# Bivariate Spatial Regression Model for Mixed-Response Lattice Data

MS Creative Component with Dr. Emily Berg

Steve Harms

Iowa State University
Department of Statistics

*stharms@iastate.edu*

April 4, 2019

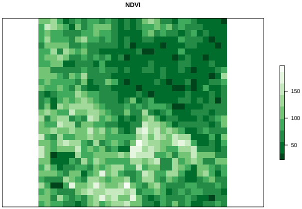
# Overview

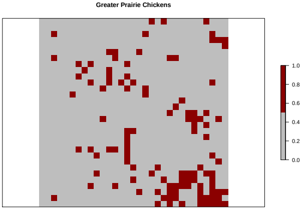
# Data Description

- 2+ response variables, possibly correlated, distributions of mixed types
- Spatial data collected on a grid (discrete-indexed)
  - Model dependence in terms of adjacency rather than distance
- Study interested in both covariates and dependence relationships
  - Need interpretable dependence parameters

Examples:

- ecology (presence/absence + environmental variables)
- agriculture (yield + categorical plant characteristics)
- weather/meteorology
- image pixel analysis



# Notation

- $\{s_i : i = 1, ..., n\}$ : locations
- $Y(s_i)$ : Binary variable (0-1) at location $s_i$
- $Z(s_i)$ : Gaussian variable at location $s_i$
- $\boldsymbol{z} : \{z(s_i) : i = 1, ..., n\}$ and $\boldsymbol{y} : \{y(s_i) : i = 1, ..., n\}$: all obs. of the same type
- $\boldsymbol{z}(\bar{s}_i) : \{z(s_j) : i \neq j\}$ and $\boldsymbol{y}(\bar{s}_i) : \{y(s_j) : i \neq j\}$: all obs. of the same type except for location $s_i$
- $N_i$ : Neighborhood of location $s_i$
  - $j \in N_i$ if $s_i$ and $s_j$ are neighboring locations
- $M_i$ : Number of neighbors ($|N_i|$) of location $s_i$
- $\delta_i$ , $\mu_i$ : Global/marginal mean parameters (node potentials)
- $\eta_y$ , $\eta_z$ : Spatial dependence parameters for variables of the same type
- $\rho$ : Dependence parameter for variables of different types

# Background - Markov Random Field Models

Markov Random fields (auto-models) for lattice data from Besag (1974):

- One response variable from a one-parameter exponential family
- Specify expectation at each location conditionally on neighboring values: $\mathrm{E}[\mathrm{x}(\mathrm{s}_i)|x(\bar{s}_i)] = \mu_i + \sum_{j \neq i} \eta_{ij} x(s_j)$

- Sain, Furrer, and Cressie (2011): Extension to multivariate auto-normal model by "stacking" the univariate lattices
  - Direct extension of Besag's univariate conditional model rather than specifying MVN covariance of response variables
- Caragea and Berg (2014): Model two tree species' presence in bivariate autologistic model
  - Both response variables are binary

# Background - Mixed Graphical Models

Graphical models for learning the edge structure given *n* realizations of a set of mixed nodes (variables):

- Chen, et al. (2014): Edge potential parameter spaces for mixed one-parameter exponential family nodes corresponding to valid joint
- Lee and Hastie (2015): Specific case where any continuous variable is conditionally Gaussian, any discrete variable is conditionally logistic regression
- Gaussian nodes:

$$p(z(s_i)|\boldsymbol{z}(\bar{s}_i),\boldsymbol{y})=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{\frac{-\sigma^{-2}}{2}\left[\frac{\mu_i+\sum_{i=1}^{n}\rho_{ij}y(s_j)+\sum_{i=1}^{n}\eta_{z,ij}z(s_j)}{\sigma^{-2}}-z(s_i)\right]^2\right\}$$

- Discrete nodes:

$$p(y(s_i)|\boldsymbol{y}(\bar{s}_i),\boldsymbol{z})=\frac{\exp\left\{y(s_i)\left(\delta_i+\sum_{i=1}^{n}\rho_{ij}z(s_j)+\sum_{i=1}^{n}\eta_{y,ij}y(s_j)\right)\right\}}{1+\exp\{y(s_i)\left(\delta_i+\sum_{i=1}^{n}\rho_{ij}z(s_j)+\sum_{i=1}^{n}\eta_{y,ij}y(s_j)\right)\}}$$

