A spatial model for rare binary events

August 7, 2016

1 Introduction

The goals of spatial binary data analysis are often to estimate covariate effects while accounting for spatial dependence and to make predictions at locations without samples. A common approach to incorporate spatial dependence in the model for binary data is related to a continuous spatial process $Z(\mathbf{s}) \in \mathbb{R}$ to the binary response $Y(\mathbf{s})$ by thresholding $Y(\mathbf{s}) = I[Z(\mathbf{s}) > c]$ where $I[\cdot]$ is an indicator function. In many spatial analyses of binary data, a Gaussian process is used to model $Z(\mathbf{s})$. This is true for both spatial probit and spatial logistic regression.

In this model, spatial dependence is determined by the joint probability that two sites simultaneously exceed the threshold c. However, when c is large, and thus $Y(\mathbf{s}) = 1$ is rare, then the asymptotic theory suggests that the Gaussian process will not do well at modeling dependence. In fact, even under strong spatial correlation for Z, it gives asymptotic independence. Therefore, this suggests that for rare binary data, the Gaussian model will not perform very well.

We propose using a latent max-stable process (de Haan, 1984) because it allows for asymptotic dependence. The max-stable process arises as the limit of the location-wise maximum of infinitely many spatial processes, and any finite-dimensional representation of a max-stable process has generalized extreme value distribution (GEV) marginal distributions. Max-stable processes are extremely flexible, but are often challenging to work with in high dimensions (Wadsworth and Tawn, 2014; Thibaud and Opitz, 2015). To address this challenge, methods have been proposed that implement composite likelihood techniques for max-stable processes (Padoan et al., 2010; Genton et al., 2011; Huser and Davison, 2014). Composite likelihoods have been used to model binary spatial data (Heagerty and Lele, 1998), but this is not using max-stable processes.

As an alternative to these composite approaches, Reich and Shaby (2012) present a hierarchical model that implements a low-rank representation for a max-stable process. We chose to use this low-rank representation for our rare binary spatial regression model. Our model builds on related work by Wang and Dey (2010) who use a GEV link for non-spatial binary data. The proposed model generalizes this to have spatial dependence.

The paper proceeds as follows. In Section 2 we present the proposed latent max-stable process for rare binary data analysis. In Section 3 we give the bivariate distribution for our model. In Section 4 we show a link between a commonly used measure of dependence between binary variables and another metric for extremal dependence. The computing for our model is outlined in Section 5. Finally, we present a simulation study in Section 6, and a data analysis on *Tamarix ramosissima* in Section 7. Lastly, in Section 8 we provide some discussion and possibilities for future research.

2 Spatial dependence for binary regression

Let $Y(\mathbf{s})$ be the binary response at spatial location $\mathbf{s}$ in a spatial domain of interest $\mathcal{D} \in \mathcal{R}^2$. We assume $Y(\mathbf{s}) = I[Z(\mathbf{s}) > 0]$ where $Z(\mathbf{s})$ is a latent continuous max-stable process. The marginal distribution of $Z(\mathbf{s})$ at site $\mathbf{s}$ is GEV with location $\mathbf{X}(\mathbf{s})^\top \boldsymbol{\beta}$, scale $\sigma > 0$, and shape ξ, where $\mathbf{X}(\mathbf{s})$ is a p-vector of spatial covariates at site $\mathbf{s}$ and $\boldsymbol{\beta}$ is a p-vector of regression coefficients. We set $\sigma = 1$ for identifiability because only the sign and not the scale of Z affects Y. If $\mathbf{X}(\mathbf{s})^\top \boldsymbol{\beta} = \mu$ for all $\mathbf{s}$, then $\mathrm{P}(Y = 1)$ is the same for all observations, and the two parameters μ and ξ are not individually identifiable, so when there are no covariates, we fix $\xi = 0$. Although $\boldsymbol{\beta}$ and ξ could be permitted to vary across space, we assume that they are constant across $\mathcal{D}$. At spatial location $\mathbf{s}$, the marginal distribution (over $Z(\mathbf{s})$) is $P[Y(\mathbf{s}) = 1] = 1 - \exp\left[-\frac{1}{z(\mathbf{s})}\right]$ where $z(\mathbf{s}) = \left[1 - \xi \mathbf{X}(\mathbf{s})^\top \boldsymbol{\beta}\right]^{1/\xi}$. This is the same as the marginal distribution given by Wang and Dey (2010).

For a finite collection of locations $\mathbf{s}_1, \ldots, \mathbf{s}_n$, denote the vector of observations $\mathbf{Y} = [Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_n)]^T$. The spatial dependence of $\mathbf{Y}$ is determined by the joint distribution of $\mathbf{Z} = [Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n)]^T$. To incorporate spatial dependence, we consider the hierarchical representation of the max-stable process proposed in Reich and Shaby (2012). Consider a set of positive stable random effect $A_1, \ldots, A_L \stackrel{iid}{\sim} \text{PS}(\alpha)$ associated with spatial knots $\mathbf{v}_1, \ldots, \mathbf{v}_L \in \mathcal{R}^2$. The hierarchical model is given by

$$\mathbf{Z}(\mathbf{s}_i)|A_1, \ldots, A_L \stackrel{indep}{\sim} \text{GEV}[\mathbf{X}(\mathbf{s}_i)^\top \boldsymbol{\beta} + \theta(\mathbf{s}_i), \alpha\theta(\mathbf{s}_i), \xi\alpha] \quad \text{and} \quad \theta(\mathbf{s}_i) = \left[\sum_{l=1}^{L} A_l w_l(\mathbf{s}_i)^{1/\alpha}\right]^\alpha \tag{1}$$

where $w_l(\mathbf{s}_i) > 0$ are a set of L weights that vary smoothly across space and satisfy $\sum_{l=1}^{L} w_l(\mathbf{s}) = 1$ for all $\mathbf{s}$, and $\alpha \in (0, 1)$ determines the strength of dependence, with α near zero giving strong dependence and $\alpha = 1$ giving joint independence.

Because the latent $\mathbf{Z}(\mathbf{s})$ are independent given the random effects $\theta(\mathbf{s})$, the binary responses are also conditionally independent. This leads to the tractable likelihood

$$Y(\mathbf{s}_i)|A_l, \ldots, A_L \stackrel{indep}{\sim} \text{Bern}[\pi(\mathbf{s}_i)] \tag{2}$$

where

$$\pi(\mathbf{s}_i) = 1 - \exp\left\{-\sum_{l=1}^{L} A_l \left(\frac{w_l(\mathbf{s}_i)}{z(\mathbf{s}_i)}\right)^{1/\alpha}\right\}. \tag{3}$$

Marginally over the A_l, this gives

$$Z(\mathbf{s}) \sim \text{GEV}(\mathbf{X}(\mathbf{s})^\top \boldsymbol{\beta}, 1, \xi), \tag{4}$$

and thus $\mathrm{P}[Y(\mathbf{s}) = 1] = 1 - \exp\left\{-\frac{1}{z(\mathbf{s})}\right\}$ where $z(\mathbf{s}) = [1 - \xi \mathbf{X}(\mathbf{s})\boldsymbol{\beta}]^{1/\xi}$.

