Rare Spatial Binary Regression

Samuel Morris (NC State)
Brian Reich (NC State)

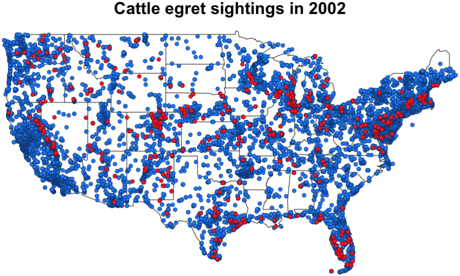
JSM 2015, Seattle

Motivation

- Modeling and predicting rare occurrences in a spatial setting
  - Rare: to mean not occurring often (e.g. 5% or less)
  - Based on binary regression, but incorporating methods for spatial extremes
- Application: Species mapping (eBirds)
  - Cornell Lab of Ornithology and National Audubon Society
  - In 2002, almost 50,000 bird sightings (starting year)
  - March 2012, more than 3.1 million observations
  - Presence/absence data for over 1750 species

Species 1: Cattle egret

Cattle egret (Bubulcus ibis):



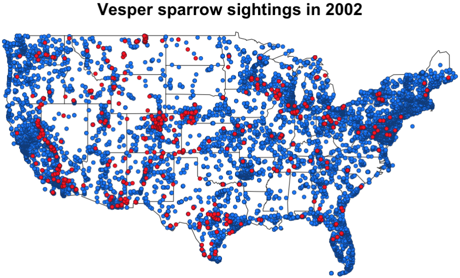
- eBird frequency: 2%
- Photo credit: Manjith Kainickara (Wikipedia)

All reported sightings in 2002. Red indicates a cattle egret sighting

Species 2: Vesper sparrow

Vesper sparrow (Pooecetes gramineus):

- eBird frequency: 1–2%
- Photo credit: Tripp Davenport (Flickr)



All reported sightings in 2002. Red indicates a vesper sparrow sighting

Binary regression

- Response is either 0 or 1
- Goal is to understand $P(Y_i = 1) = g(\mathbf{X}_i\boldsymbol{\beta})$ where
  - $\mathbf{X}_i$ is a p-vector of covariates for response i
  - $\boldsymbol{\beta}$ is a p-vector of regression parameters
  - $g(\cdot) : \mathcal{R} \to (0, 1)$

Binary regression

- Common link functions:
  - Logit:

$$g(\mathbf{X}_i\beta) = \frac{\exp(\mathbf{X}_i\beta)}{1+\exp(\mathbf{X}_i\beta)}$$

  - Probit:

$$g(\mathbf{X}_i\beta) = \Phi(\mathbf{X}_i\beta)$$

    where Φ is the standard normal CDF

  - Cloglog:

$$g(\mathbf{X}_i\beta) = 1 - \exp[\exp(\mathbf{X}_i\beta)]$$

Generalized extreme value (Wang and Dey, 2010)

- Link function is defined as

$$g(z_i) = 1 - \exp(-z_i)$$

where

$$z_i = \begin{cases} (1 - \xi \mathbf{X}_i \boldsymbol{\beta})^{-1/\xi} & \xi \neq 0 \\ \exp(-\mathbf{X}_i \boldsymbol{\beta}) & \xi = 0 \end{cases}$$

is standardized to give unit Fréchet distribution.

- Note: The cloglog link is a special case when $\xi = 0$

Spatial setting: Logit and probit

- For logit and probit settings:
  - Assume an underlying Gaussian process for the latent variable

$$\mathbf{Z} \sim \mathrm{N}_n(\mathbf{X}\beta, \mathbf{\Sigma})$$

    where
    - $\mathbf{X}$ is an $n \times p$ matrix of covariates
    - β is defined as before
    - $\mathbf{\Sigma}$ is an $n \times n$ positive-definite covariance matrix
  - Conditional on $z(\mathbf{s}_i)$

$$Y(\mathbf{s}_i) \overset{ind}{\sim} \mathrm{Bern}\{g[z(\mathbf{s}_i)]\}$$

- Problem: Asymptotic dependence for a multivariate Gaussian distribution is 0 unless correlation is 1.

