

A COMPARISON OF  
ESTIMATION METHODS  
FOR SPATIAL DATA  
ANALYSIS  
WITH DISCRETE DATA ON  
A LATTICE

Monica C. Jackson  
University of Maryland  
Department of Mathematics

# Introduction

- Spatial data analysis involves data that represents points (e.g. soil core samples) or regions (e.g. counties)
- Spatial models on a lattice are analogous to time-series models in the sense that, when building models for data on a lattice, there is not a realization occurring between locations or regions.
- In spatial data, we do not have the unidirectional flow of time that occurs with time series.

# Problems

- There are certain practical problems that are associated with the models we propose to study. These problems include:
  1. handling of edge data
  2. large numbers of zeros
  3. outliers

# Research Objective

- Simulate Spatial data
- Determine if method of generating data affect method of estimating
- How well are trends detected
- Problem data sets

## **Motivation:**

A need for a more comprehensive comparison

# Review of Previous Work

- Noel Cressie: Gaussian models
- Ferrandiz et. al: auto-Poisson models
- Kaiser and Cressie: Winsorized auto-Poisson model
- Breslow and Clayton: GLMM models

# Auto Models

- Gaussian

$$f(z_i | z_j : j \neq i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{(z_i - \eta_i)^2}{2\sigma_i^2}}, \text{ for } i = 1, \dots, n \quad (1)$$

where,

$z_i$  represents the response observed at lattice location  $i$ .

$z_j$  represents the response observed at lattice locations  $j$

$j$  is predefined neighbors of  $i$

$\sigma_i$  represents the conditional variance.

$$\eta_i = \mu_i + \sum_{j \neq i}^n \theta_{ij}(z_j - \mu_j), \text{ for } i = 1, \dots, n \quad (2)$$

where,  $\theta_{ij}$  models the spatial dependence.

$\theta_{ij} = 0$  if location  $j$  is not a neighbor of location  $i$ . We also define  $\theta_{ii} = 0$ .

- Poisson

$$P(z_i|z_j) = \exp(-\eta_i) \eta_i^{z_i} / z_i! \quad (3)$$

where

$$\eta_i = \exp\left(\mu_i + \sum_{j=1}^n \theta_{ij} z_j\right). \quad (4)$$

$\Theta$  is the  $n \times n$  spatial dependence matrix with elements  $\theta_{ij}$  where  $i = 1 \dots n$  and  $j = 1 \dots n$ .

$\theta_{ij} = \theta_{ji}$ ,  $\theta_{ii} = 0$ , and  $\theta_{ij} = 0$  if location  $j$  is not a predefined neighbor of location  $i$ .

$\mu_i$  is the elements of the  $n \times 1$  vector,  $\mu$ . Where,  $\mu = \mathbf{X}\beta$ , and  $\mathbf{X}$  is an  $n \times k$  covariate matrix with  $k$  covariates and slope,  $\beta$ .

Ferrandiz et. al. (1995) use this model to predict trends in cancer mortality in Valencia, Spain.

- Winsorized auto-Poisson model (Kaiser)

A possible Winsorization could be defined as

$$Y = Z I_{(Z \leq R)} + R I_{(Z > R)} \quad (5)$$

where  $R$  is a fixed integer value on  $(0, \infty)$  and  $I_{(Z \leq R)}$  represents the indicator function which takes on value 1 if it's condition (e.g.  $Z \leq R$ ) is met and 0 otherwise.

Generally, it is sufficient to choose  $R$  such that  $R \geq 3 \times \eta_m$  where  $\eta_m = \max\{\eta_1, \dots, \eta_n\}$ .

- auto-binomial

The auto-Poisson model is an approximation to the auto-binomial when  $n_i$  (the number of individuals in a region) is large and the parameter  $p_i$  is small. The conditional probability mass function (pmf) of this model is defined as

$$P(z_i | z_j : j \neq i) = \binom{n_i}{z_i} p_i^{z_i} [1 - p_i]^{n_i - z_i} \quad (6)$$

where

$$p_i = \frac{\exp(\mu_i + \sum_{j=1}^n \theta_{ij} z_j)}{1 + \exp(\mu_i + \sum_{j=1}^n \theta_{ij} z_j)}. \quad (7)$$

$n_i$  is defined as being the size of the population at location  $i$  (e.g. the number of people who live a given county).  $\theta_{ij}$  and  $\mu_i$  are as previously defined.

