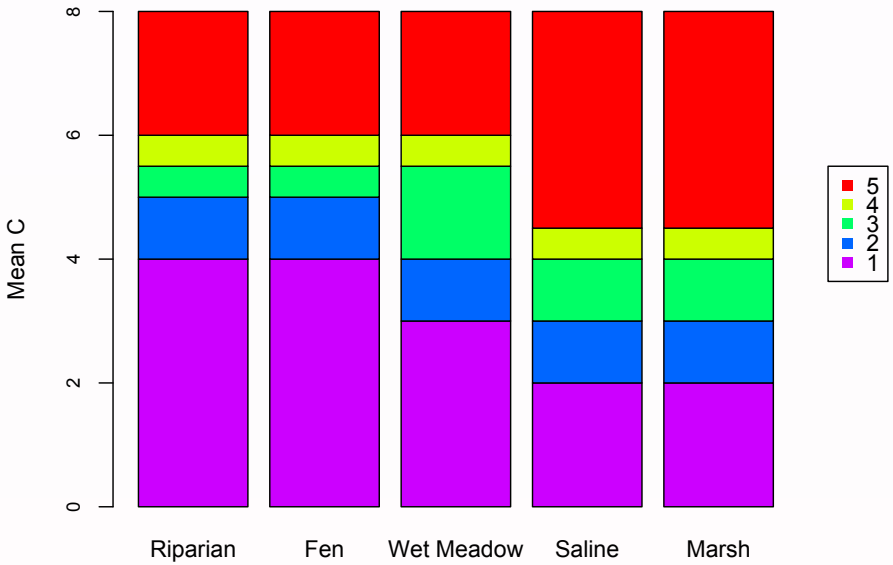
Posterior quantiles of the latent wetland condition variable



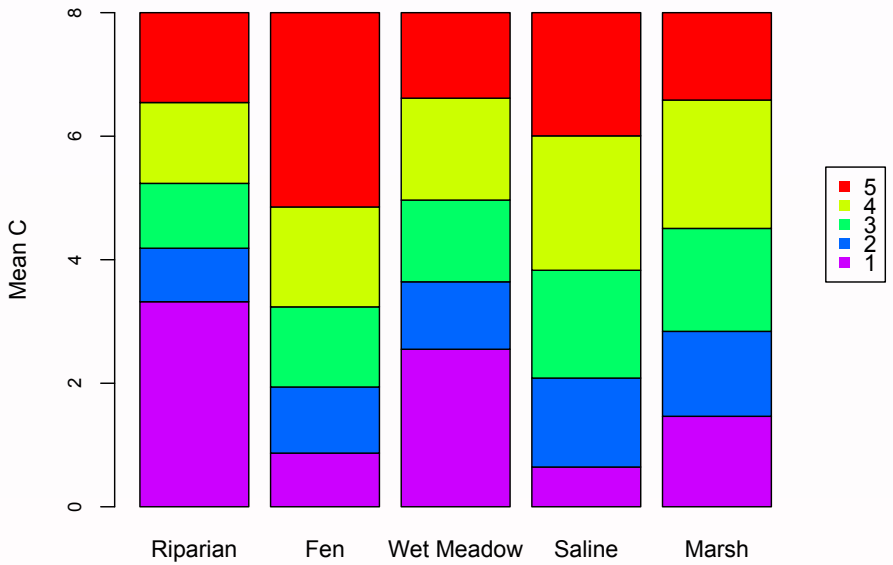


Threshold values for floristic quality assessment (mean C)

Threshold values proposed by Lemly et al. 2009



Threshold values from data driven method



Conclusions

- Multivariate multilevel latent variable model
- Discrete or continuous response
- A comprehensive model to assess wetlands or other resources
- Paper: Schliep and Hoeting (2014), *Journal of Agricultural, Biological, and Environmental Statistics*, Volume 18, Number 4, 492–513.


Questions?