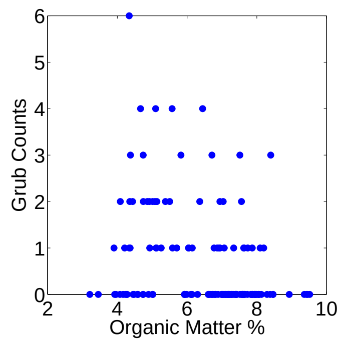
MAXIMUM LIKELIHOOD FOR SPATIALLY CORRELATED DISCRETE DATA

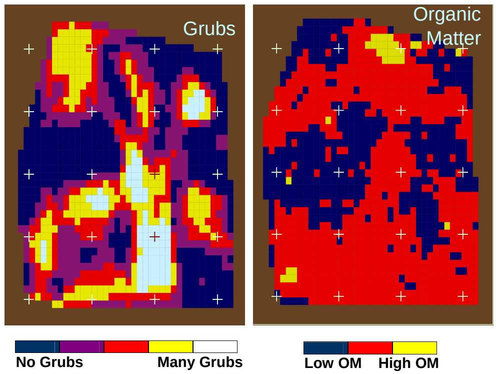
Lisa Madsen

January 29, 2007

JAPANESE BEETLE DATA



JAPANESE BEETLE DATA



JAPANESE BEETLE DATA

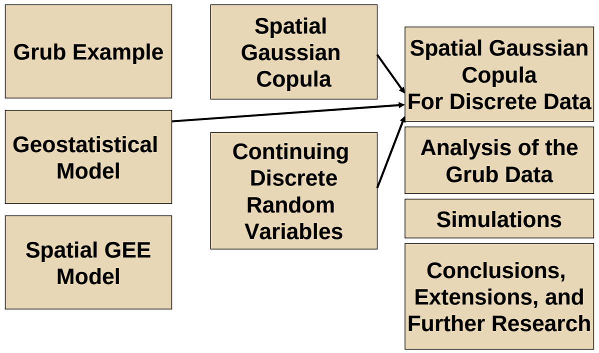
Model

The data are overdispersed counts. A sensible model is negative binomial with mean given by a function of organic matter.

$$E(Y_i) = \exp(\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3)$$

where Y_i is the ith grub count and x_i is the percent organic matter at that location.

OUTLINE



THE GEOSTATISTICAL MODEL

For Normal Data

$$\begin{aligned} \boldsymbol{Y} &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \\ \boldsymbol{\epsilon} &\sim N(\mathbf{0}, \boldsymbol{\Sigma}) \end{aligned}$$

Covariance matrix $\boldsymbol{\Sigma}$ is constructed from a *spatial covariogram*, a function depending on distance and a vector of parameters.

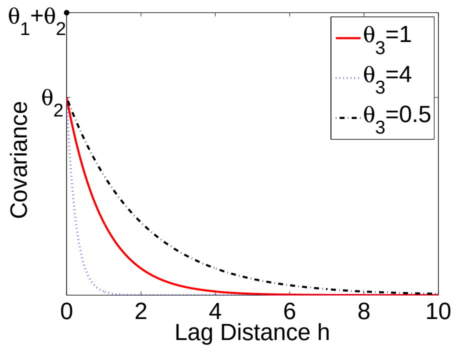
THE EXPONENTIAL COVARIOGRAM MODEL

If h_{ij} =distance between locations of Y_i and Y_j,

$$\operatorname{cov}(Y_i, Y_j) = \Sigma_{ij} = \begin{cases} \theta_1 + \theta_2, & h_{ij} = 0 \\ \theta_2 \exp(-\theta_3 h_{ij}), & h_{ij} > 0 \end{cases}$$

$$\begin{aligned} \theta_1 &= \text{nugget (measurement error)} \\ \theta_2 &= \text{partial sill} \\ \theta_3 &= \text{decay (reciprocal of range)} \end{aligned}$$

THE EXPONENTIAL COVARIOGRAM MODEL



THE GEOSTATISTICAL LIKELIHOOD

Combining the covariogram model with the normal assumption yields a likelihood

$$f(\boldsymbol{Y}|\boldsymbol{\beta},\boldsymbol{\theta}) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}(\theta)|^{1/2}} \exp\left[\frac{1}{2}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})\right]$$

from which we can find maximum likelihood estimates for the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$.