Many weight functions are possible, but the weights must be constrained so that $\sum_{l=1}^{L} w_l(\mathbf{s}_i) = 1$ for $i = 1, \ldots, n$ to preserve the marginal GEV distribution. For example, Reich and Shaby (2012) take the weights to be scaled Gaussian kernels with knots $\mathbf{v}_l$,

$$w_l(\mathbf{s}_i) = \frac{\exp\left[-0.5\left(||\mathbf{s}_i - \mathbf{v}_l||/\rho\right)^2\right]}{\sum_{j=1}^{L} \exp\left[-0.5\left(||\mathbf{s}_i - \mathbf{v}_j||/\rho\right)^2\right]} \tag{5}$$

where $||\mathbf{s}_i - \mathbf{v}_l||$ is the distance between site $\mathbf{s}_i$ and knot $\mathbf{v}_l$, and the kernel bandwidth $\rho > 0$ determines the spatial range of the dependence, with large ρ giving long-range dependence and vice versa.

After marginalizing out the positive stable random effects, the joint distribution of $\mathbf{Z}$ is

$$G(\mathbf{z}) = P\left[Z(\mathbf{s}_1) < z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n) < z(\mathbf{s}_n)\right] = \exp\left\{-\sum_{l=1}^{L}\left[\sum_{i=1}^{n}\left(\frac{w_l(\mathbf{s}_i)}{z(\mathbf{s}_i)}\right)^{1/\alpha}\right]^{\alpha}\right\}, \tag{6}$$

where $G(\cdot)$ is the CDF of a multivariate GEV distribution. This is a special case of the multivariate GEV distribution with asymmetric Laplace dependence function (Tawn, 1990).

3 Joint distribution

We give an exact expression in the case where there are only two spatial locations which is useful for constructing a pairwise composite likelihood (Padoan et al., 2010) and studying spatial dependence. When

$n = 2$, the probability mass function is given by

$$\mathrm{P}[Y(\mathbf{s}_i) = y_i, Y(\mathbf{s}_j) = y_j] = \begin{cases} \varphi(\mathbf{z}) & y_i = 0, y_j = 0 \\ \exp\left\{-\dfrac{1}{z(\mathbf{s}_i)}\right\} - \varphi(\mathbf{z}), & y_i = 1, y_j = 0 \\ \exp\left\{-\dfrac{1}{z(\mathbf{s}_j)}\right\} - \varphi(\mathbf{z}), & y_i = 0, y_j = 1 \\ 1 - \exp\left\{-\dfrac{1}{z(\mathbf{s}_i)}\right\} - \exp\left\{-\dfrac{1}{z(\mathbf{s}_j)}\right\} + \varphi(\mathbf{z}), & y_i = 1, y_j = 1 \end{cases} \tag{7}$$

where $\varphi(\mathbf{z}) = \exp\left\{-\sum_{l=1}^{L}\left[\left(\frac{w_l(\mathbf{s}_i)}{z(\mathbf{s}_i)}\right)^{1/\alpha} + \left(\frac{w_l(\mathbf{s}_j)}{z(\mathbf{s}_j)}\right)^{1/\alpha}\right]^{\alpha}\right\}$. For more than two locations, we are also able to compute the exact likelihood when the n is large but the number of events $K = \sum_{i=1}^{n} Y(\mathbf{s}_i)$ is small, as might be expected for very rare events, see Appendix A.2.

4 Quantifying spatial dependence

Assume that Z_1 and Z_2 are both GEV$(\beta, 1, 1)$ so that $\mathrm{P}(Y_i = 1)$ decreases to zero as β increases. A common measure of dependence between binary variables is Cohen's Kappa (Cohen, 1960),

$$\kappa(\beta) = \frac{P_A - P_E}{1 - P_E} \tag{8}$$

where P_A is the joint probability of agreement $P(Y_1 = Y_2)$ and P_E is the joint probability of agreement under an assumption of independence $P(Y_i = 1)^2 + P(Y_i = 0)^2$. For the spatial model,

$$\begin{aligned} P_A(\beta) &= 1 - 2\exp\left\{-\frac{1}{\beta}\right\} + 2\exp\left\{-\frac{\vartheta(\mathbf{s}_1, \mathbf{s}_2)}{\beta}\right\} \\ P_E(\beta) &= 1 - 2\exp\left\{-\frac{1}{\beta}\right\} + 2\exp\left\{-\frac{2}{\beta}\right\}, \end{aligned}$$

and

$$\kappa(\beta) = \frac{P_A(\beta) - P_E(\beta)}{1 - P_E(\beta)} = \frac{\exp\left\{-\frac{\vartheta(\mathbf{s}_1,\mathbf{s}_2)-1}{\beta}\right\} - \exp\left\{-\frac{1}{\beta}\right\}}{1 - \exp\left\{-\frac{1}{\beta}\right\}} \tag{9}$$

where $\vartheta(\mathbf{s}_i, \mathbf{s}_j) = \sum_{l=1}^{L} \left[w_l(\mathbf{s}_i)^{1/\alpha} + w_l(\mathbf{s}_j)^{1/\alpha}\right]^{\alpha}$ is the pairwise extremal coefficient given by Reich and Shaby (2012). To measure extremal dependence, let $\beta \to \infty$ so that events are increasingly rare. Then,

$$\kappa = \lim_{\beta \to \infty} \kappa(\beta) = 2 - \vartheta(\mathbf{s}_1, \mathbf{s}_2) \tag{10}$$

which is the same as the χ statistic of Coles (2001), a commonly used measure of extremal dependence.

5 Computation

For small K, we can evaluate the likelihood directly. When K is large, we use Markov chain Monte Carlo (MCMC) methods with the random effects model to explore the posterior distribution. To overcome challenges with evaluating the positive stable density, we follow Reich and Shaby (2012) and introduce a set of auxiliary variables $B_1, \ldots, B_L$ following the auxiliary variable technique of Stephenson (2009) (for more details, see Appendix A.3 of Reich and Shaby (2012)). So, the hierarchical model is given by

$$\begin{aligned} Y(\mathbf{s}_i) | \pi(\mathbf{s}_i) &\overset{indep}{\sim} \text{Bern}[\pi(\mathbf{s}_i)] \\ \pi(\mathbf{s}_i) &= 1 - \exp\left\{ - \sum_{l=1}^{L} A_l \left(\frac{w_l(\mathbf{s}_i)}{z(\mathbf{s}_i)} \right)^{1/\alpha} \right\} \\ A_l &\sim \text{PS}(\alpha) \end{aligned} \tag{11}$$

with priors $\boldsymbol{\beta} \sim \text{N}(\mathbf{0}, \sigma^2_\beta \mathbf{I}_p)$, $\xi \sim \text{N}(0, \sigma^2_\xi)$, $\rho \sim \text{Unif}(\rho_l, \rho_u)$, and $\alpha \sim \text{Beta}(a_\alpha, b_\alpha)$. The model parameters are updated using Metropolis Hastings (MH) update steps, and the random effects $A_1, \ldots, A_L$, and auxiliary variables $B_1, \ldots, B_L$ are updated using Hamiltonian Monte Carlo (HMC) update steps. The code for this is available online through `https://github.com/sammorris81/rare-binary`.