# Generalized Linear Mixed Models (GLMM)

- Generalized linear models (McCullagh and Nelder) extend the usual linear models to cases of Poisson, Binomial, and other exponential family distributions by means of a link function and well established approaches to estimation and hypothesis testing.

Namely, we can define  $\eta_i$  as the mean of the *i*th point on the lattice and

$$\log \eta_i = \mu_i + \varepsilon_i \quad (8)$$

where  $\mu_i$  may be defined in terms of a limited number of covariates and  $\varepsilon_i$  is normally distributed error term. We can define  $\Sigma = (\varepsilon_1, \dots, \varepsilon_n)$  as  $N(0, \Sigma)$  where  $\Sigma$  has a spatial covariance structure.

# Research Process

- Simulate: auto-Poisson and auto-binomial
- Validate: Hypothesis testing
- Estimate: GLM with pseudolikelihood, Transforming to approximately Gaussian, GLMM
- Contaminate: Zeros, edges, outliers

# Simulating Spatial Data

- Using the Auto-Poisson and Auto-binomial

The algorithm that we are using is similar to the Gibbs sampling algorithm (Gilks). To obtain a sample from a joint distribution  $p(z_1, z_2, \dots, z_d)$  the algorithm is implemented in the following iterative manner, where  $z_i^d$  refers to random variable  $z_i$  at iteration  $d$  where  $d = 1 \dots D$  and we define  $j$  as the neighbors of  $i$ :

Step 1: Sample  $z_1^{d+1}$  from  $p(z_1|z_2^d, \dots, z_n^d)$

Step 2: Sample  $z_2^{d+1}$  from  $p(z_2|z_1^{d+1}, z_3^d, \dots, z_n^d)$

⋮

Step  $g$ : Sample  $z_n^{d+1}$  from  $p(z_n|z_1^{d+1}, \dots, z_{n-1}^{d+1})$ .  
The process is completed after a predetermined number of steps  $g$  have been completed for  $D$  iterations.

More specifically, for the auto-Poisson:

Step 1: Sample  $z_1^{d+1}$  from

$$P(z_1|z_j : j \neq 1) = \exp(-\eta_1)\eta_1^{z_1}/z_1!$$

where

$$\eta_1 = \exp(\mu_1 + \sum_{j=1}^n \theta_{1j}z_j)$$

Step 2: Sample  $z_2^{d+1}$  from

$$P(z_2|z_j : j \neq 2) = \exp(-\eta_2)\eta_2^{z_2}/z_2!$$

where

$$\eta_2 = \exp(\mu_2 + \sum_{j=1}^n \theta_{2j}z_j)$$

: Step g: Sample  $z_n^{d+1}$  from

$$P(z_n|z_j : j \neq n) \exp(-\eta_n)\eta_n^{z_n}/z_n!$$

where

$$\eta_n = \exp(\mu_n + \sum_{j=1}^n \theta_{nj}z_j)$$

The proximity index between the  $i$ th and  $j$ th locations is.

$$a_{ij} = \frac{\sqrt{N_i N_j}}{d_{ij}}, \quad (9)$$

where  $N_i$  and  $N_j$  represent subsets of the total population at location  $i$  and  $j$  respectively (e.g. the number of people in Bangladesh in village  $i$  and  $j$ ).

$d_{ij}$  represents the distance between location  $i$  and  $j$ .

location  $i$  as a neighbor of location  $j$  if  $a_{ij} > a$  where  $a$  is threshold.

We will place one covariate in the model corresponding to column coordinate or "longitude". This covariate will cause an east-west trend in the data.

The conditional mean for the  $i$ th region given the realization of the neighbors is

$$\log \eta_i = \alpha + \log N_i + \sum_{s=1}^k \beta_s x_i + \sum_{j=1}^n \theta_{ij} z_j. \quad (10)$$

where  $\alpha$  is the intercept,  $x_i$  is column coordinate of region  $i$ ,  $N_i =$  size (e.g. population at risk) of region  $i$ , and  $z_j$  is the count at location  $j$  (recall location  $j$  is a neighbor of location  $i$ ).