# Bivariate Mixed MRF Model

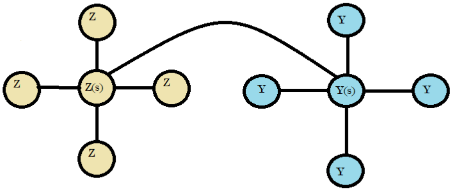
## Mixed Bivariate Markov Random Fields

Goal: Combine the interpretive properties of auto-models with compatible node-conditional specifications from mixed graphical models

- View each response variable at each location as a node, combine with (fixed/known) set of edges into an undirected graph with $2n$ total nodes
- Use this graph structure in the context of auto-models

# Assumptions

- Simplifying assumptions for the models in this work:
  - Four-nearest neighbor structure (pairwise-only dependence): $M = 4$
  - Binary and Gaussian nodes are only connected at each location
  - Isotropy: $\eta_{y,ij} = \eta_y$ and $\eta_{z,ij} = \eta_z$
  - Interior locations only: each node has same number of neighbors (no edge effects)
  - Common conditional variance for Gaussian nodes (variance $= \sigma^2 \forall z(s_i)$)



# Conditional Distributions

- Gaussian node at location $s_i$ has conditional distribution

$$p(z(s_i)|\boldsymbol{z}(\bar{s}_i), y(s_i)) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ \frac{-1}{2\sigma^2} \left[z(s_i) - \alpha(s_i)\right]^2 \right\}$$

where

$$\alpha(s_i) = \mu_i + \frac{\eta_z}{M} \sum_{j \in N_i} (z(s_j) - \mu_j) + \rho \left( y(s_i) - \frac{\exp(\delta_i)}{1 + \exp(\delta_i)} \right)$$

# Conditional Distributions

- Binary node at location $s_i$ has conditional distribution

$$p(y(s_i)|\boldsymbol{y}(\bar{s}_i), z(s_i)) = \frac{\exp\{y(s_i)\,(\phi(s_i))\}}{1 + \exp\{y(s_i)\,(\phi(s_i))\}}$$

  where

$$\phi(s_i) = \delta_i + \frac{\eta_y}{M}\sum_{j \in N_i}\left[y(s_j) - \frac{\exp(\delta_j)}{1 + \exp(\delta_j)}\right] + \frac{\rho}{\sigma^2}(z(s_i) - \mu_i)$$

- "Centering" in the auto-regressive terms allow separation of global and local effects for a more sensible interpretation of $\eta_y$, $\eta_z$, and $\rho$

# Properties of the Conditional Specification

- $\mu_i$ and $\frac{\exp(\delta_j)}{1+\exp(\delta_i)}$ are independence expectations, i.e., marginal means
- Incorporate covariates $\boldsymbol{X_y}, \boldsymbol{X_z}$ by setting $\mu_i = \boldsymbol{X_{z,i}\beta_z}$ and $\delta_i = \boldsymbol{X_{y,i}\beta_y}$
- The remaining terms in $\alpha(s_i)$ and $\phi(s_i)$ govern the clustering or "local" behavior due to dependence/auto-correlation
- Full set of node-conditionals combine into a negpotential function

$$\boldsymbol{Q(y,z)} = \sum_{i=1}^{n}\left[\frac{\mu_i}{\sigma^2}z(s_i) - \frac{1}{2\sigma^2}z(s_i)^2 + \delta_i y(s_i)\right] + \frac{\rho}{\sigma^2}\sum_{i=1}^{n}\left[(z(s_i)-\mu_i)\left(y(s_i) - \frac{\exp(\delta_i)}{1+\exp(\delta_i)}\right)\right]$$
$$- \frac{1}{2\sigma^2}\sum_{i=1}^{n}\sum_{j\in N_i}\left[\frac{\eta_z}{M}(z(s_i)-\mu_i)(z(s_j)-\mu_j)\right]$$
$$+ \frac{1}{2}\sum_{i=1}^{n}\sum_{j\in N_i}\left[\frac{\eta_y}{M}\left(y(s_i) - \frac{\exp(\delta_i)}{1+\exp(\delta_i)}\right)\left(y(s_j) - \frac{\exp(\delta_j)}{1+\exp(\delta_j)}\right)\right]$$

