Spatial Dependency

The semivariance is a measure of spatial dissimilarity between all pairs of values generally used in geostatistics. It may be understood as opposite to correlation, which measures the degree of similarity between observations. As the correlation between observed values decreases, the semivariance increases with increasing separation distance.

The variogram is a statistic widely used for estimating the parameters involved in the specification of the covariance structure. The variogram is defined as follows:

$$C(\mathbf{u}) = \frac{1}{2} var[Y(\mathbf{x}+\mathbf{u}) - Y(\mathbf{x})] \text{ for any random process } \{Y(\mathbf{x}) \colon \mathbf{x} \epsilon R^2\}.$$

A non-parametric estimator of the variogram is the empirical variogram:

$$\hat{C}(\mathbf{u}) = \frac{1}{2|N(u)|} \sum \left\{Y(\mathbf{x}_i) - Y(\mathbf{x}_j)\right\}^2$$

Where the sum is over N(**u**)={(i,j): x_i-x_j=**u**} and |N(**u**)| is the number of distinct elements of N(**u**).

The variogram of **Figure 1** suggests that observations that are not far apart from each other (i.e. closer together) are less dissimilar (i.e. more similar) than observations that are far apart. Because the shape of the empirical variogram indicates the presence of spatial dependency in the data, it makes sense to analyze the data using techniques from spatial analysis.

Spatial Analysis

The generalized linear geostatistical model (GLGM) takes the form:

$$Y_i|U(s_i)\sim Bernoulli[p(s_i)]$$

$$logit[p(s_i)] = \mu + \beta X(s_i) + U(s_i)$$

$$cov[U(s_i), U(s_j)] = \sigma^2 \rho[(s_i - s_j), \varphi]$$

Here Y_i is the observed data at location s_i with associated covariates $X(s_i)$. The probability that $Y_i = 1|U(s_i)$ (loss to follow-up given the random process at location s_i) is denoted as $p(s_i)$. The log odds of $p(s_i)$ is modeled as a linear additive function that includes an intercept term μ, regression coefficients β associated to the covariates $X(s_i)$, and a random spatial process $U(s_i)$ at location s_i. The random Gaussian field *U* has joint multivariate normal distribution given by **U**~N(0,**Σ**) with $\mathbf{U} = [U(s_1), \dots, (s_n)]^T$ and $\Sigma_{ij} = cov[U(s_i), U(s_j)] = \sigma^2 \rho[(s_i - s_j), \varphi]$. The correlation function ρ was chosen to be exponential, which can be defined with a single scale parameter φ (usually referred as the range) that controls the rate at which the correlation decays with distance. Parameter σ^2 is the spatial variance.

The model parameters were estimated using Bayesian inference with Markov Chain Monte Carlo (MCMC) algorithms. The Bayesian framework requires specifying prior distribution for the model parameters. The priors were assigned the following distributions:

$$\mu \sim N\left(0, \sigma_\mu^2\right)$$

$$\beta \sim N\left(0, \sigma_\beta^2\right)$$

$$\sigma^2 \sim Scaled\ Inverse - \chi^2(\nu, \tau^2)$$

$$\varphi \sim exponential(\lambda)$$

The hyperparameters for the prior distributions of µ and β were chosen to be the maximum likelihood estimators (MLE). The value of the hyperparameters for the spatial parameters were the following: $\nu = 1$, $\tau^2 = 4$, and $\lambda = 4$. Finally, statistical inference was based on the posterior distribution of the model parameters and the Gaussian process.

References

Christensen, O. F., & Waagepetersen, R. (2002). Bayesian prediction of spatial count data using generalized linear mixed models. Biometrics, 58(2), 280-286.

Christensen, O.F. and Ribeiro Jr., P.J. (2002) geoRglm: A package for generalized linear spatial models. R-NEWS, Vol 2, No 2, 26-28.

Diggle, P. J., Tawn, J. A., and Moyeed, R. A. (1998). Model-based geostatistics. Journal of the Royal Statistical Society: Series C (Applied Statistics), 47(3), 299-350.

Zhang, H. (2002). On estimation and prediction for spatial generalized linear mixed models. Biometrics, 58(1), 129-136