Large values of  $\beta$  will force a large trend effect (means get larger as you go from west to east) while values of  $\beta = 0$  imply no trend.

We define  $\theta_{ij}$  as follows

$$\theta_{ij} = \gamma a_{ij} \quad (11)$$

where,  $\gamma$  is an interaction parameter that allows for the spatial dependence to be proportional to the proximity index  $a_{ij}$ . This mean will be used in auto-Poisson and auto-binomial model. We truncate our random variables in the auto-Poisson model as described earlier.

# Convergence and Initialization

- choice of starting values for the algorithm is not an extreme concern but may impact number of iterations in burn-in period.
- It is well established that (Gilks et. al.) any Monte Carlo Markov Chain (MCMC) sampler will “forget” its starting values once the algorithm has run enough, also known as a “burn-in” period.
- We run the algorithm for different burn-in periods.
- it is possible to obtain convergence much faster with initial values related more closely to the data set we are trying to generate.
- the algorithm described is repeated for each realization needed.

## Using GLMM

- This method is known as the “simultaneous” method while the previous method was the “conditional” method.
- The difference being that the correlation structure is defined in the error term for the simultaneous models and defined in the mean for the conditional models.

First consider the GLMM for the Poisson distribution. Recall from equation (8) that we defined the mean of a location  $i$  as

$$\log \eta_i = \mu_i + \varepsilon_i. \quad (12)$$

We now will define  $\mu_i$  as the first part of equation 10 as

$$\mu_i = \alpha + \log N_i + \sum_{s=1}^k \beta_s x_i. \quad (13)$$

It is now through the  $n \times 1$  vector,  $\varepsilon$  with elements  $\varepsilon_i$ , that we define the spatial dependence structure of the data.

Define

$$\boldsymbol{\varepsilon} \sim \text{Gau}(\mathbf{0}, \boldsymbol{\Sigma}). \quad (14)$$

Now define a correlation structure that is similar to the spatial dependence structure (the  $\theta_{ij}z_j$  term ) defined in the latter part of equation (10). First we define the  $n \times n$  covariance matrix,  $\boldsymbol{\Sigma}$ , as

$$\boldsymbol{\Sigma} = \left[ (\mathbf{I} - \gamma \mathbf{A})^T \mathbf{D}^{-1} (\mathbf{I} - \gamma \mathbf{A}) \right]^{-1} \sigma^2 \quad (15)$$

where  $\gamma$  and  $\sigma$  are scalar parameters,

$\mathbf{D}$  is a diagonal  $n \times n$  matrix with elements  $1/N_i$  and  $\mathbf{A}$  is an  $n \times n$  matrix with elements  $a_{ij}$  as defined in equation (9).

Now, the elements of the matrix  $\gamma \mathbf{A}$  in equation (15) is equivalent to  $\theta_{ij}$  in equation (11).

To generate multivariate data with covariance matrix  $\Sigma$  we use the standard method of generating multivariate Gaussian random variables. The Cholesky decomposition of  $\Sigma$  is defined as

$$\Sigma = \mathbf{L}\mathbf{L}^T \quad (16)$$

where  $\mathbf{L}$  is a lower triangular matrix.  $\epsilon$  can now be generated from  $\mathbf{L}$  by

$$\epsilon = \mathbf{L}\xi \quad (17)$$

where

$$\xi_1 \dots \xi_n \sim \text{Gau}(0, 1) \text{ i.i.d.}$$

Now we have that,

$$E[\epsilon] = E[\mathbf{L}\xi] = \mathbf{L}E[\xi] = \mathbf{0}$$

and

$$\text{Var}[\epsilon] = \text{Var}[\mathbf{L}\xi] = \mathbf{L}\text{Var}[\xi]\mathbf{L}^T = \Sigma$$

This gives us our desired distribution.

- The computing cost of using this algorithm is on the order of  $(n \times n)^3$
- We did consider other algorithms such as the Circulant Embedding Method (as described by Kozintsev ) which defines  $\sigma$  by a Fourier transform matrix broken into real and imaginary parts.

# The Algorithm

Step 1: Generate an  $n \times 1$  vector,  $\varepsilon$ , where  $\varepsilon \sim \text{Gau}(0, \Sigma)$ .