6 Simulation study

For our simulation study, we generate $n_m = 50$ datasets under 12 different simulation settings to explore the impact of sample size, sampling technique, and misspecification of link function. We generate data assuming three possible types of underlying process. For each of the underlying processes, we generate complete datasets on a 100×100 rectangular grid of $n = 10,000$ locations. If a dataset is generated with $K < 50$, it is discarded and a new dataset is generated. This is done to guarantee that datasets have no less than 0.5% rarity. For model fitting, we select a subsample and use the remaining sites to evaluate predictive performance.

6.1 Latent processes

The first process is a latent max-stable process that uses the GEV link described in (1) with knots on a 50×50 regularly spaced grid on $[0, 1] \times [0, 1]$. For this process, we set $\alpha = 0.35$, $\rho = 0.1$, and $\beta_0 \approx 2.97$ which gives $K = 500$, on average. Because there are no covariates, we set $\xi = 0$. We then set $Y(\mathbf{s}) = I[Z(\mathbf{s}) > 0]$.

For the second process, we generate a latent variable from a spatial Gaussian process with a mean of $\text{logit}(0.05) \approx -2.94$ and an exponential covariance given by

$$\text{cov}(\mathbf{s}_1, \mathbf{s}_2) = \tau^2_{\text{Gau}} \exp\left\{ -\frac{||\mathbf{s}_1 - \mathbf{s}_2||}{\rho_{\text{Gau}}} \right\} \qquad (12)$$

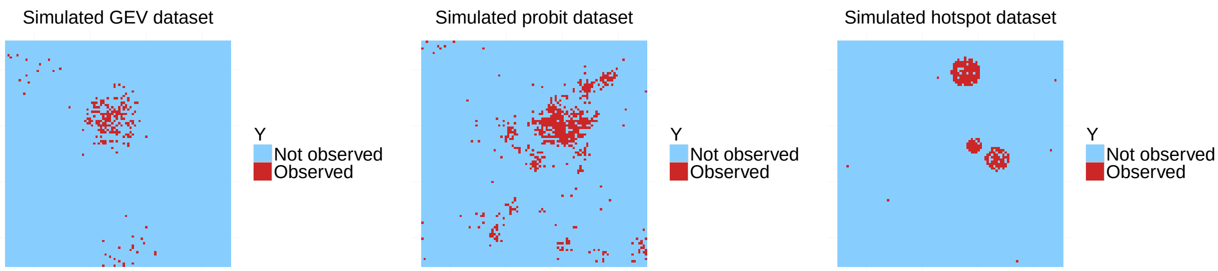

Figure 1: One simulated dataset from spatial GEV (left), spatial logistic (center), and hotspot (right).

where $\tau_{\text{Gau}} = 10$ and $\rho_{\text{Gau}} = 0.1$. Finally, we generate $Y(\mathbf{s}_i) \overset{ind}{\sim} \text{Bern}[\pi(\mathbf{s}_i)]$ where $\pi(\mathbf{s}_i) = \dfrac{\exp\{z(\mathbf{s})\}}{1 + \exp\{z(\mathbf{s})\}}$

For the third process, we generate data using a hotspot method. For this process, we first generate hotspots throughout the space. Let n_{hs} be the number of hotspots in the space. Then $n_{\text{hs}} - 1 \sim \text{Poisson}(2)$. This generation scheme ensures that every dataset has at least one hotspot. We generate the hotspot locations $\mathbf{h}_1, \ldots, \mathbf{h}_{n_{\text{hs}}} \sim \text{Unif}(0, 1)^2$. Let B_h be a circle of radius of radius r_h around hotspot $h = 1, \ldots, n_{\text{hs}}$. The r_h differ for each hotspot and are generated i.i.d. from a $\text{Unif}(0.03, 0.08)$ distribution. We set $\text{P}[Y(\mathbf{s}_i) = 1] = 0.85$ for all $\mathbf{s}_i$ in B_h, and $\text{P}[Y(\mathbf{s}_i)] = 0.0005$ for all $\mathbf{s}_i$ outside of B_h. These settings are selected to give an average of approximately $K = 500$ for the datasets. Figure 1 gives an example dataset from each of the data settings.

6.2 Methods

For each dataset, we fit the model using three different models, the proposed spatial GEV model, a spatial probit model, and a spatial logistic model. Logistic and probit methods represent two of the more common spatial techniques for binary data, we chose to compare our method to them. One way these methods differ from our proposed method is that they assume the underlying process is Gaussian. In this case, we assume that $Z(\mathbf{s})$ follows a Gaussian process with mean $\mathbf{X}(s)^\top\boldsymbol{\beta}$. The marginal distributions are given by

$$P\left[Y(\mathbf{s})=1\right]=\begin{cases}\dfrac{\exp\left[\mathbf{X}^\top(\mathbf{s})\boldsymbol{\beta}+\mathbf{W}(\mathbf{s})\boldsymbol{\epsilon}\right]}{1+\exp\left[\mathbf{X}^\top(\mathbf{s})\boldsymbol{\beta}+\mathbf{W}(\mathbf{s})\boldsymbol{\epsilon}\right]}, & \text{logistic}\\ \Phi\left[\mathbf{X}^\top\boldsymbol{\beta}(\mathbf{s})+\mathbf{W}(\mathbf{s})\boldsymbol{\epsilon}\right], & \text{probit}\end{cases} \tag{13}$$

where $\boldsymbol{\epsilon}\sim\mathrm{N}(\mathbf{0},\tau_L^2\mathbf{I}_L)$ are Gaussian random effects at the knot locations, and $\mathbf{W}(\mathbf{s})$ are a set of L basis functions given to recreate the Gaussian process at all sites. We use our own code for the spatial probit model, but we use the `spGLM` function in the `spBayes` package (Finley et al., 2015) to fit the spatial logistic model. For the probit model, we use

$$\mathbf{W}_l(\mathbf{s}_i)=\frac{\exp\left[-\left(||\mathbf{s}_i-\mathbf{v}_l||/\rho\right)^2\right]}{\sqrt{\sum_{j=1}^{L}\exp\left[-\left(||\mathbf{s}_i-\mathbf{v}_j||/\rho\right)^2\right]^2}}. \tag{14}$$

For the logistic model, the $\mathbf{W}_l(\mathbf{s}_i)$ are the default implementation from the `spGLM`. Timings

6.3 Sampling technique

We subsample the generated data using $n_s=100$, 250 initial locations for two different sampling designs. The first is a two-stage spatially-adaptive cluster technique (CLU) taken from Pacifici et al. (2016). In this design, if an initial location is occupied, we also include the four rook neighbor (north, east, south, and west) sites in the sample. For the second design, we use a simple random sample (SRS) with the same number of sites included in the cluster sample. For the GEV setting, when $n_s=100$, there are on average 117 sites and at most 142 sites in a sample, and when $n_s=250$, there are on average 286 sites and at most 332 sites in a sample. For the logistic setting, when $n_s=100$, there are on average 118 sites and at most 147 sites in a sample, and when $n_s=250$, there are on average 290 sites and at most 330 sites in a sample. For the hotspot setting, when $n_s = 100$, there are on average 110 sites and at most 128 sites in a sample, and when $n_s = 250$, there are on average 275 sites and at most 306 sites in a sample.