- $p(\boldsymbol{y}, \boldsymbol{z}) = \frac{\exp(\boldsymbol{Q(y,z)})}{\int_\Omega \exp(\boldsymbol{Q(y,z)})d\boldsymbol{y}d\boldsymbol{z}}$ is a valid joint distribution if $\int_\Omega \exp(\boldsymbol{Q(y,z)})d\boldsymbol{y}d\boldsymbol{z} < \infty$, which holds if $\eta_z < 1$ and $\sigma^2 > 0$

# Model Breakdown

- When dependence becomes too strong, no longer have clearly separated marginal (global) and conditional (local) effects
  - Large $\eta_z$ or $\eta_y$ cause local dependence to overwhelm global properties
  - Data are heavily clustered/chaotic
- Kaiser (2007) quantifies boundaries for "nicely behaved" models:
  - Auto-logistic models: $\eta_y < 4$
  - Auto-normal models: Breakdown as $\eta_z \to 1$, especially $\eta_z \geq 0.99$
- In bivariate case, we have 3 different dependence parameters ($\eta_y$,$\eta_z$, $\rho$) and want all parameters to reflect appropriate properties of the data
- My rule of thumb: $\rho > \sqrt{\sigma^2(1-\eta_z)(4-\eta_y)}$ implies <u>strong dependence</u> for bivariate model

# Estimation and Inference

# Maximum Pseudo-likelihood Estimation

- For most practical applications, maximizing full likelihood requires repeated evaluation of high-dimensional integral in full joint distribution
- Instead, minimize the negative of the log pseudo-likelihood $\tilde{\ell}$, which only uses lower-dimensional node-conditionals:

$$\tilde{\ell} = \sum_{i=1}^{n} \log\left(p(z(s_i)|\boldsymbol{z}(\bar{s}_i), y(s_i))\right) + \sum_{i=1}^{n} \log\left(p(y(s_i)|\boldsymbol{y}(\bar{s}_i), z(s_i))\right)$$

- Hughes, et al. (2011) show loss in efficiency for MPLE is negligible when the lattice is reasonably large ($\sim$ 400 points)
- Exact or asymptotic distributions for parameter estimates are not straightforward from pseudo-likelihood methods

# Simulation and Inference

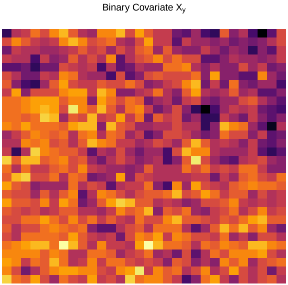
- Use Gibbs sampling to simulate from full set of specified node-conditional distributions
  - Simulate sequentially node-by-node from each conditional distribution, update, iterate until convergence
- Relatively simple simulation algorithm enables ready implementation of parametric bootstrap for inference:
  1. For a set of parameters $\hat{\boldsymbol{\Theta}}$ estimated from the original data, use Gibbs sampling to simulate bootstrap replicates from $\hat{\boldsymbol{\Theta}}$
  2. Obtain MPLE estimates for each bootstrapped dataset to get an empirical distribution of parameters, $\theta^* \in \boldsymbol{\Theta}^*$
  3. Lower and upper bounds for a 100(1-$\alpha$)% confidence interval for a parameter $\theta \in \boldsymbol{\Theta}$ are:
     $L_\alpha = 2\hat{\theta} - \theta^*_{1-\alpha/2}$ and $U_\alpha = 2\hat{\theta} - \theta^*_{\alpha/2}$

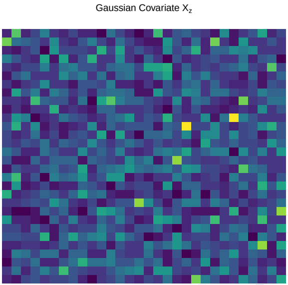
# Simulation Studies

# Study 1: MPLE Estimation Set-up

- Goal: Compare performance for full model vs. previous reduced models
- 3 different dependence regimes (weak, moderate, strong)
- For each dependence regime, obtain MPLE estimates (8 parameters for full model) for 1,000 simulated datasets on a 30 x 30 lattice

