SIMULATING DISCRETE SPATIALLY CORRELATED POISSON DATA ON A LATTICE

Monica C. Jackson\textsuperscript{1} $^\S$, Kimberly Flagg Sellers\textsuperscript{2}

\textsuperscript{1} Department of Mathematics and Statistics
American University
Washington, DC 20016-8050, USA
e-mail: monica@american.edu

\textsuperscript{2} Department of Mathematics
Georgetown University
Washington, DC 20057, USA
e-mail: kfs7@georgetown.edu

Abstract: Spatial data analyses assess patterns in data and present results as graphs, maps, and model parameter estimates. Several spatial models are available, but there are poor guidelines for choosing among them. We simulate data under two methods for discrete Poisson data on a lattice. A Gibbs sampling algorithm is used to generate conditionally-specified data and a generalized linear mixed model with random effects having spatial correlation structure for simultaneously-specified data. The limitations derived on the generation of spatial data sets are also presented.

AMS Subject Classification: 62H10
Key Words: spatial statistics, spatial correlations

1. Introduction

Spatial data analysis tries to reduce spatial patterns in data to relevant summaries, usually presented in the form of graphs, maps, and estimates of model parameters (see Pocock et al [27] and MacNab and Dean [22]). Before beginning a spatial statistical study, it is necessary to determine whether the locations are: i) regularly- or irregularly-spaced, ii) represent points or regions,
and iii) whether the random variable of interest is a continuous or discrete. While models that explicitly model spatial relationships have been available for over a decade, one can continue to find numerous studies in the biological or social sciences which ignore the spatial component. Pielou [26] gives an example, where the autocorrelation is the main focus of the study. Pielou studies the distribution of diseased trees among healthy trees.

Kaluzy et al [20, p. 315] define lattice data as “spatial data with observations associated with irregularly or regularly spaced regions”. In spatial data, we do not have the unidirectional flow of time that occurs with time series. Instead, spatial models are often built on nearest neighbors. The simultaneously and conditionally-specified models are the two common types of models used to handle spatially correlated lattice data. We focus on discrete data collected on a lattice. Under this format, regions (or lattice models) are typically linked or correlated with their neighbors. Neighbors must be clearly defined to determine associated correlation. This can be a particular challenge in the case of an irregular lattice. Factors such as distance or border-sharing can help determine a region’s neighbors. For discrete data, there are several models including generalized linear mixed models (GLMM) and conditionally-specified Poisson and binomial models. This paper will investigate some of the simulation issues related to these models.

2. Lattice Models

Simultaneously-specified models and conditionally-specified models are two primary methods to define spatial models. Cressie [8, p. 405] states that simultaneously-specified models are analogous to the way autoregressive models are specified in time series. In contrast its conditionally-specified models are used when one is interested in how a location interacts with their predefined neighbors (see Royle [31]). In conditionally-specified models the spatial correlation is defined through the mean, whereas in simultaneously-specified models the spatial correlation lies in the variance-covariance structure. Most discrete data are analyzed using a transformation for approximate Normality and then models based on the Gaussian distribution are used. Diggle et al [10] suggest another method which avoids transforming data but uses linear kriging methodology.

Brook [5] is credited by Cressie [8, p. 408] for being the first to cite the difference between the simultaneously-specified models and the conditionally-specified models. Brook's description is based on the one-dimensional Markov
Chain with neighbors defined as the first-order nearest neighbor, i.e. for the conditionally-specified model

\[ P(z_i | z_1, z_2, z_3, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n) = P(z_i | z_{i-1}, z_{i+1}), \text{ for } i = 1, \ldots, n. \]  

(1)

More specifically, if there was a spatial transect with equally-spaced locations, 1, 2 and 3, then the first-order nearest neighbors of location 2 are locations 1 and 3. The simultaneously-specified models makes no such restriction on defining neighbors. Brooks's results extend to more general neighbor structures. Kaiser and Cressie [18] look at a Markov random field approach to define simultaneously-specified and conditionally-specified distributions for two dimensional data. They use a joint distribution through the specification of the conditional distributions. A benefit to their method is that spatial dependencies are specified site by site through a conditional mean. However, they state that a drawback to their method are unresolved parameterization issues.