6.4 Priors

For all models, we only include an intercept term β_0 in the model, and the prior for the intercept is $\beta_0 \sim \text{N}(0, 10)$. Additionally, for all models, the prior for the bandwidth is $\rho \sim \text{Unif}(0.001, 1)$. In all methods, we place knots at all data points. For the GEV method, the prior for the spatial dependence parameter is $\alpha \sim \text{Beta}(2, 5)$. We select this prior because it gives greater weight to $\alpha < 0.5$, which is the point at which spatial dependence becomes fairly week, but also avoids values below 0.1 which can lead to numerical problems. We fix $\xi = 0$ because we do not include any covariates. For both the spatial probit and logistic models, the prior on the variance term for the random effects is $\text{IG}(0.1, 0.1)$ where $\text{IG}(\cdot)$ is an Inverse Gamma distribution. For all models, we run the MCMC sampler for 25,000 iterations with a burn-in period of 20,000 iterations. Convergence is assessed through visual inspection of traceplots.

6.5 Model comparisons

For each dataset, we fit the model using the n_s observations as a training set, and validate the model's predictive power at the remaining grid points. Let $\mathbf{s}_j^*$ be the jth site in the validation set. From the posterior distributions of the parameters we can calculate $P[Y(\mathbf{s})_j^* = 1]$. To obtain $\hat{P}[Y(\mathbf{s}_j^*) = 1]$, we take the average of the posterior distribution for each j. We consider a few different metrics for comparing model performance. One score is the Brier scores (Gneiting and Raftery, 2007, BS). The Brier score for predicting an occurrence at site $\mathbf{s}$ is given by $\{I[Y(\mathbf{s}) = 1] - \hat{P}[Y(\mathbf{s}) = 1]\}^2$ where $I[Y(\mathbf{s}) = 1]$ is an indicator function indicating that an event occurred at site $\mathbf{s}$. We average the Brier scores over all test sites, and a lower score indicates a better fit. The Brier score equally penalizes false negatives and false positives, but in the case of rare data, this may not be the best metric due to the unbalanced nature of the data. Therefore, we also consider the receiver operating characteristic (ROC) curve, and the area under the ROC curve (AUROC) for the different methods and settings. The ROC curve and AUROC are obtained via the `ROCR` (Sing et al., 2005) package in `R` (R Core Team, 2016). We then average AUCs across all datasets for each method and setting to obtain a single AUC for each combination of method and setting.

6.6 Results

Overall, we find that using a spatial probit model actually performs quite well in all cases. Table 1 gives the Brier scores and AUC for each of the methods. In Figure 2 – Figure 4, for each setting we present the vertically averaged ROC curve for each simulation method. Looking at Brier scores, we see that our model is outperformed by the probit model in all cases, by the logistic models in many settings. For AUROC methods, in a few of the settings, we do demonstrate a very small improvement over the probit and logistic models. Because these results are somewhat surprising, we also considered the performance metrics by rareness of the data. Figure 5 uses a Loess smoother from `ggplot` for the Brier scores and AUROC plotted against the rareness of the dataset for the GEV data setting with cluster samples of size $n_s = 250$. This plot shares a key feature with the same plot from many of the other settings in that as rareness increases, the AUROC for GEV is higher than the AUROC for the probit and logistic methods.

7 Data analysis

We compare our method to the spatial probit and logit for mapping the probability of the occurrence of *Tamarix ramosissima*, a plant species, for a 1-km^2 study region of PR China (Smith et al., 2012). The Chinese Academy of Forestry conducted a full census of the area, and the true occupancy of the species are plotted in Figure 6. The region is split into 10-m $\times$ 10-m grid cells, and *Tamarix ramosissima* can be found

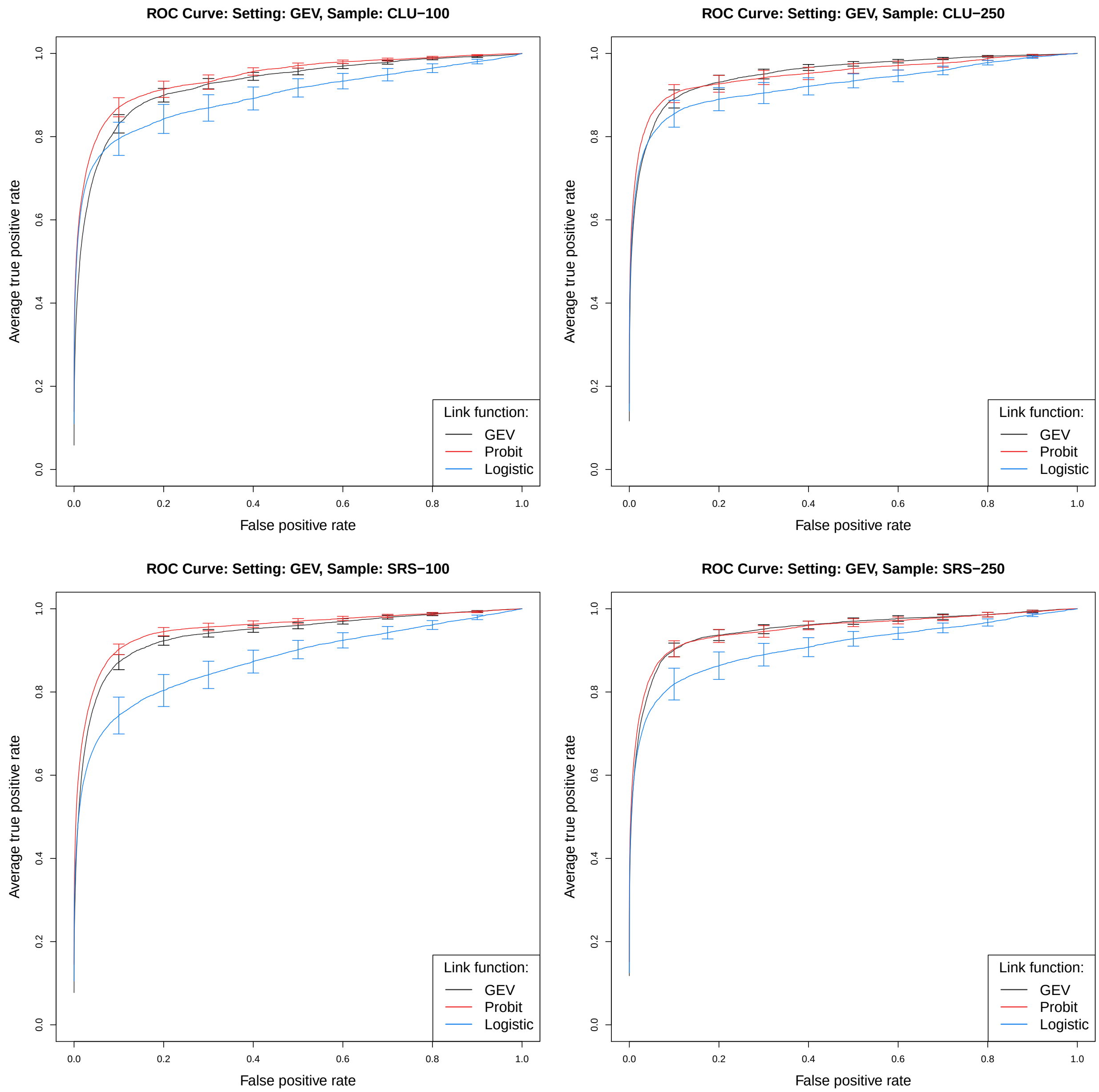

Figure 2: Vertically averaged ROC curves for GEV simulation setting.