THE SPATIAL GEE MODEL

Some History

- Liang and Zeger's (1986) pioneering paper in Biometrika introduced GEEs for longitudinal data.
- Zeger (1988) developed GEE analysis for a time series of counts using a latent process model.
- McShane, Albert, and Palmatier (1997) adapted Zeger's model and analysis to spatially correlated count data.
- Gotway and Stroup (1997) used GEEs to model and predict spatially correlated binary and count data.
- Lin and Clayton (2005) develop asymptotic theory for GEE estimators of parameters in a spatial logistic regression model

THE LATENT PROCESS SPATIAL GEE MODEL

A latent process, typically lognormal, is used to model the spatial correlation. A conditionally independent discrete process, typically Poisson for counts, is assumed to model the data. Let

$$\begin{aligned}
s &= \text{spatial location} \\
\boldsymbol{x}(s) &= \text{vector of known covariates at location } s \\
\boldsymbol{\beta} &= \text{vector of unknown regression coefficients} \\
Z(s) &\sim \text{lognormal with } E[Z(s)] = 1,\ \mathrm{var}[Z(s)] = \sigma^2 \\
Y(s)|Z(\cdot) &\sim \text{independent Poisson}\{\exp[\boldsymbol{x}'(s)\boldsymbol{\beta}]\cdot Z(s)\}.
\end{aligned}$$

THE LATENT PROCESS SPATIAL GEE MODEL

Marginal Moments

The marginal moments of lognormal-Poisson $Y(s)$,

$$\begin{aligned}
E[Y(s)] &= \exp[\boldsymbol{x}'(s)\beta] \\
\mathrm{var}[Y(s)] &= E[Y(s)] + \sigma^2 E[Y(s)]^2,
\end{aligned}$$

closely resemble those of a negative binomial process: If W is distributed as negative binomial, then

$$\mathrm{var}(W) = E(W) + \frac{1}{k}E(W)^2$$

for some $k > 0$.

THE LATENT PROCESS SPATIAL GEE MODEL
Correlations

The latent process $Z(\cdot)$ carries the spatial correlation.

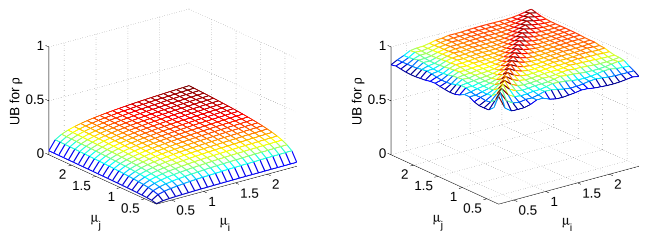
$$\mathrm{corr}[Z(s), Z(s+h)] = \rho_Z(h),$$

which induces correlation among the $Y(s)$:

$$\begin{aligned} &\mathrm{corr}[Y(s), Y(s+h)] \\ &= \rho_Z(h)\{1+\sigma^{-2}E[Y(s)]^{-1}\}\{1+\sigma^{-2}E[Y(s+h)]^{-1}\}. \end{aligned}$$

These correlations are severely limited compared to those possible between negative binomial random variables.

THE LATENT PROCESS SPATIAL GEE MODEL
Limits to Correlation



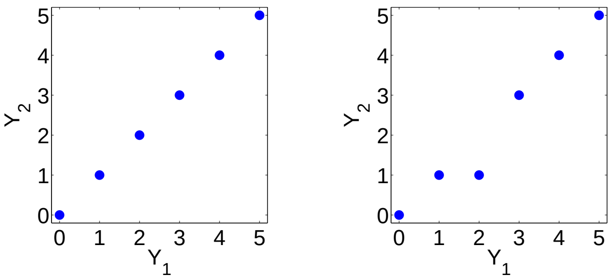
(a) Lognormal-Poisson (b) Negative binomial

THE LATENT PROCESS SPATIAL GEE MODEL

The latent process model may underestimate correlations among the data. When correlations are underestimated, standard errors are also underestimated.