Spatial setting: GEV

- If we believe the underlying distribution is extremal, then the dependence structure should match
- Multivariate GEV distributions are more challenging to work with than multivariate Gaussian distributions
- Interested in the asymptotic dependence (i.e. dependence in the tail of the distribution):
  - Extremal index: Effective number of independent replications
  - χ-statistic:

$$\chi = \lim_{c \to c^*} P[Y(\mathbf{s}_2) > c \mid Y(\mathbf{s}_1) > c]$$

where c^* is the upper limit of the support of Y

Max-stable processes

- Max-stable process is the extremal analogue to the Gaussian process
- Dependence structures are very flexible, but can be very challenging to work with in high dimensions
    - Pairwise composite likelihood (Padoan et al., 2010)
    - Recent work allows for higher dimensions (Engelke et. al, 2014; Wadsworth and Tawn, 2014)

Dimension reduction

- Problem: For very large n computational challenges arise
- Consider a set of $L << n$ knots $\mathbf{v}_1, \dots, \mathbf{v}_L$
- We assume that the latent variables at the n locations can be represented by a function of L random effects
  - Logit and probit methods use Gaussian predictive process
  - For the GEV, we propose using the hierarchical model for extremes by Reich and Shaby (2012)

Random effects representation

- Logit and probit use a Gaussian predictive process

$$\begin{bmatrix} \mathbf{z}_n \\ \mathbf{z}_L \end{bmatrix} \sim \mathrm{N}_{n+L}\left(\begin{bmatrix} \mathbf{X}_n \\ \mathbf{X}_L \end{bmatrix} \boldsymbol{\beta}, \begin{bmatrix} \Sigma_{nn} & \Sigma_{nL} \\ \Sigma_{Ln} & \Sigma_{LL} \end{bmatrix} \right)$$

- We fit the model using the latent variables at the knot locations
- Use distribution of $\mathbf{z}_n|\mathbf{z}_L$ to get back distribution at all sites

Max-stable processes: A hierarchical representation (Reich & Shaby, 2012)

- Let $\widetilde{\mathbf{Y}} \sim \mathrm{GEV}_n[\mu(\mathbf{s}), \sigma(\mathbf{s}), \xi(\mathbf{s})]$ be a realization from multivariate generalized extreme value distribution
- Consider a set of L knots, $\mathbf{v}_1, \dots, \mathbf{v}_L$
- Model the spatial dependence using

$$\theta(\mathbf{s}) = \left[\sum_{l=1}^{L} A_l w_l(\mathbf{s})^{1/\alpha}\right]^{\alpha}$$

  where
  - A_l are i.i.d. positive stable random effects
  - $w_l(\mathbf{s})$ are a set of non-negative scaled kernel basis functions, scaled so that $\sum_{l=1}^{L} w_l(\mathbf{s}) = 1$
  - $\alpha \in (0, 1)$ is a parameter controlling strength of spatial dependence (0: high, 1: independent)

Max-stable processes: A hierarchical representation (Reich & Shaby, 2012)

- When conditioning on θ

$$\widetilde{Y}(\mathbf{s}_i) \mid A_l \overset{ind}{\sim} \text{GEV}[\mu^*(\mathbf{s}_i), \sigma^*(\mathbf{s}_i), \xi^*(\mathbf{s}_i)]$$

$$A_l \overset{iid}{\sim} \text{PS}(\alpha)$$

where

- $\mu^*(\mathbf{s}_i) = \mu(\mathbf{s}) + \frac{\sigma(\mathbf{s})}{\xi(\mathbf{s})}[\theta(\mathbf{s})^{\xi(\mathbf{s})} - 1]$
- $\sigma^*(\mathbf{s}_i) = \alpha\sigma(\mathbf{s})\theta(\mathbf{s})^{\xi(\mathbf{s})}$
- $\xi^*(\mathbf{s}) = \alpha\xi(\mathbf{s})$

Proposed Method

- Fit a hierarchical random effects model using MCMC
    - Extends model from Reich and Shaby (2012)
    - Using random walk Metropolis-Hastings algorithm
    - Pairwise composite likelihood estimates used for initial values and hyperparameters in MCMC