Figure: Covariates used in simulation: Normal with spatial correlation $\eta_z = 0.9$, $\mu = 1$, $\sigma^2 = 1$ (left) and $Gamma(3, 4)$ (right)



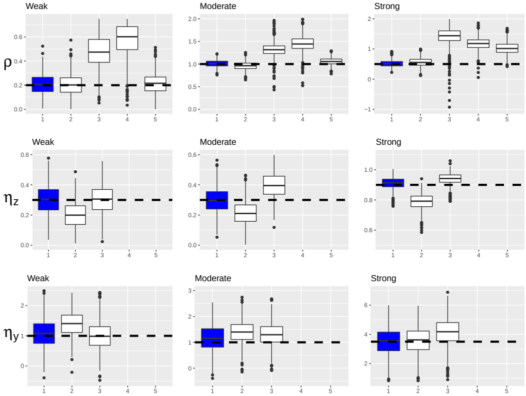



| Parameter Regimes Used to Simulate Data | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Dependence** | $\rho$ | $\eta_z$ | $\eta_y$ | $\beta_{0y}$ | $\beta_{1y}$ | $\beta_{0z}$ | $\beta_{1z}$ | $\sigma^2$ |
| Weak | 0.2 | 0.3 | 1 | -1 | 0.5 | 1 | 0.5 | 1 |
| Moderate | 1 | 0.3 | 1 | -1 | 0.5 | 1 | 0.5 | 1 |
| Strong | 0.5 | 0.9 | 3.5 | -1 | 0.5 | 1 | 0.5 | 1 |

| **Model** | **Gaussian Cond. Mean Parameter** $\alpha(s_i)$ |
|---|---|
| Full (1) | $\beta_0+\beta_1 X_{z,i}+\frac{\eta_z}{4}\sum_{j\in N_i}\left(z(s_j)-\mu_j\right)+\rho\left(y(s_i)-\frac{\exp(\delta_i)}{1+\exp(\delta_i)}\right)$ |
| Constant Mean (2) | $\beta_0+\frac{\eta_z}{4}\sum_{j\in N_i}\left(z(s_j)-\mu_j\right)+\rho\left(y(s_i)-\frac{\exp(\delta_i)}{1+\exp(\delta_i)}\right)$ |
| Univariate Spatial (3) | $\beta_0+\beta_1 X_{z,i}+\beta_\rho Y(s_i)+\frac{\eta_z}{4}\sum_{j\in N_i}\left(z(s_j)-\mu_j\right)$ |
| Univariate Non-Spatial (4) | $\beta_0+\beta_1 X_{z,i}+\beta_\rho Y(s_i)$ |
| Bivariate Non-Spatial (5) | $\beta_0+\beta_1 X_{z,i}+\rho\left(y(s_i)-\frac{\exp(\delta_i)}{1+\exp(\delta_i)}\right)$ |

# MPLE Study Results - Dependence

- Only bivariate spatial models (1 and 2) capture the dependence structure adequately
- Strong dependence (far right column): only full bivariate spatial model obtains unbiased estimates for all three dependence parameters



# MPLE Study Results - Covariates

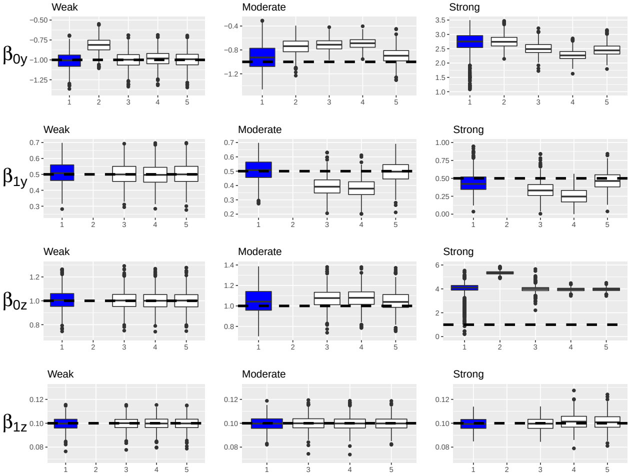



Figure: Mean Parameter Estimates for Simulated Data

# MPLE Study Summary

Summary

- Weak/Moderate dependence: all models provide unbiased estimates of the regression coefficients $\beta$
- Strong dependence: only full model is remotely close to estimating the true parameter values
- The strong dependence regime "breaks" the data-generating mechanism:
  - Expected marginal behavior is not observed due to local dependence effects