BRASH ASSERTION

Correlation is not an appropriate measure of dependence for discrete random variables. In fact it's only appropriate for *normal* random variables.



(a) Perfect correlation (b) Almost perfect correlation

THE BIVARIATE GAUSSIAN COPULA

Let $Y_1 \sim F_1$ and $Y_2 \sim F_2$ be continuous random variables. The *Gaussian copula* defines a joint distribution function

$$C(y_1, y_2; \delta) = \Phi_\delta \left[\Phi^{-1}(F_1(y_1)), \Phi^{-1}(F_2(y_2))\right].$$

$$\begin{aligned} \Phi &= \text{standard normal cdf} \\ \Phi_\delta &= \text{bivariate normal cdf with correlation } \delta \end{aligned}$$

Maximum correlation between Y_1 and Y_2 is achieved by setting $\delta = 1$.

THE MULTIVARIATE GAUSSIAN COPULA

The bivariate Gaussian copula can be generalized. For $i = 1 \ldots n$, let $Y_i \sim F_i$ be continuous random variables and

$$\begin{aligned} \Phi &= \text{standard normal cdf} \\ \Phi_{\boldsymbol{\Sigma}} &= \text{multivariate Gaussian cdf with covariance matrix } \boldsymbol{\Sigma}. \\ \boldsymbol{\Sigma} &= \text{a correlation matrix} \end{aligned}$$

A joint distribution function is

$$C(y_1, \ldots, y_n; \boldsymbol{\Sigma}) = \Phi_{\boldsymbol{\Sigma}} \left[\Phi^{-1}(F_1(y_1)), \ldots \Phi^{-1}(F_n(y_n))\right].$$

THE MULTIVARIATE GAUSSIAN COPULA

Joint Density

Differentiating the distribution function yields a joint density for random variables Y_i with marginal density f_i:

$$c(\boldsymbol{y}; \boldsymbol{\Sigma}) = |\boldsymbol{\Sigma}|^{-1/2} \exp\left[-\frac{1}{2}\boldsymbol{z}'\boldsymbol{\Sigma}^{-1}\boldsymbol{z}\right] \exp\left[\frac{1}{2}\boldsymbol{z}'\boldsymbol{z}\right] \cdot \prod_{i=1}^{n} f_i(y_i)$$

where $\boldsymbol{z} = \left[\Phi^{-1}\{F_1(y_1)\}, \ldots, \Phi^{-1}\{F_n(y_n)\}\right]'$.

$\boldsymbol{\Sigma}$ determines the dependence structure.

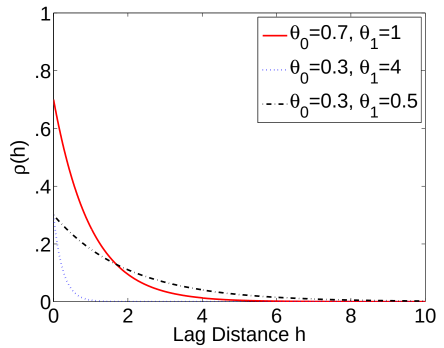
THE SPATIAL GAUSSIAN COPULA

Bring non-normal $Y_1, \ldots, Y_n$ into the geostatistical framework by modeling the Gaussian copula's $\mathbf{\Sigma}$ as a spatial correlation matrix,

$$\Sigma_{ij} = \rho(h_{ij}) = \begin{cases} \theta_0 \exp(-h_{ij}\ \theta_1), & i \neq j \\ 1, & i = j \end{cases}$$

where h_{ij} is the distance between the locations of Y_i and Y_j, and $\theta_0 \in (0,1]$ and $\theta_1 > 0$ are parameters.