Hierarchical MCMC

- Data:

$$Y(\mathbf{s}_i)|g[z(\mathbf{s}_i)] \overset{ind}{\sim} \text{Bern}\{g[z(\mathbf{s}_i)]\}$$

- Latent process:

$$g[z(\mathbf{s}_i)] = 1 - \exp\left\{\sum_{l=1}^{L} A_l \left[\frac{w_l(\mathbf{s}_i)}{z_i}\right]^{1/\alpha}\right\}$$

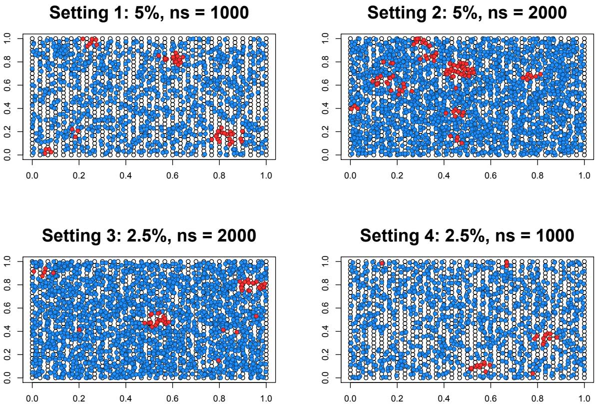
where

$$z_i = \begin{cases} (1 - \xi \mathbf{X}_i \beta)^{-1/\xi} & \xi \neq 0 \\ \exp(-\mathbf{X}_i \beta) & \xi = 0 \end{cases}$$

$$A_l \overset{iid}{\sim} \text{PS}(\alpha)$$

Simulation study: Settings

- Conducting a simulation study with 50 datasets generated from our model with strong spatial dependence with knots on a 31×31 grid
- Looking at impact of number of observations as well as prevalence



Simulation study: Methods

- Fitting Bayesian models for spatial probit, spatial logit (`spBayes::spGLM`) and spatial GEV (fixing $\xi = 0$ for identifiability)
- Model fit using 75% of the observations as a training set and 25% for cross-validation
- Measuring performance with the Brier score (Gneiting and Raftery, 2007)

Simulation study: Preliminary results and future work

- Preliminary results:
  - Our model demonstrates some improvement over the spatial probit model (75%–85% reduction in Brier score)
  - Using adaptive MCMC with `spBayes::spGLM` is very slow with this many knots, and still waiting on results (on order of 5–6 times longer on a single-threaded BLAS)
- Future work:
  - Cluster size and smoothness of the field impacts which methods do better
  - Exploring impact of reducing number of knots
  - Data analysis with eBirds data

Questions

- Questions?
- Thank you for your attention.

References I

- Coles, S., Heffernan, J. and Tawn, J. (1999) Dependence Measures for Extreme Value Analyses. *Extremes*, **2**, 339–365.

- Engelke, S., Malinowski, A., Kabluchko, Z. and Schlather, M. (2014) Estimation of Hüsler-Reiss distributions and Brown-Resnick processes. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, **77**, 239–265.

- Padoan, S. A., Ribatet, M. and Sisson, S. A. (2010) Likelihood-Based Inference for Max-Stable Processes. *Journal of the American Statistical Association*, **105**, 263–277.

- Reich, B. J. and Shaby, B. A. (2012) A hierarchical max-stable spatial model for extreme precipitation. *The Annals of Applied Statistics*, **6**, 1430–1451.

- Sullivan, B. L., Wood, C. L., Iliff, M. J., Bonney, R. E., Fink, D. and Kelling, S. (2009) eBird: A citizen- based bird observation network in the biological sciences. *Biological Conservation*, **142**, 2282–2292.

References II

- Wadsworth, J. L. and Tawn, J. a. (2014) Efficient inference for spatial extreme value processes associated to log-Gaussian random func- tions. *Biometrika*, **101**, 1–15.

- Wang, X. and Dey, D. K. (2010) Generalized extreme value regression for binary response data: An ap- plication to B2B electronic payments system adoption. *The Annals of Applied Statistics*, **4**, 2000–2023.