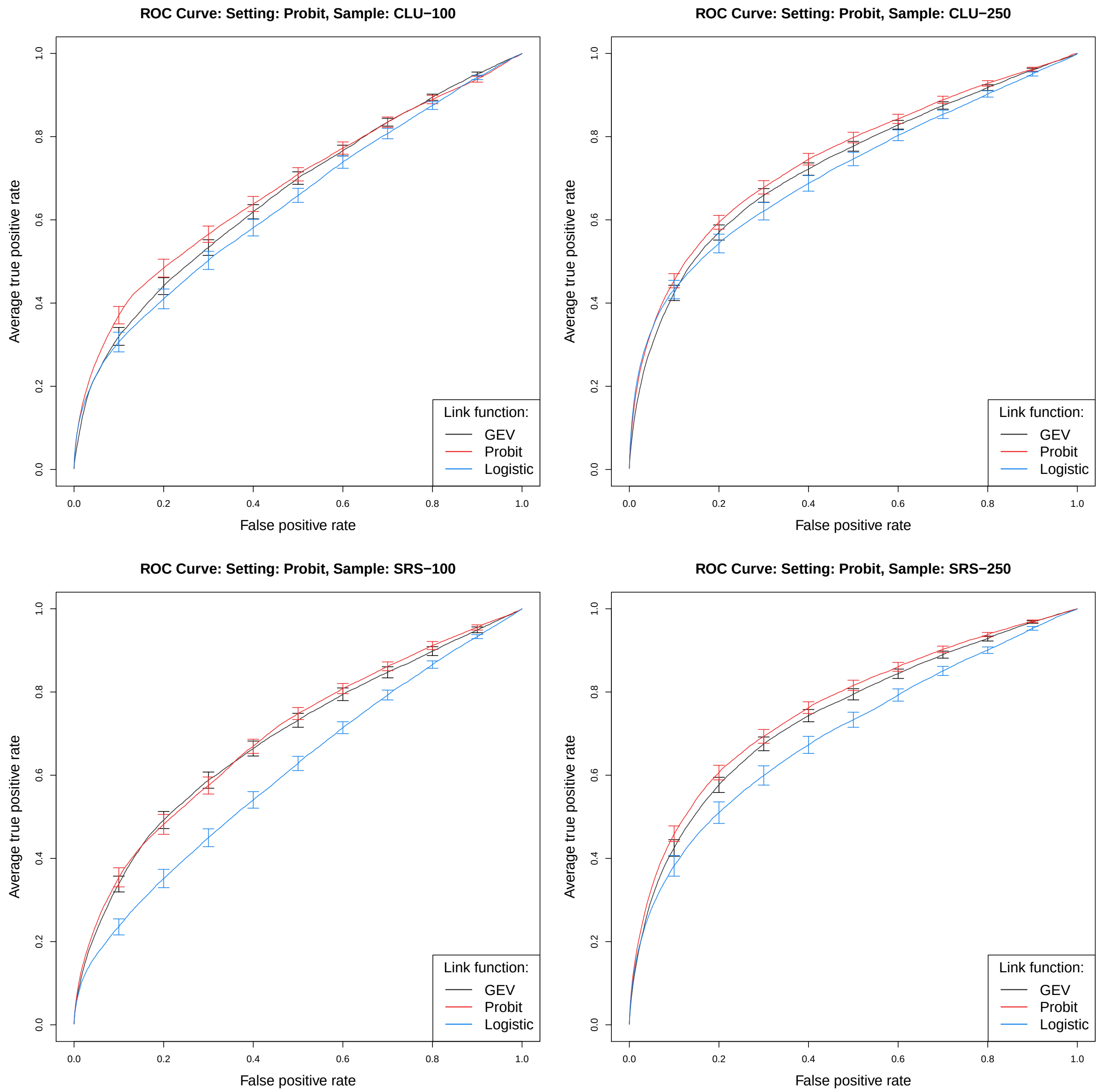

Figure 3: Vertically averaged ROC curves for probit simulation setting.

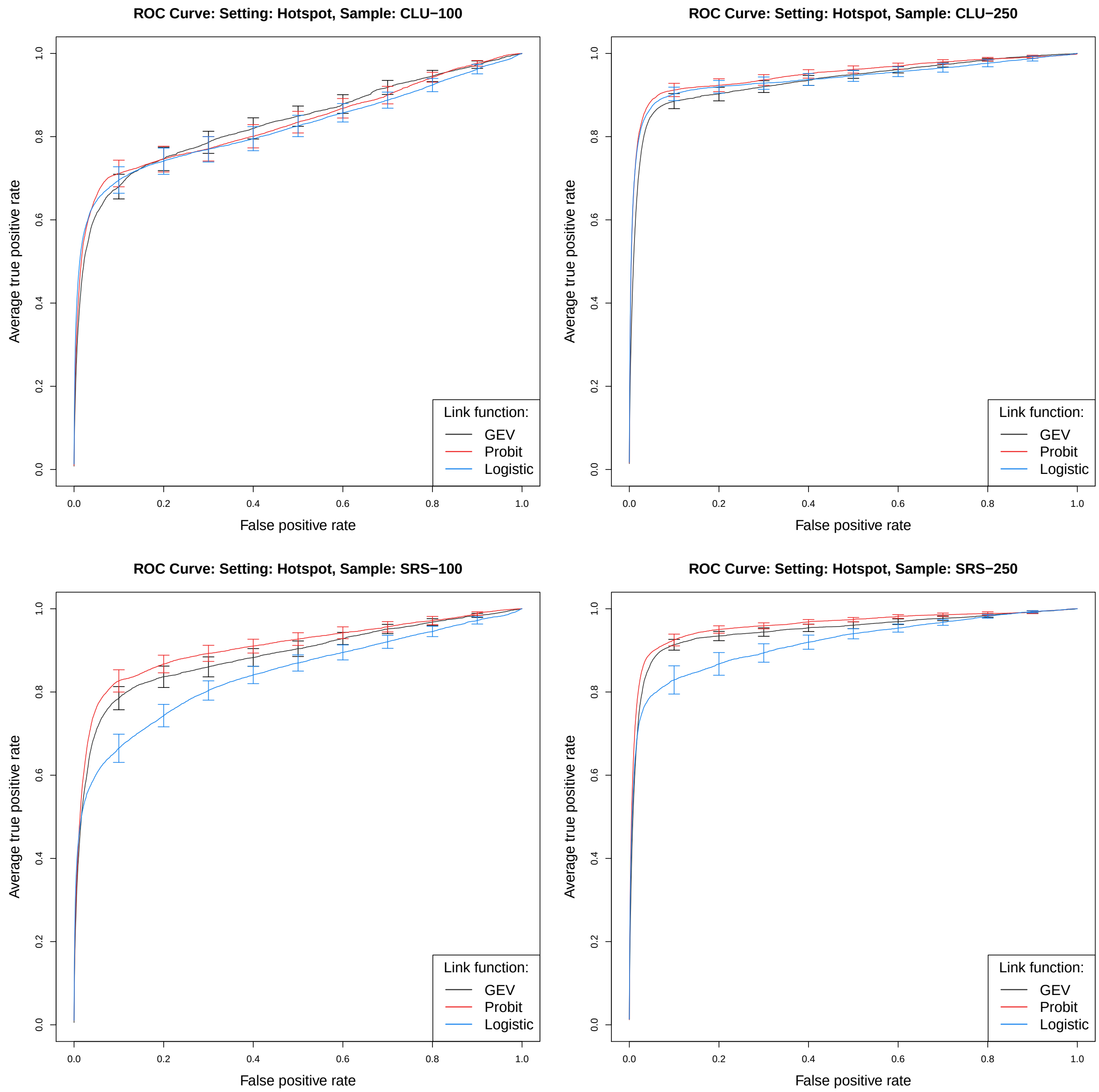

Figure 4: Vertically averaged ROC curves for hotpost simulation setting.