Step 2: Add the values obtained in step 1 to the mean vector defined by equation (13).

Step 3: Generate Poisson variables using  $\log(\mu_i)$  from step 2.

- For the logistic link function

Step 1: Generate an  $n \times 1$  vector,  $\varepsilon$ , where  $\varepsilon \sim \text{Gau}(0, \Sigma)$ .

Step 2: Add the values obtained in step 1 to the mean vector defined by equation (13) where our link function is

$$\log \frac{p_i}{1 - p_i}.$$

Step 3: Generate Binomial random variables using  $n_i$  and  $p_i$  derived in step 2.

# Criteria for Goodness of Fit

- We define  $\mathbf{Z}^r$  (where  $r = 1 \dots m$ ) to represent each independent realization generated.
- Follow the methods of Ferrandiz et. al. to develop a goodness of fit test.

With each  $\mathbf{Z}$  we can calculate

$$\chi_r^2 = \sum_{i=1}^n \frac{(z_i - \hat{\mu}_i)^2}{\hat{\mu}_i}. \quad (18)$$

and generate a distribution:

Step 1: Generate a realization of the auto-Poisson data which we will define as  $\mathbf{Z}^0$ .

Step 2: Fit the model.

Step 3: Generate  $r$  additional replications using the fitted parameters ( $\mathbf{Z}^1, \dots, \mathbf{Z}^r$ ).

Step 4: Calculate the  $\chi^2$  values for each dataset (which we now call  $\chi^{2[0]}, \dots, \chi^{2[r]}$ ) using equation (18) by estimating  $\hat{\mu}$  in equation (18) by

$$\hat{\mu}_i = \frac{1}{r + 1} \sum_{w=0}^r z_i^w. \quad (19)$$

.

Step 5: Reject the model (distribution is not auto-Poisson) at a significance level,  $\alpha$  when  $\chi^2[0]$  is greater than  $(1-\alpha)$  % of the chi-square values from the simulated data  $(\chi^2[1], \dots, \chi^2[r])$ .

Step 6: repeat steps 1-5 to obtain  $m$  independent realizations.

More precisely, for our problem:

Step 1: we generated one data set with the following parameters:  $\gamma = .1$ ,  $\beta = .01$  (see equation 11 and 10 ),  $a = .02$  (our threshold).

Step 2: Once the data was generated the data was fit. We obtained the following parameter estimates:  $\hat{\gamma} = .12$  and  $\hat{\beta} = .0089$ .

Step 3: We generated 100 additional replications with the parameter estimates obtained in step 2.

Step 4: We calculated  $\chi^2[0], \dots, \chi^2[100]$

Step 5: We rejected the model if  $\chi^2[0]$  was greater than 95 % of the chi-square values from the simulated data ( $\chi^2[1], \dots, \chi^2[100]$ ).

Step 6: We repeated steps 1-5 100 times to obtain 100 independent realizations.

## Replications

We solve the following equation for  $m$  to obtain the total number of replications  $m$  needed.  $\Pi$  represents coverage probability, with

$$\text{Standard error} = \sqrt{\frac{\Pi(1 - \Pi)}{m}} < .005. \quad (20)$$

For our simulations we use 1900 replications.

# Analyzing Correlated Discrete Data

The three methods of estimation we used are

- GLM with pseudo-likelihood
- using MLE with data transformed to Gaussian
- and using GLMM.

## Using GLM with pseudo-likelihood

Equation (10) be written as a GLM if

$$q_i = \sum_{j=1}^n a_{ij} z_j \quad (21)$$

then define independent Poisson random variables as,

$$z_i \sim \text{Poisson}(\eta_i) \quad (22)$$

where,

$$\log \eta_i = \alpha + \log N_i + \sum_{s=1}^k \beta_s x_i + \gamma q_i. \quad (23)$$

Where  $\alpha$  is an intercept. We now have an approximation to equation (10). Ferrandiz used this method and estimated all parameters ( $\gamma$  and  $\beta$ )

To verify simulations we simulated 1900 data sets under the null hypothesis shown in tables and .