    $\Longrightarrow$ Caution for parametric bootstrap if we encounter strong dependence

# Study 2: Bootstrap C.I. Coverage

Goal: Investigate empirical vs. nominal confidence interval coverage under different dependence

Set-up:

- No covariates
- 2 dependence regimes: weak/small variance and strong/large variance
- 30 x 30 lattice
- 95% confidence intervals via parametric bootstrap using 500 bootstrap replicates for 1,000 simulated datasets

# Study 2: Bootstrap C.I. Coverage

- Generate $m = 1,000$ datasets from the true parameter values, $\Theta = \{\rho, \eta_z, \eta_y, \delta, \mu, \sigma^2\}$
- For each of the $i = 1, ..., 1000$ simulated datasets:
  - Estimate the parameters using MPLE to get $\hat{\Theta}_i$
  - Use $\hat{\Theta}_i$ to generate 500 bootstrap replicates for each dataset
  - For each of the bootstrap replicates, obtain MPLE estimates of the parameters and collect into empirical distribution $\Theta_i^*$ of $\Theta_i$ for data set $i$
  - Obtain $C = 95\%$ confidence intervals for $\theta \in \Theta$ by taking $U_i = 2 \cdot \hat{\theta}_i - \theta^*_{i,.025}$ and $L_i = 2 \cdot \hat{\theta}_i - \theta^*_{i,.975}$ as the bounds.
- For a Monte Carlo sample of size $m = 1000$, the empirical coverage is $C^* = (\frac{1}{1000}) \sum_{i=1}^{1000} \mathbb{1}\{\theta \in (L_i, U_i)\}$

# Study 2: Bootstrap C.I. Coverage

- Under weak or moderate dependence conditions, empirical coverage is close to 95%
- Slight undercoverage matches the results for the bivariate autologistic model from Caragea and Berg (2014)

| Weak Dependence | | | | |
|---|---|---|---|---|
| **Parameter** | **True Value** | **MC Estimate** | **MC Std. Error** | **95% CI Coverage** |
| $\rho$ | 0.2 | 0.202 | 0.073 | 92.9% |
| $\eta_z$ | 0.5 | 0.493 | 0.083 | 93.2% |
| $\eta_y$ | 1 | 0.959 | 0.419 | 94.5% |
| $\delta$ | 0.5 | 0.508 | 0.090 | 94.1% |
| $\mu$ | 1 | 1.001 | 0.052 | 93.6% |
| $\sigma^2$ | 1 | 0.996 | 0.053 | 92.9% |

Table: Bootstrap confidence interval coverage for weak dependence regime

# Study 2: Bootstrap C.I. Coverage

- Strong dependence: empirical coverage is further from the nominal 95% value for the unbiased dependence parameters
- Unable to obtain unbiased estimates or reliable confidence intervals for mean parameters $\delta$ and $\mu$

| **Strong Dependence** | | | | |
|---|---|---|---|---|
| **Parameter** | **True Value** | **MC Estimate** | **MC Std. Error** | **95% CI Coverage** |
| $\rho$ | 5 | 5.024 | 0.911 | 96.9% |
| $\eta_z$ | 0.9 | 0.901 | 0.042 | 86.5% |
| $\eta_y$ | 3 | 3.031 | 0.710 | 98.0% |
| $\delta$ | 0.5 | -1.683 | 1.103 | 49.4% |
| $\mu$ | 1 | -17.845 | 10.419 | 50.0% |
| $\sigma^2$ | 100 | 99.622 | 5.582 | 92.6% |

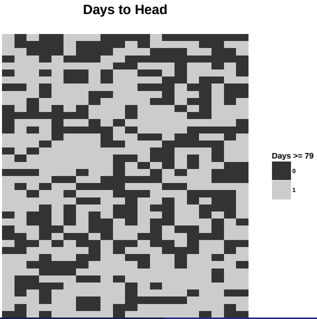
Table: Bootstrap confidence interval coverage for strong dependence regime