Table 1: Brier scores ($\times 100$) [SE] and AUROC [SE] for GEV, Probit, and Logistic methods from the simulation study.

			BS			AUROC		
Setting	n	Sample Type	GEV	Probit	Logistic	GEV	Probit	Logistic
GEV	100	CLU	3.10 [0.27]	2.45 [0.19]	2.79 [0.25]	0.926 [0.009]	0.942 [0.009]	0.900 [0.020]
		SRS	2.92 [0.20]	2.54 [0.18]	2.92 [0.25]	0.938 [0.007]	0.951 [0.007]	0.879 [0.021]
	250	CLU	2.18 [0.15]	1.87 [0.13]	2.05 [0.14]	0.951 [0.008]	0.948 [0.011]	0.922 [0.017]
		SRS	2.29 [0.15]	2.06 [0.13]	2.26 [0.15]	0.949 [0.009]	0.949 [0.010]	0.908 [0.020]
Logistic	100	CLU	5.29 [0.25]	4.94 [0.23]	5.10 [0.25]	0.659 [0.012]	0.676 [0.014]	0.643 [0.013]
		SRS	5.32 [0.23]	5.09 [0.24]	5.34 [0.26]	0.690 [0.012]	0.693 [0.012]	0.613 [0.012]
	250	CLU	4.81 [0.21]	4.55 [0.21]	4.66 [0.22]	0.731 [0.010]	0.749 [0.010]	0.714 [0.014]
		SRS	4.86 [0.22]	4.63 [0.20]	5.01 [0.23]	0.742 [0.010]	0.760 [0.010]	0.698 [0.015]
Hotspot	100	CLU	2.29 [0.17]	2.01 [0.15]	1.81 [0.12]	0.841 [0.016]	0.833 [0.019]	0.824 [0.020]
		SRS	2.09 [0.13]	1.87 [0.12]	2.13 [0.15]	0.885 [0.015]	0.906 [0.013]	0.844 [0.015]
	250	CLU	1.65 [0.11]	1.25 [0.08]	1.40 [0.09]	0.934 [0.009]	0.949 [0.008]	0.939 [0.011]
		SRS	1.53 [0.10]	1.31 [0.08]	1.63 [0.11]	0.947 [0.007]	0.960 [0.005]	0.918 [0.015]

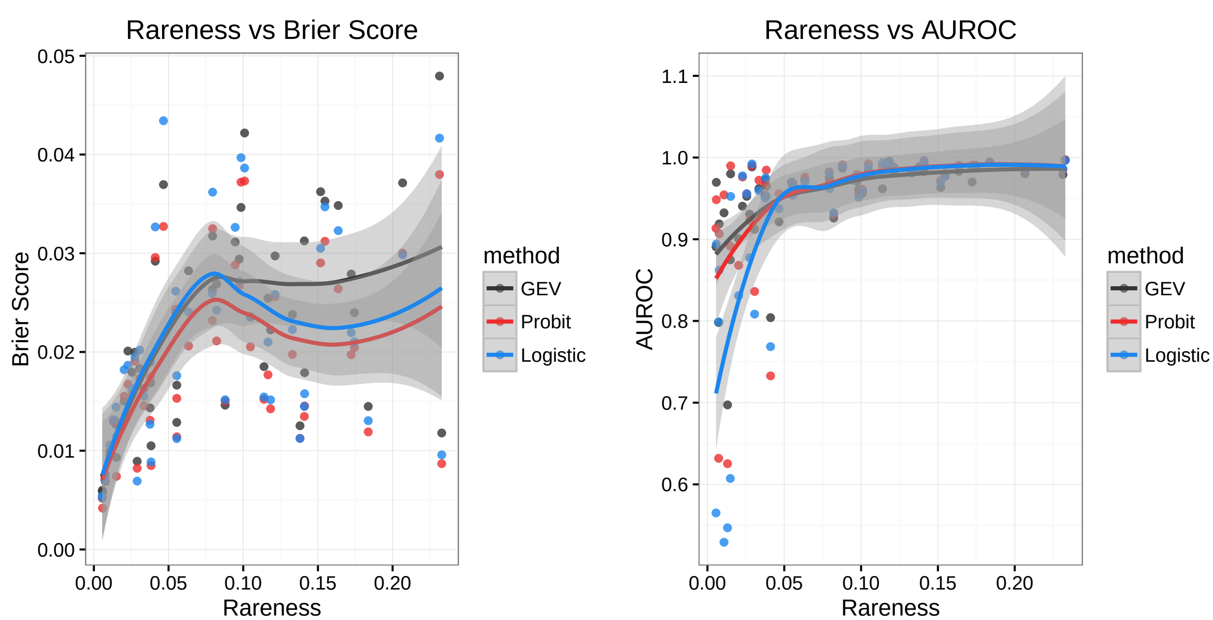

Figure 5: Smooth of BS (left) and AUROC (right) by rareness for GEV link and CLU-250 sampling.

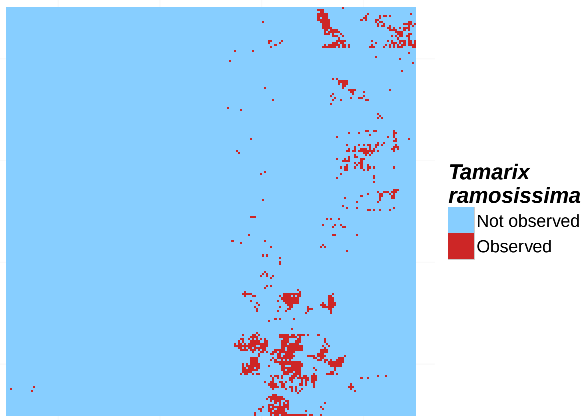

Figure 6: True occupancy of *Tamarix ramosissima* from a 1-km^2 study region of PR China.

in approximately 6% of the grid cells.

7.1 Methods

For the data analysis, we generate 100 subsamples using the CLU and SRS sampling methods with $n_s = 100, 250$ initial locations. For each subsample, we fit the spatial GEV, spatial probit, and spatial logistic models. Knot placement, prior distributions, and MCMC details for the data analysis are the same as the simulation study. To compare models, we use similar metrics as in the simulation study, but we average the metrics over subsamples.

7.2 Results

As with the simulation study, we find that in most cases the spatial probit model gives the best performance. Table 2 give summary Brier scores ($\times 100$) and AUROC for the *Tamarix ramosissima*. Figure 7 gives the vertically averaged ROC curves for each method and sampling setting.