Method of simulation	Using auto-Poisson	Using GLMM
True Null % significant for $\beta$	$\beta = 0$ $\gamma = .01$ $\beta = 0$ 2%	$\beta = 0$ $\gamma = .01$ $\beta = 0$ 5%
True Null % significant for $\beta$	$\beta = 0$ $\gamma = .05$ $\beta = 0$ 5%	$\beta = 0$ $\gamma = .05$ $\beta = 0$ 3%
True Null % significant for $\beta$	$\beta = .01$ $\gamma = 0$ $\gamma = 0$ 2%	$\beta = .01$ $\gamma = 0$ $\gamma = 0$ 4%
True Null % significant for $\gamma$	$\beta = .05$ $\gamma = 0$ $\gamma = 0$ 4%	$\beta = .05$ $\gamma = 0$ $\gamma = 0$ 4%

Method of simulation	Using auto-binomial	Using GLMM
True Null % significant for $\beta$	$\beta = 0$ $\gamma = .01$ $\beta = 0$ 5%	$\beta = 0$ $\gamma = .01$ $\beta = 0$ 4%
True Null % significant for $\beta$	$\beta = 0$ $\gamma = .05$ $\beta = 0$ 4%	$\beta = 0$ $\gamma = .05$ $\beta = 0$ 4%
True Null % significant for $\beta$	$\beta = .01$ $\gamma = 0$ $\gamma = 0$ 3%	$\beta = .01$ $\gamma = 0$ $\gamma = 0$ 5%
True Null % significant for $\gamma$	$\beta = .05$ $\gamma = 0$ $\gamma = 0$ 5%	$\beta = .05$ $\gamma = 0$ $\gamma = 0$ 4%

# Using MLE with data transformed to Gaussian

Freeman-Tukey square-root transformation

$$F_i = \sqrt{\frac{1000z_i}{N_i}} + \sqrt{\frac{1000(z_i + 1)}{N_i}}, \quad (24)$$

- Compute the maximum likelihood estimates

The log-likelihood,  $L$ , function from the conditional Gaussian distribution

$$L = -\frac{n}{2} \log 2\pi + \quad (25)$$

$$\frac{1}{2} \log |(\mathbf{I} - \Theta)^{-1} \Sigma| + \quad (26)$$

$$\frac{1}{2} (\mathbf{Z} - \boldsymbol{\mu})' \Sigma^{-1} (\mathbf{I} - \Theta) (\mathbf{Z} - \boldsymbol{\mu}) \quad (27)$$

$$(28)$$

where  $\mathbf{Z}$ ,  $\boldsymbol{\mu}$ ,  $\mathbf{I}$ ,  $\Theta$  and  $\mathbf{I} - \Theta$  are as previously defined.

We multiply equation (25) by  $-1$  to get the negative log-likelihood.

Thus we will now minimize equation (25) with respect to  $\boldsymbol{\mu}$ ,  $\boldsymbol{\Theta}$ , and  $\boldsymbol{\Sigma}$ .

In place of the counts,  $z_i$ , we use the Freeman-Tukey transformed rates,  $\mathbf{F}$  ( $\mathbf{F}$  is an  $n \times 1$  vector with elements  $F_i$ ), per 1000 people as shown in equation (24). Now, equation (25) is reparameterized and written as

$$L = \frac{n}{2} \log 2\pi + \frac{n}{2} \log \tau^2 - \frac{1}{2} \log |\mathbf{D}^{-1}(\mathbf{I} - \gamma\mathbf{A})| + \frac{1}{2}(\mathbf{F} - \mathbf{XB})'\mathbf{D}^{-1}(\mathbf{I} - \gamma\mathbf{A})(\mathbf{F} - \mathbf{XB}) \quad (26)$$

where from equation 25 we have  $\boldsymbol{\mu} = \mathbf{XB}$ ,

$$\Sigma = \tau^2 \mathbf{D},$$

$$\mathbf{D} = \begin{pmatrix} N_1 & 0 & 0 & \dots & 0 \\ 0 & N_2 & 0 & \dots & 0 \\ 0 & 0 & \dots & \dots & \vdots \\ \vdots & \vdots & \dots & \dots & \vdots \\ 0 & \dots & \dots & 0 & N_n \end{pmatrix}, \quad (30)$$

$\Theta = \gamma \mathbf{A}$ . Also  $\mathbf{X}$ ,  $\mathbf{A}$ ,  $\beta$  and  $N_i$  (for  $i = 1 \dots n$ ) are as defined earlier.