Rathbun [28] shows a method of defining neighbors based on other influences besides distance between two locations (e.g. how far a location is from a necessary water source) for locations not on a regular grid. Sometimes distance matrices are used in models to illustrate the effect that locations have on one another (Cressie [8]).

For example, consider the matrix

\[
\Theta = \begin{pmatrix}
\theta_{11} & \theta_{12} & \theta_{13} & \cdots & \theta_{19} \\
\theta_{21} & \theta_{22} & \theta_{23} & \cdots & \theta_{29} \\
\theta_{31} & \theta_{32} & \ddots & \ddots & \vdots \\
\vdots & \vdots & \ddots & \ddots & \vdots \\
\theta_{91} & \cdots & \cdots & \theta_{98} & \theta_{99}
\end{pmatrix}
\]  

(2)

where \( \theta_{ij} \) (for \( i, j = 1, \ldots, 9 \)) represents the relative strength of the spatial dependence between locations \( i \) and \( j \). The stronger the dependence between the two locations, the higher the value of \( \theta_{ij} \). It is not always true that \( \theta_{ij} = \theta_{ji} \), meaning that the effect that location \( i \) has on location \( j \) may not be the same as the effect that location \( j \) has on \( i \). For example, let location \( i \) represent a large metropolitan city and location \( j \) represent a small neighboring city. The larger city may have a stronger effect on the smaller city since residents of a small city are more likely to travel into the larger city for needs such as entertainment, thus causing \( \theta_{ij} < \theta_{ji} \). In streams, dependence between points upstream and downstream is apt to be symmetric for some kinds of measurements but not others. So careful thought to the system and variables being studied is necessary in formulating a reasonable spatial dependence structure. The interpretation of the matrix in equation (2) has a different meaning based on the spatial model's
simultaneous or conditional definition. Both cases are presented in the following section.

3. Conditionally-Specified versus Simultaneously-Specified Model

Ripley [29, p. 90] states that all simultaneously-specified Gaussian models can be expressed as conditionally-specified models by defining

$$\Omega = \Theta + \Theta' - \Theta'\Theta. \quad (3)$$

Where, $\Omega$ is a $n \times n$ spatial dependence matrix for the conditionally-specified case and $\Theta$ is a $n \times n$ spatial dependence matrix for the simultaneously-specified case. Cressie [8, p. 409] shows that not all conditionally-specified models can be expressed as simultaneously-specified models, i.e. the reverse is not generally true. However, in some specific cases where the lattice is finite, it is feasible to express a conditionally-specified Gaussian model as a simultaneously-specified model by defining

$$\Theta = I - L', \quad (4)$$

where $LL'$ is the Cholesky decomposition of $I - \Omega$ (see Ripley [29, p. 90]). Arnold et al [1, Chapter 14] describe a simple example of the difference between the simultaneously-specified models and conditionally-specified models using a linear normal model in two dimensions. Based on this discussion, one might normally consider the simultaneous model to be preferable, but the conditional model may be more intuitive to a scientist in some application. Because of the symmetry requirement for $\Theta$, the conditional model may involve fewer parameters. Sections 3.1-3.4 describe some common spatial models considered in spatial analysis; see Jackson [16] for further details.

3.1. Conditionally-Specified Poisson Model

One can use the conditionally-specified Poisson model to describe counts of interest. It is common with disease data and almost always with abundance data. This auto model is defined as

$$P(zi|zj) = \frac{\exp(-\eta_i)\eta_{i}^{z_i}}{z_i!}, \quad (5)$$

where

$$\eta_i = \exp\left(\mu_i + \sum_{j=1}^{n} \omega_{ij}z_j\right). \quad (6)$$
Here, $\Omega = (\omega_{ij})$ is the $n \times n$ spatial dependence matrix of locations $i$ and $j$, where $i = 1 \ldots n$ and $j = 1 \ldots n$. We also have the added restrictions that $\omega_{ij} = \omega_{ji}, \omega_{ii} = 0$, and $\omega_{ij} = 0$ if location $j$ is not a predefined neighbor of location $i$. Since this is a conditional model the $\mu = (\mu_i)$ is an $n \times 1$ vector where typically $\mu$ is modeled as $X\beta$, and $X$ is an $n \times k$ covariate matrix with $k$ covariates and slope, $\beta$. This model has been used in cancer studies. For example, Ferrandiz et al [11] use this model to predict trends in cancer mortality in Valencia, Spain, where the spatial distribution of bladder, colon, prostate, and stomach mortality counts were studied. They found that the concentration of nitrate in drinking water was significantly associated with stomach cancer mortality. They use an offset to adjust for the number of people at risk.