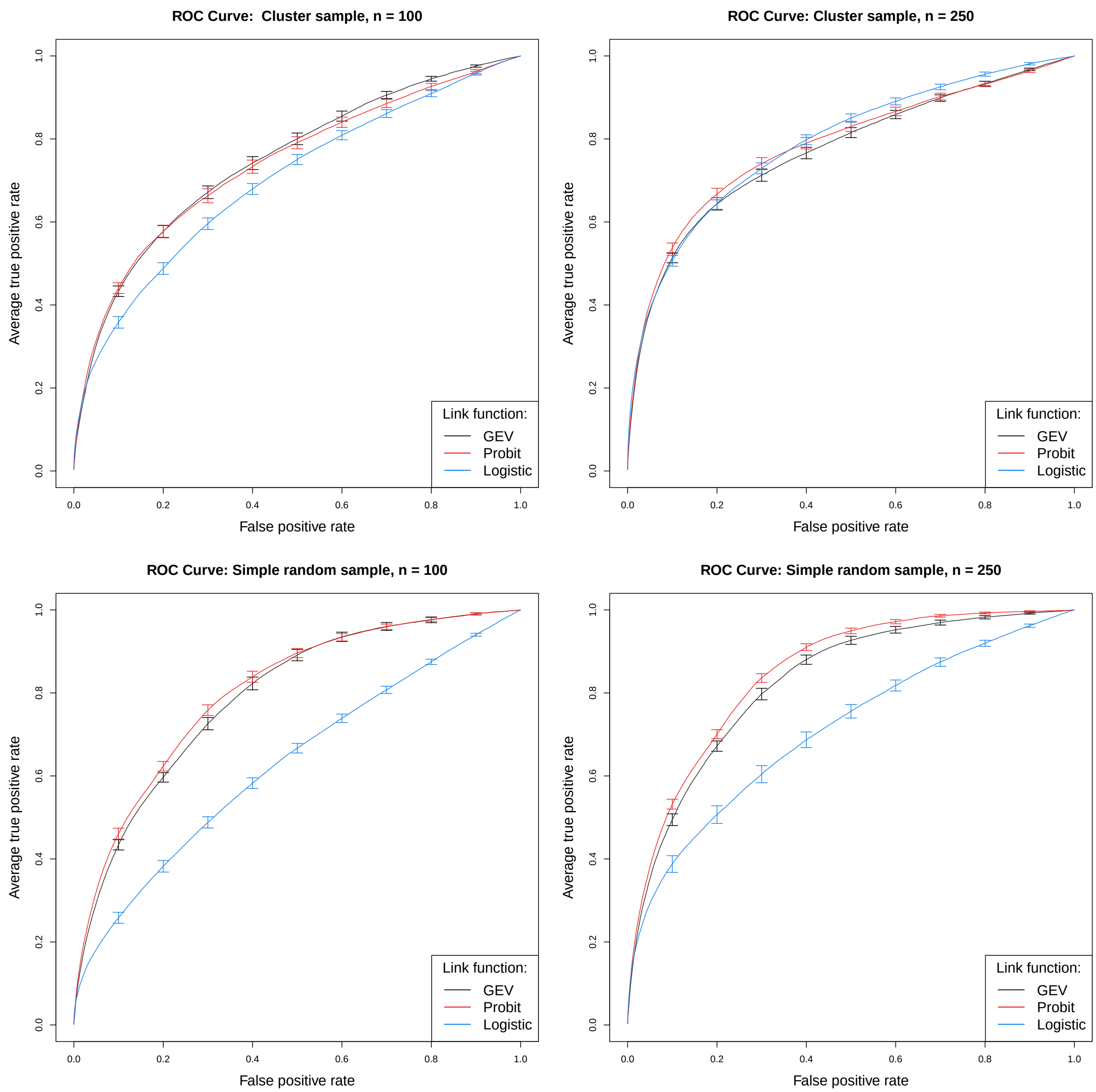

Figure 7: Vertically averaged ROC curves for *Tamarix ramosissima*.

Table 2: Brier scores ($\times 100$) [SE] and AUROC [SE] for GEV, Probit, and Logistic methods for *Tamarix ramosissima*.

		BS			AUROC		
n	Sample Type	GEV	Probit	Logistic	GEV	Probit	Logistic
100	CLU	5.15 [0.05]	5.08 [0.04]	5.35 [0.02]	0.747 [0.009]	0.742 [0.010]	0.701 [0.009]
	SRS	5.03 [0.04]	4.97 [0.04]	5.51 [0.02]	0.796 [0.006]	0.802 [0.006]	0.631 [0.008]
250	CLU	4.78 [0.03]	4.68 [0.03]	4.98 [0.04]	0.772 [0.009]	0.783 [0.009]	0.790 [0.008]
	SRS	4.83 [0.04]	4.75 [0.03]	5.17 [0.05]	0.826 [0.007]	0.848 [0.005]	0.712 [0.013]

8 Discussion and future research

In this paper, we present a max-stable spatial method for rare binary data. The principal finding in this paper is that the spatial probit model tends to outperform the proposed model. This finding is surprising given that the max-stable process is the theoretically justified spatial process for extreme value distributions, and it leads to possible research questions in the future.

It is unusual that the spatial probit model should outperform the proposed model, particularly when the data are generated directly from the proposed model. One possible explanation is that for the simulated data, there is a wide range of rarity in the data (GEV: 0.5% – 35.9%, Logistic: 1.4% – 14.4%, and Hotspot: 0.5% – 6.8%). Given that for both the GEV and logistic data settings, we have a number of datasets with a relatively high rate of occurrence, it is possible that probit is a competitive method. In particular, it may be useful to have a slightly more restrictive data generation strategy (i.e. restrict datasets to $K < 500$).

Acknowledgments

A Appendices

A.1 Binary regression using the GEV link

Here, we provide a brief review of the the GEV link of Wang and Dey (2010). Let $Y_i \in \{0, 1\}, i = 1, \ldots, n$ be a collection of i.i.d. binary responses. It is assumed that $Y_i = I(z_i > 0)$ where $I(\cdot)$ is an indicator function, $z_i = [1 - \xi \mathbf{X}_i \boldsymbol{\beta}]^{1/\xi}$ is a latent variable following a GEV$(1, 1, 1)$ distribution, $\mathbf{X}_i$ is the associated p-vector of covariates with first element equal to one for the intercept, and $\boldsymbol{\beta}$ is a p-vector of regression coefficients. Then, $Y_i \overset{ind}{\sim} \text{Bern}(\pi_i)$ where $\pi_i = 1 - \exp\left(-\frac{1}{z_i}\right)$.