To obtain parameter estimates for  $\gamma, \beta$ , and  $\tau$  it is necessary to use graphical and computational methods as shown by Cressie and Chan .

First, we assume  $\gamma$  is fixed to a value.

Next the maximum likelihood estimate of  $\beta$  when  $\gamma$  is fixed becomes

$$\hat{\beta} = (\mathbf{X}'\mathbf{D}^{-1}(\mathbf{I} - \gamma\mathbf{A})\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}^{-1}(\mathbf{I} - \gamma\mathbf{A})\mathbf{F}. \quad (31)$$

and the maximum likelihood estimate of  $\tau^2$  when  $\gamma$  is fixed becomes

$$\begin{aligned} \hat{\tau}^2 &= \mathbf{F}'\mathbf{D}^{-1}(\mathbf{I} - \gamma\mathbf{A}) \\ &\quad \times \{\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{D}^{-1}(\mathbf{I} - \gamma\mathbf{A})\mathbf{X})^{-1}\mathbf{X}'(32) \\ &\quad \times \mathbf{D}^{-1}(\mathbf{I} - \gamma\mathbf{A})\}\mathbf{F}/n. \quad (33) \end{aligned}$$

Now we substitute equation (33) and equation (31) into equation (29) and use a Splus program to compute the new negative log-likelihood,  $L(\gamma)$ , as a function of  $\gamma$ .

The maximum likelihood estimate of  $\gamma$  is obtained when the negative log-likelihood is minimized.

Once we have  $\hat{\gamma}$  we can construct a  $100(1 - \alpha)$  confidence interval for  $\hat{\gamma}$  from  $L(\gamma)$  where  $L(\gamma)$  is the negative log-likelihood as a function of  $\gamma$  and  $100(1 - \alpha)$  is our significance level. To do this Cressie uses a method developed by Whittle. First he notes that when  $\gamma$  is the true value,

$$P \left\{ L(\gamma) \leq L(\hat{\gamma}) + (n/(n - r - 2))\chi_1^2(\alpha)/2 \right\} \simeq 1 - \alpha \quad (34)$$

Here,  $\alpha$  is the level of the test,  $\chi_1^2(\alpha)$  is the upper  $100(1 - \alpha)\%$  point on the chi-squared distribution with one degree of freedom,  $n$  is the number of counts, and  $r$  is the number of parameters to be fit (i.e. the parameters defined in the mean model, in our case it is the trend). The test of

$$H_0 : \gamma = 0 \text{ versus } H_1 : \gamma \neq 0 \quad (35)$$

is rejected (i.e. spatial dependence is detected) when  $L(0)$  is not contained in the confidence interval. A hypothesis test for

$$H_0 : \beta = 0 \text{ versus } H_1 : \beta \neq 0 \quad (36)$$

was formed the same way. Except the matrix  $X$  in equation 31 was modified to reflect our null hypothesis. We then used a likelihood ratio statistic of

$$2 \left( L(\hat{\beta}, \hat{\sigma}^2, \hat{\gamma}) - L(\hat{\beta}_0, \hat{\sigma}_0^2, \hat{\gamma}_0) \right) \quad (37)$$

where  $\hat{\beta}_0, \hat{\sigma}_0^2$ , and  $\hat{\gamma}_0$  are the values under the null. The value in equation 37 was compared to  $\chi_1^2$ . All computations were performed by a program written in Splus

To verify our simulations we used our 1900 simulated data sets and then transformed the data to Gaussian and test the null hypothesis. The results are shown in the tables.

Method of simulation	Using auto-Poisson	Using GLMM
True Null % significant for $\beta$	$\beta = 0$ $\gamma = .01$ $\beta = 0$ 1%	$\beta = 0$ $\gamma = .01$ $\beta = 0$ 2%
True Null % significant for $\beta$	$\beta = 0$ $\gamma = .05$ $\beta = 0$ 3%	$\beta = 0$ $\gamma = .05$ $\beta = 0$ 3%
True Null % significant for $\beta$	$\beta = .01$ $\gamma = 0$ $\gamma = 0$ 4%	$\beta = .01$ $\gamma = 0$ $\gamma = 0$ 3%
True Null % significant for $\gamma$	$\beta = .05$ $\gamma = 0$ $\gamma = 0$ 4%	$\beta = .05$ $\gamma = 0$ $\gamma = 0$ 6%