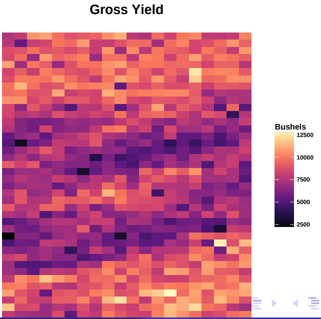
# Illustrative Examples

# Illustrative Example 1: Model Estimation and Selection

# Agricultural Field Trials - Set-Up

- Lado, et al. (2013) used basic spatial adjustments to aid in prediction of wheat yields
- Re-purpose yield data to implement bivariate mixed MRF model
- Goal: Evaluate model performance for full model vs. reduced models
    - Gaussian response $Z(s_i)$ : Gross Yield (1000s of bushels)
    - Binary response $Y(s_i)$: Slow/Fast Growth (1 if $>$ 78 days to head)
    - Thousand-Kernel Weight (g) used a covariate for both responses
    - Fields arranged on a 40 x 20 lattice



# Agricultural Field Trials - Estimation Results

Regression Coefficient Estimates (and 95% Confidence Intervals):

- Estimates are similar across models
- In 3 out of 4 models, Thousand-Kernel Weight has significant relationship with both response variables

| **Model** | **Binary (Y) Intercept ($\beta_{0,y}$)** | **1000-Kernel Wt $\beta_{1,y}$** | **Gaussian (Z) Intercept ($\beta_{0,z}$)** | **1000-Kernel Wt ($\beta_{1,z}$)** |
|---|---|---|---|---|
| Full | 6.591<br>(5.109 , 7.794) | -0.137<br>(-0.162, -0.106) | 6.842<br>(5.915, 7.606) | 0.022<br>(0.008 , 0.036) |
| Bivariate<br>No Covariates | 0.336<br>(0.130 , 0.534) | - | 8.014<br>(7.496 , 8.384) | - |
| Univariate | 6.600<br>(4.319 , 7.104) | -0.135<br>(-0.145 , -0.087) | 7.065<br>(6.969 , 8.370) | 0.019<br>(-0.011 , 0.019) |
| Univariate<br>Non-Spatial | 6.561<br>(5.177 , 7.818) | -0.134<br>(-0.161 , -0.105) | 5.330<br>(4.435 , 6.250) | 0.057<br>(0.038 , 0.077) |
| Bivariate<br>Non-Spatial | 6.873<br>(5.328 , 8.266) | -0.141<br>(-0.171 , -0.110) | 5.442<br>(4.485 , 6.507) | 0.055<br>(0.032 , 0.076) |

# Agricultural Field Trials - Estimation Results

Dependence Parameters:

- Estimates are quite different between bivariate and univariate models
- Moderate within-type (spatial) dependence
- Significant cross-dependence for 2/3 models

| **Model** | $\rho$ | $\eta_z$ | $\eta_y$ |
|---|---|---|---|
| Full Bivariate | 0.299 (0.127 , 0.474) | 0.888 (0.795 , 1.008) | 0.846 (0.000 , 1.947) |
| Bivariate No Covariates | 0.126 (-0.026 , 0.264) | 0.900 (0.809, 0.996) | 1.001 (0.135 , 2.001) |
| Univariate Spatial | - | 0.618 (0.634 , 0.921) | 1.232 (0.622 , 2.710) |
| Univariate Non-Spatial | - | - | - |
| Bivariate Non-Spatial | 0.233 (0.282 , 0.487) | - | - |

# Agricultural Field Trials - Model Selection

- We use continuous-ranked probability score to compare in-sample predictive capability of each model:

$$CRPS(F(x(s_i)|\boldsymbol{x}(\bar{s}_i), \hat{\boldsymbol{\Theta}}), x_i) = \int_{-\infty}^{\infty} (F(x(s_i)) - \mathbb{1}_{[x(s_i) \geq x]})^2 dx(s_i)$$

  where $F(x(s_i)|\boldsymbol{x}(\bar{s}_i), \hat{\boldsymbol{\Theta}})$ is the conditional CDF at each node
- Sum CRPS over all points to get CRPS(M) for the model
- Corresponding skill score for model $M$ vs. constant mean model $c$ is

$$Skill = 100 \cdot \left[1 - \frac{CRPS(M)}{CRPS(c)}\right]$$

- Minimize CRPS or maximize Skill to select best model

# Agricultural Field Trials - Model Selection

| **Model** | $CRPS_y$ | $Skill_y$ | $CRPS_z$ | $Skill_z$ |
|---|---|---|---|---|
| Full | 0.20 | 16.74 | 0.66 | 34.23 |
| Intercept-Only Spatial | 0.24 | 1.62 | 0.66 | 33.46 |
| Univariate | 0.21 | 14.82 | 0.70 | 29.57 |
| Univariate Non-Spatial | 0.22 | 8.11 | 0.97 | 2.59 |
| Bivariate Non-Spatial | 0.22 | 10.44 | 0.96 | 3.37 |