A.2 Derivation of the likelihood

We use the hierarchical max-stable spatial model given by Reich and Shaby (2012). If at each margin, $Z_i \sim \text{GEV}(1,1,1)$, then $Z_i|\theta_i \overset{indep}{\sim} \text{GEV}(\theta, \alpha\theta, \alpha)$. We reorder the data such that $Y_1 = \ldots = Y_K = 1$, and $Y_{K+1} = \ldots = Y_n = 0$. Then the joint likelihood conditional on the random effect θ is

$$
\begin{aligned}
P(Y_1 = y_1, \ldots, Y_n = y_n) &= \prod_{i \le K} \left\{ 1 - \exp\left[-\left(\frac{\theta_i}{z_i}\right)^{1/\alpha} \right] \right\} \prod_{i > K} \exp\left[-\left(\frac{\theta_i}{z_i}\right)^{1/\alpha} \right] \\
&= \exp\left[-\sum_{i=K+1}^{n} \left(\frac{\theta_i}{z_i}\right)^{1/\alpha} \right] - \exp\left[-\sum_{i=K+1}^{n} \left(\frac{\theta_i}{z_i}\right)^{1/\alpha} \right] \sum_{i=1}^{K} \exp\left[-\left(\frac{\theta_i}{z_i}\right)^{1/\alpha} \right] \\
&\quad + \exp\left[-\sum_{i=K+1}^{n} \left(\frac{\theta_i}{z_i}\right)^{1/\alpha} \right] \sum_{1 < i < j \le K} \left\{ \exp\left[-\left(\frac{\theta_i}{z_i}\right)^{1/\alpha} - \left(\frac{\theta_j}{z_j}\right)^{1/\alpha} \right] \right\} \\
&\quad + \cdots + (-1)^K \exp\left[-\sum_{i=1}^{n} \left(\frac{\theta_i}{z_i}\right)^{1/\alpha} \right] \qquad (15)
\end{aligned}
$$

Finally marginalizing over the random effect, we obtain

$$
\begin{aligned}
P(Y_1 = y_1, \ldots, Y_n = y_n) &= \int G(\mathbf{z}|\mathbf{A}) p(\mathbf{A}|\alpha) d\mathbf{A}. \\
&= \int \exp\left[-\sum_{i=K+1}^{n} \left(\frac{\theta_i}{z_i}\right)^{1/\alpha} \right] - \exp\left[-\sum_{i=K+1}^{n} \left(\frac{\theta_i}{z_i}\right)^{1/\alpha} \right] \sum_{i=1}^{K} \exp\left[-\left(\frac{\theta_i}{z_i}\right)^{1/\alpha} \right] \\
&\quad + \exp\left[-\sum_{i=K+1}^{n} \left(\frac{\theta_i}{z_i}\right)^{1/\alpha} \right] \sum_{1 < i < j \le K} \left\{ \exp\left[-\left(\frac{\theta_i}{z_i}\right)^{1/\alpha} - \left(\frac{\theta_j}{z_j}\right)^{1/\alpha} \right] \right\} \\
&\quad + \cdots + (-1)^K \exp\left[-\sum_{i=1}^{n} \left(\frac{\theta_i}{z_i}\right)^{1/\alpha} \right] p(\mathbf{A}|\alpha) d\mathbf{A}. \qquad (16)
\end{aligned}
$$

Consider the first term in the summation,

$$\int \exp\left\{-\sum_{i=K+1}^{n}\left(\frac{\theta_i}{z_i}\right)^{1/\alpha}\right\} p(\mathbf{A}|\alpha)d\mathbf{A} = \int \exp\left\{-\sum_{i=K+1}^{n}\left(\frac{\left[\left(\sum_{l=1}^{L} A_l w_l(\mathbf{s}_i)^{1/\alpha}\right)^{\alpha}\right]}{z_i}\right)^{1/\alpha}\right\} p(\mathbf{A}|\alpha)d\mathbf{A}$$
$$= \int \exp\left\{-\sum_{i=K+1}^{n}\sum_{l=1}^{L} A_l \left(\frac{w_l(\mathbf{s}_i)}{z_i}\right)^{1/\alpha}\right\} p(\mathbf{A}|\alpha)d\mathbf{A}$$
$$= \exp\left\{-\sum_{l=1}^{L}\left[\sum_{i=K+1}^{n}\left(\frac{w_l(\mathbf{s}_i)}{z_i}\right)^{1/\alpha}\right]^{\alpha}\right\}. \tag{17}$$

The remaining terms in equation (16) are straightforward to obtain, and after integrating out the random effect, the joint density for $K = 0, 1, 2$ is given by

$$P(Y_1 = y_1, \ldots, Y_n = y_n) = \begin{cases} G(\mathbf{z}) & K = 0 \\ G(\mathbf{z}_{(1)}) - G(\mathbf{z}) & K = 1 \\ G(\mathbf{z}_{(12)}) - G(\mathbf{z}_{(1)}) - G(\mathbf{z}_{(2)}) + G(\mathbf{z}) & K = 2 \end{cases} \tag{18}$$

where

$$G[\mathbf{z}_{(1)}] = P[Z(\mathbf{s}_2) < z(\mathbf{s}_2), \ldots, Z(\mathbf{s}_n) < z(\mathbf{s}_n)]$$
$$G[\mathbf{z}_{(2)}] = P[Z(\mathbf{s}_1) < z(\mathbf{s}_1), Z(\mathbf{s}_3) < z(\mathbf{s}_3), \ldots, Z(\mathbf{s}_n) < z(\mathbf{s}_n)]$$
$$G[\mathbf{z}_{(12)}] = P[Z(\mathbf{s}_3) < z(\mathbf{s}_3), \ldots, Z(\mathbf{s}_n) < z(\mathbf{s}_n)].$$

Similar expressions can be derived for all K, but become cumbersome for large K.

References

Cohen, J. (1960) A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, **20**, 37–46.

Coles, S. (2001) *An Introduction to Statistical Modeling of Extreme Values*. Lecture Notes in Control and Information Sciences. London: Springer.

Finley, A. O., Banerjee, S. and Gelfand, A. E. (2015) spBayes for Large Univariate and Multivariate Point-Referenced Spatio-Temporal Data Models. *Journal of Statistical Software*, **63**.

Genton, M. G., Ma, Y. and Sang, H. (2011) On the likelihood function of Gaussian max-stable processes. *Biometrika*, **98**, 481–488.

Gneiting, T. and Raftery, A. E. (2007) Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, **102**, 359–378.

de Haan, L. (1984) A Spectral Representation for Max-stable Processes. *The Annals of Probability*, **12**, 1194–1204.

Heagerty, P. and Lele, S. (1998) A Composite Likelihood Approach to Binary Spatial Data. *Journal of the American Statistical Association*, **1459**, 1099–1111.

Huser, R. and Davison, A. C. (2014) Space-time modelling of extreme events. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **76**, 439–461.

Pacifici, K., Reich, B. J., Dorazio, R. M. and Conroy, M. J. (2016) Occupancy estimation for rare species using a spatially-adaptive sampling design. *Methods in Ecology and Evolution*, **7**, 285–293.

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.

R Core Team (2016) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL`https://www.R-project.org/`.

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.

Sing, T., Sander, O., Beerenwinkel, N. and Lengauer, T. (2005) ROCR: visualizing classifier performance in R. *Bioinformatics*, **21**, 3940–3941.

Smith, D. R., Yuancai, L., Walter, C. A. and Young, J. A. (2012) Incorporating predicted species distribution in adaptive and conventional sampling designs. In *Design and Analysis of Long-term Ecological Monitoring Studies* (eds. R. A. Gitzen, J. J. Millspaugh, A. B. Cooper and D. S. Licht), chap. 17, 381–396. Cambridge University Press.

Stephenson, A. G. (2009) High-Dimensional Parametric Modelling of Multivariate Extreme Events. *Australian & New Zealand Journal of Statistics*, **51**, 77–88.

Tawn, J. A. (1990) Modelling multivariate extreme value distributions. *Biometrika*, **77**, 245–253.

Thibaud, E. and Opitz, T. (2015) Efficient inference and simulation for elliptical Pareto processes. *Biometrika*, **102**, 855–870.

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

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