Method of simulation	Using auto-binomial	Using GLMM
True Null % significant for $\beta$	$\beta = 0$ $\gamma = .01$ $\beta = 0$ 3%	$\beta = 0$ $\gamma = .01$ $\beta = 0$ 5%
True Null % significant for $\beta$	$\beta = 0$ $\gamma = .05$ $\beta = 0$ 3%	$\beta = 0$ $\gamma = .05$ $\beta = 0$ 5%
True Null % significant for $\beta$	$\beta = .01$ $\gamma = 0$ $\gamma = 0$ 4%	$\beta = .01$ $\gamma = 0$ $\gamma = 0$ 5%
True Null % significant for $\gamma$	$\beta = .05$ $\gamma = 0$ $\gamma = 0$ 4%	$\beta = .05$ $\gamma = 0$ $\gamma = 0$ 5%

# Using GLMM

The GLMM is defined as

$$g(\eta) = \mathbf{X}\beta + \mathbf{Z}\varepsilon \quad (38)$$

where the matrix,  $\mathbf{X}$ , defines the fixed effect part, the matrix  $\mathbf{Z}$  defines the random-effect part,  $\beta$  is a fixed effect unknown constant,  $\varepsilon$  are random variables drawn from a distribution, and  $g()$  is the link function (e.g.  $\log(\eta)$  for the Poisson case and  $\log(p/(1-p))$  for the binomial case). More specifically, in our case we have for the auto-Poisson case that,

$$\log(\eta_i) = \alpha + \sum_{s=1}^k \beta_s x_i + \varepsilon_i. \quad (39)$$

and for the auto-binomial case,

$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha + \sum_{s=1}^k \beta_s x_i + \varepsilon_i. \quad (40)$$

where,

$$\varepsilon \sim \text{Gau}(\mathbf{0}, \Sigma).$$

- $\Sigma$ , is where the spatial correlation structure is defined.

Common structures are:

Spherical

$$f(d_{ij}) = [1 - 1.5(d_{ij}/\rho) + 0.5(d_{ij}/\rho)^3]I(d_{ij} < \rho) \quad (41)$$

Exponential

$$f(d_{ij}) = [\exp(-d_{ij}/\rho)] \quad (42)$$

Gaussian

$$f(d_{ij}) = [\exp(-d_{ij}^2/\rho^2)] \quad (43)$$

and Linear

$$f(d_{ij}) = (1 - \rho d_{ij})I(d_{ij} < \rho) \quad (44)$$

where  $I(\cdot)$  is the identify function and  $\rho$  is the range at which locations are no longer correlated.

- Splus allows for you to construct your own correlation structures for estimation.
- Liang and Zeger developed a method to perform parameter estimation called Generalized Estimating Equations (GEE). GEE provides a method to analyze correlated data that otherwise could have been modeled as GLM. Release 6.12 and later of SAS implements Liang and Zeger's GEE method.

To verify our simulations we show the results of the null hypothesis test in the following tables. Note that the GLMM method of estimation appears conservative in our simulations. Meaning observed type I error is less than nominal type I error.

Method of simulation	Using auto-Poisson	Using GLMM
True Null % significant for $\beta$	$\beta = 0$ $\gamma = .01$ $\beta = 0$ 3%	$\beta = 0$ $\gamma = .01$ $\beta = 0$ 2%
True Null % significant for $\beta$	$\beta = 0$ $\gamma = .05$ $\beta = 0$ 4%	$\beta = 0$ $\gamma = .05$ $\beta = 0$ 2%
True Null % significant for $\beta$	$\beta = .01$ $\gamma = 0$ $\gamma = 0$ 4%	$\beta = .01$ $\gamma = 0$ $\gamma = 0$ 1%
True Null % significant for $\gamma$	$\beta = .05$ $\gamma = 0$ $\gamma = 0$ 3%	$\beta = .05$ $\gamma = 0$ $\gamma = 0$ 4%