A SPATIAL CORRELATION FUNCTION



RECAP

- Observations Y_i with cdf F_i and density f_i, $i = 1, \ldots, n$
- $E(Y_i)$ depends on unknown parameter vector $\boldsymbol{\beta}$ and known covariates $\boldsymbol{x}_i$
- Joint density $c(y_1, \ldots, y_n; \boldsymbol{\beta}, \boldsymbol{\theta}) =$
  $|\mathbf{\Sigma}(\boldsymbol{\theta})|^{-1/2} \exp\left[-\frac{1}{2}\boldsymbol{z}'\mathbf{\Sigma}(\boldsymbol{\theta})^{-1}\boldsymbol{z}\right] \exp\left[\frac{1}{2}\boldsymbol{z}'\boldsymbol{z}\right] \cdot \prod_{i=1}^{n} f_i(y_i)$

The joint density forms a likelihood for the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ which can be maximized to obtain MLEs.

But...how does this work for *discrete* data?

CONTINUING DISCRETE RANDOM VARIABLES

Denuit and Lambert (2005):

Associate with discrete Y_i a continuous random variable

$$Y_i^* = Y_i - U_i$$

where $U_i \sim \text{Uniform}(0,1)$ independent of Y_i and of U_j for $j \neq i$.

CONTINUING DISCRETE RANDOM VARIABLES

$Y_i^* = Y_i - U_i$ is a continuous random variable with distribution function

$$F_i^*(y) = F_i([y]) + (y - [y])Pr\{Y_i = [y+1]\}$$

and density

$$f_i^*(y) = Pr\{Y_i = [y+1]\}$$

where $[y]$ denotes the integer part of y.

CONTINUING DISCRETE RANDOM VARIABLES

A couple of observations:

- $Y_i^* = Y_i - U_i$ if and only if $Y_i = [Y_i^* + 1]$, so no information is lost by continuing Y_i.
- Distribution and density functions

$$\begin{aligned} F_i^*(y) &= F_i([y]) + (y - [y])Pr\{Y_i = [y+1]\} \\ f_i^*(y) &= Pr\{Y_i = [y+1]\} \end{aligned}$$

depend on only the parameters of the distribution of Y_i.

THE SPATIAL GAUSSIAN COPULA FOR DISCRETE DATA

The spatial Gaussian copula joint density for $Y_1^*, \ldots, Y_n^*$,

$$c(\boldsymbol{y}; \boldsymbol{\beta}, \boldsymbol{\theta}) =$$

$$|\boldsymbol{\Sigma}(\boldsymbol{\theta})|^{-1/2} \exp\left[-\frac{1}{2}\boldsymbol{y}'\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\boldsymbol{y}\right] \exp\left[\frac{1}{2}\boldsymbol{y}'\boldsymbol{y}\right] \cdot \prod_{i=1}^{n} f_i^*(y_i),$$

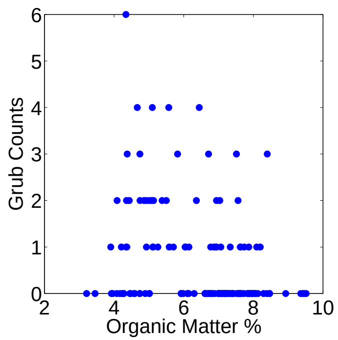
gives a log-likelihood $L(\boldsymbol{\beta}, \boldsymbol{\theta}; \boldsymbol{Y}, \boldsymbol{U}) = \log[c(\boldsymbol{Y}^*; \boldsymbol{\beta}, \boldsymbol{\theta})]$.