- Full bivariate spatial model with covariate has best in-sample prediction for both response variables
- Spatial models outperform non-spatial models for Gaussian response (Gross Yield)
- Covariate (Thousand-Kernel Weight) important for fitting the binary data

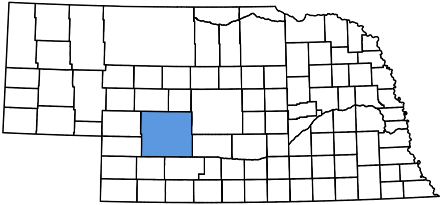
# Illustrative Example 2: Strong Spatial Dependence

# Greater Prairie Chickens - Data

Setting: Lincoln County (Western Nebraska)

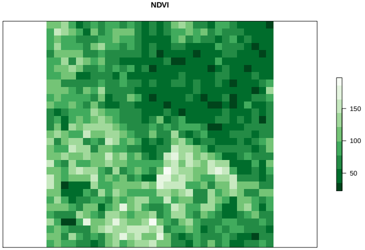
Data from three sources:

- Greater Prairie Chicken observations (March-May, since 1990) from Cornell Lab of Ornithology
- Vegetative Cover: April 2018 NDVI (Normalized Difference Vegetation Index) collected via satellite, from USDA National Agricultural Statistics Survey's VegScape API
- Soil properties (covariates) from NRCS Web Soil Survey



# Greater Prairie Chickens - Data

- Data sources overlaid/combined/aggregated onto 31 x 31 lattice (each point ~2 square miles)
- NDVI (continuous response) has very strong spatial autocorrelation
- Prairie Chicken nesting sites (binary response) has visible spatial autocorrelation, but somewhat sparse



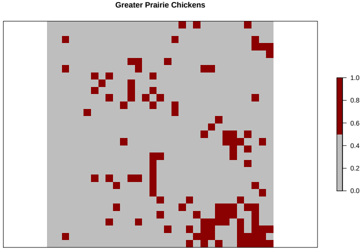




Figure: NDVI (left) and Greater Prairie Chicken observations (right)

# Greater Prairie Chickens - Methodology

Methodology:

- NDVI as continuous response $Z(s_i)$
- Greater Prairie Chicken observations (1 if observed since 1990, 0 if not) as binary response $Y(s_i)$
- Exploratory analysis to select possible soil covariates
- Estimate the full bivariate MRF model use 90% confidence intervals for inference
- Can we find any evidence for:
    - Either response being related to soil properties?
    - Spatial dependence or cross-dependence among NDVI and nesting sites?

# Greater Prairie Chickens - Estimation Results

| Parameter | Estimate | Std. Err. | 90% Bootstrap C.I. |
|---|---|---|---|
| Y-Intercept($\beta_{0,y}$) | -0.417 | 1.962 | (-2.343 , 4.183) |
| Wind Erodibility Index ($\beta_{1,y}$) | -0.005 | 0.006 | (-0.017 , 0.002) |
| % Sand ($\beta_{2,y}$) | 0.024 | 0.018 | ( 0.003 , 0.063) |
| Water Supply ($\beta_{3,y}$) | 0.019 | 0.045 | (-0.061 , 0.083) |
| Z-Intercept($\beta_{0,z}$) | 204.543 | 15.970 | (-43.172 , 66.560) |

Covariates:

- Soil variables not significantly related with either response, even with reduced confidence level
- Standard errors quite large for all variables except % Sand composition as covariate for Greater Prairie Chicken presence