Method of simulation	Using auto-binomial	Using GLMM
True Null % significant for $\beta$	$\beta = 0$ $\gamma = .01$ $\beta = 0$ 4%	$\beta = 0$ $\gamma = .01$ $\beta = 0$ 4%
True Null % significant for $\beta$	$\beta = 0$ $\gamma = .05$ $\beta = 0$ 6%	$\beta = 0$ $\gamma = .05$ $\beta = 0$ 5%
True Null % significant for $\beta$	$\beta = .01$ $\gamma = 0$ $\gamma = 0$ 3%	$\beta = .01$ $\gamma = 0$ $\gamma = 0$ 3%
True Null % significant for $\gamma$	$\beta = .05$ $\gamma = 0$ $\gamma = 0$ 4%	$\beta = .05$ $\gamma = 0$ $\gamma = 0$ 5%

## Mean Squared Error

$$MSE_{\beta} = E[(\hat{\beta} - \beta)^2] \quad (45)$$

where  $\hat{\beta}$  is an estimator of  $\beta$ , and

$$MSE_{\gamma} = E[(\hat{\gamma} - \gamma)^2], \quad (46)$$

where  $\hat{\gamma}$  is an estimator of the true value  $\gamma$ .

We estimated the MSEs using 1900 replications to compute the size of the error for each parameter, i.e.

$$M\hat{S}E_{\beta} = \frac{1}{1900} \sum_{i=1}^{1900} (\hat{\beta}_i - \beta)^2, \quad (47)$$

and

$$M\hat{S}E_{\gamma} = \frac{1}{1900} \sum_{i=1}^{1900} (\hat{\gamma}_i - \gamma)^2. \quad (48)$$

# Results

- Winsorization on Data Simulations

Let  $z_i$ ,  $i = 1, \dots, n$  be independent Poisson( $v_i$ ) random variables, respectively. Also, let  $v_m = \max_{i=1, \dots, n} v_i$ . Then,

$$P(\{z_1 \cup \dots \cup z_n\} \geq 3v_m) = (49)$$

$$P(\text{at least one of } \{z_1, \dots, z_n\} \geq 3v_m) = (50)$$

$$1 - P(\text{none of } \{z_1, \dots, z_n\} \geq 3v_m). \quad (51)$$

$$(52)$$

Now, before we can finish the calculation we need to define the following: Let,

$$x_i = \begin{cases} 1 & \text{if } z_i \geq 3v_m \\ 0 & \text{otherwise} \end{cases}$$

It is clear that  $x_i \sim \text{Bernoulli}(p_i)$  where,

$$p_i = P(z_i \geq 3v_m) \quad (53)$$

$$= 1 - P(z_i < 3v_m) \quad (54)$$

$$= 1 - P(z_i \leq \lceil 3v_m - 1 \rceil) \quad (55)$$

$$= 1 - \sum_{z=0}^{\lceil 3v_m - 1 \rceil} \frac{e^{-v_i v_i z}}{z!}. \quad (56)$$

Now, continuing from equation (52), we have that

$$= 1 - P(\text{none of } \{z_1, \dots, z_n\} \geq 3v_m) \quad (57)$$

$$= 1 - P(x_1 = 0, x_2 = 0, \dots, x_n = 0) \quad (58)$$

The  $x_i$ 's are independent of each other since the correlated means that are produced during the Gibbs sampler in step 1 have no bearing on any of the individual locations. Thus,

$$= 1 - P(x_1 = 0, x_2 = 0, \dots, x_n = 0) \quad (59)$$

$$= 1 - [P(x_1 = 0) \times P(x_2 = 0) \times \dots \times P(x_n = 0)]$$

$$= 1 - \prod_{i=1}^n (1 - p_i). \quad (60)$$

$$V = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 1 \\ 2 \\ 3 \\ \vdots \\ 1 \\ 2 \\ 3 \end{pmatrix}$$

$x$	$P(\mathbf{Z} \geq x \times v_m)$
1	1
2	.72
3	.037
4	.0005
5	$4.11 \times 10^{-6}$
6	$1.8 \times 10^{-8}$
7	$5.57 \times 10^{-11}$
8	$1.02 \times 10^{-13}$
9	0
10	0

# Future Research

- Irregularly spaced lattices.
- Consider the issue of prediction on spatial data sets
- Overdispersion-happens when your variance is greater than your mean.