THE SPATIAL GAUSSIAN COPULA FOR DISCRETE DATA

Since $L(\boldsymbol{\beta}, \boldsymbol{\theta}; \boldsymbol{Y}, \boldsymbol{U})$ depends on $\boldsymbol{U}$, MLEs will be

$$(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}) = E_{\boldsymbol{U}} \left\{ \underset{\boldsymbol{\beta},\boldsymbol{\theta}}{\arg\max} \, [L(\boldsymbol{\beta}, \boldsymbol{\theta}; \boldsymbol{Y}, \boldsymbol{U})] \right\}.$$

ANALYSIS OF THE GRUB DATA



ANALYSIS OF THE GRUB DATA

Model

$$\begin{aligned}
Y_i &\sim \text{Negative Binomial, } i = 1 \ldots 143 \\
E(Y_i|x_i) &= \mu_i = \exp(\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3) \\
\mathrm{var}(Y_i) &= \mu_i \left(\frac{1+\phi}{\phi} \right) \\
\mathrm{corr}(Y_i, Y_j) &= \begin{cases} \theta_0 \exp(-h_{ij}\, \theta_1), & i \neq j \\ 1, & i = j \end{cases}
\end{aligned}$$

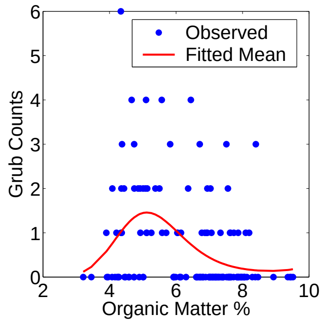
ANALYSIS OF THE GRUB DATA

Method

1. Generate $U_1 \ldots U_n$ ~iid U$(0,1)$ and form $Y_i^* = Y_i - U_i$.
2. Find $(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\theta}}) = \arg\max_{\boldsymbol{\beta},\boldsymbol{\theta}} [L(\boldsymbol{\beta}, \boldsymbol{\theta}; \boldsymbol{Y}, \boldsymbol{U})]$ and approximation of negative Hessian of L at maximum.
3. Repeat steps 1 and 2 several times.
4. $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\theta}}$ are averages of the $(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\theta}})$.
5. Standard errors are square roots of the diagonal elements of the average approximated Hessian.

ANALYSIS OF THE GRUB DATA

Fitted Mean Function



ANALYSIS OF THE GRUB DATA

Parameter Estimates

Parameter	Estimate	Standard Error	Nominal 95% Confidence Interval
β_0	-25.2514	10.7285	$(-46.28, -4.22)$
β_1	12.3951	5.2811	$(2.04, 22.75)$
β_2	-1.9097	0.8452	$(-3.57, -0.253)$
β_3	0.0911	0.0441	$(0.005, 0.1776)$

SIMULATIONS

- $n = 143$ with spatial locations from grub data
- $\mu_i = \exp(\beta_0)$, where $\beta_0 = 1$
- Data generated using software package discsim2.1 (www.stat.oregonstate.edu/people/lmadsen)
- About 10% of the pairs (Y_i, Y_j) had correlations exceeding the lognormal-Poisson upper bound.
- MLE and GEE estimates of β_0 were calculated.

SIMULATIONS

Results

Procedure	Bias	Variance	Nominal 95% Confidence Coverage
Spatial GEE	-0.01	0.01	0.69
MLE	-0.01	0.01	0.91

CONCLUSIONS

- Latent variable spatial GEE model can dangerously underestimate variance.
- Spatial Gaussian copula makes it easy to model spatial dependence for non-normal data.
- ML method is easier to work with than GEE method.

GENERALIZATIONS TO THE MODEL

- The method can be used for any non-normal marginals and any correlation structure.
- It is not necessary that all Y_i share the same marginal distribution. For example, data could be overdispersed in some regions and underdispersed in others.
- For the negative binomial marginal model, ϕ could be allowed to vary.

FURTHER RESEARCH

- More simulations to assess performance in a variety of situations.
- More applications.
- Asymptotic details.
- Generating highly correlated discrete data.

ACKNOWLEDGEMENTS

The research presented here has been partially funded by the U.S. Environmental Protection Agency Grant #CR-829095, the Science To Achieve Results (STAR) Program. It has not been subjected to the Agency's review and therefore does not necessarily reflect the views of the Agency, and no official endorsement should be inferred.

Thanks to Clif Johnson for his extensive help figuring out how to run the simulations on the College of Engineering's Beowulf cluster.

THANK YOU!