# Greater Prairie Chickens - Estimation Results

| Parameter | Estimate | Std. Err. | 90% Bootstrap C.I. |
|---|---|---|---|
| $\eta_y$ (Presence-Absence) | 3.613 | 0.697 | (2.356 , 4.614) |
| $\eta_z$ (NDVI) | 0.991 | 0.017 | (0.955 , 1.012) |
| $\rho$ (Cross-dependence) | 0.446 | 0.260 | (0.000 , 0.844) |
| $\sigma^2$ | 16.765 | 0.949 | (15.246 , 18.245) |

Dependence Parameters:

- Moderate/Strong spatial correlation (near 4) for Greater Prairie Chicken presence
- NDVI spatial correlation is near the boundary of the parameter space (very strong dependence)
- Some cross-dependence among response variables
  - Weak evidence supporting a positive dependence between vegetative cover and probability of observing a Greater Prairie Chicken nest
- In-sample predictive capability was poor for Prairie Chicken presence-absence (Skill $< 0$)

# Conclusions and Future Work

# Summary

The proposed bivariate Gaussian-Binary Markov random field model:

- Provides interpretable dependence parameters in terms of conditional means by modeling dependence directly
- Allows for flexible specification of covariates related to each response
- Has easy-to-implement estimation and inference procedures from conditional specification
- Outperforms previous univariate auto-models in simulations and application
- Introduces mixed-response multivariate auto-models that can be extended to more complex applications

# Conclusions and Future Work

Areas for future work:

- Formally quantifying/defining/interpreting/remedying strong dependence settings
- Expanded/irregular neighborhood structures
- Extension to other distributions and higher-dimensional responses
- Faster/more exact inference methods

QUESTIONS?

# References


Besag, J. (1974). Spatial Interaction and the Statistical Analysis of Lattice Systems. Journal of the Royal Statistical Society. Series B (Methodological), 36(2), 192-236.

Besag, J. (1975). Statistical Analysis of Non-Lattice Data. Journal of the Royal Statistical Society. Series D (The Statistician), 24(3), 179-195.

Caragea, P.C. and Berg, E. (2014). A centered bivariate spatial regression model for binary data with an application to presettlement vegetation data in the Midwestern United States. J. Agric. Biol. Environ. Stat. 19, 453471.

Chen, S., Witten, D. M., and Shojaie, A. (2014). Selection and estimation for mixed graphical models. Biometrika, 102(1), 47-64.

GBIF.org (31 October 2018). GBIF Occurrence Download. https://doi.org/10.15468/

Hardouin, C. and Cressie, N. (2018). Two-scale Spatial Models for Binary Data. N. Stat Methods Appl (2018) 27: 1.

Hughes, J., Haran, M. and Caragea, P. C. (2011), Autologistic models for binary data on a lattice. Environmetrics, 22: 857-871.

# References


Kaiser, M.S. (2007). Statistical Dependence in Markov Random Field Models. Statistics Preprints, 57. https://lib.dr.iastate.edu/stat_las_preprints/57 .

Kaiser, M. S., Caragea, P. C., and Furakawa, K. (2012). Centered Parameterizations and Dependence Limitations in Markov Random Field Models. Journal of Statistical Planning and Inference, Volume 142, Issue 7, 2012, 1855-1863, ISSN 0378-3758.

Lado, B., Matus, I., Rodrguez, A., Inostroza, L., Poland, J., Belzile, F., del Pozo, A., Quincke, M., Castro, M., von Zitzewitz, J. (2013). Increased genomic prediction accuracy in wheat breeding through spatial adjustment of field trial data. G3 (Bethesda, Md.), 3(12), 2105-14. doi:10.1534/g3.113.007807

Lee, J. and Hastie, T.J. (2015). Learning the Structure of Mixed Graphical Models. Journal of Computational and Graphical Statistics, 24:1, 230-253.

Sain, S., Furrer, R., and Cressie, N. (2011). A spatial analysis of multivariate output from regional climate models. The Annals of Applied Statistics, 5(1), 150-175.

Soil Survey Staff, Natural Resources Conservation Service, United States Department of Agriculture. Web Soil Survey. Available online at the following link: https://websoilsurvey.sc.egov.usda.gov/. (Accessed: 31 October 2018).

Yang, Z., Yu, G., Di, L., Zhang, B., Han, W., and Mueller, R. (2013). Web service-based vegetation condition monitoring system - VegScape, Geoscience and Remote Sensing Symposium (IGARSS) 2013 IEEE International, 3638-3641.