3.2. Restriction on the Conditionally-Specified Poisson Model

The main problem with the conditionally-specified Poisson model is that only negative values of the parameter $\omega_{ij}$ are allowed in the model. That result is a consequence of the conditions listed in the Hamersley-Clifford Theorem (see Cressie [8, p. 417]). Assume that a count of zero is possible at all locations. Therefore, $P(0) > 0$ where $P(\cdot)$ represents the conditional probability of an event. More formally, define $x_1, x_2, \ldots, x_n$ to be random events occurring at sites $1, \ldots, n$ respectively. Then a conditional probability can be given by

$$P(x_n|x_1, x_2, \ldots, x_{n-1}).$$

We define the sample space $\xi = \{Z : P(Z) > 0\}$. Now, for $Z \in \xi$, we may legitimately define a function as

$$Q(Z) = \log \left\{ \frac{P(Z)}{P(0)} \right\}; \quad (7)$$

this is the most general form of $Q(Z)$. By defining $Z_i$ as the realizations $(z_1, \ldots, z_{i-1}, 0, z_{i+1}, \ldots, z_n)$, we can similarly define $Q(Z_i)$ via equation (7) to get

$$Q(Z) - Q(Z_i) = \log \left\{ \frac{P(Z)}{P(0)} \right\} - \log \left\{ \frac{P(Z_i)}{P(0)} \right\}$$

$$= \log \left\{ \frac{P(Z)/P(0)}{P(Z_i)/P(0)} \right\} = \log \left\{ \frac{P(Z)}{P(Z_i)} \right\},$$

thus obtaining

$$\exp \{Q(Z) - Q(Z_i)\} = \frac{P(Z)}{P(Z_i)}. \quad (8)$$
Besag [2] shows there is an expansion of $Q(Z)$ defined as
\[ Q(Z) = \sum_{1 \leq i \leq n} z_i G_i(z_i) + \sum_{1 \leq i \leq j \leq n} z_i z_j G_{i,j}(z_i, z_j), \]  
(9)
where $G_{i,j}(\cdot) = 0$ unless sites $i$ and $j$ are neighbors. If sites $i$ and $j$ are neighbors, $G_{i,j}(\cdot)$ takes on a value which is not uniquely defined. More specifically for the conditionally-specified defined in equations (5) and (6),
\[ Q(Z) = \sum \left\{ \mu_i - \ln (z_i!) \right\} + \sum \sum \omega_{ij} z_i z_j \] 
(10)
(see Brook [5]). Since the range of $Z$ is infinite, we must assure that $\exp Q(Z)$ is summable over $Z$.

Clearly, $\exp(Q(Z))$ is summable whenever $\omega_{ij} \leq 0$. To show this consider a simple realization, where $z_1$ and $z_2$ takes on a positive integers on $(0, \infty)$ and all other random variables, $z_3, \ldots, z_n$, take on 0. The exponential of equation (10) can now be written as
\[ \exp \left( Q(z) \right) = \exp \left( \frac{\mu_1 + \mu_2 + \omega_{12} z_1 z_2}{z_1! z_2!} \right). \] 
(11)
Now we see equation (11) will converge over all non-negative integers of $z_1$ and $z_2$ if $\omega_{ij} \leq 0$. Therefore the conditionally-specified Poisson model requires a negative spatial dependence. Positive values for $\omega_{ij}$ will cause the probability sum to diverge to infinity, thus making parameter estimation impossible. There are some methods around this problem if the conditionally-specified Poisson model is to be used with positive dependence. One method is Winsorization, which is described in the next section.

### 3.3. Winsorized Conditionally-Specified Poisson Model

The Winsorized conditionally-specified Poisson model (Kaiser [19]) is effective in modeling data sets which consist of positively and negatively dependent observations. Consider a $n \times 1$ vector, $Y = (y_i)_{n \times 1}$, where the elements are random variables whose support is a set of nonnegative integers. A possible Winsorization could be defined as follows
\[ Y = Z I(Z \leq R) + R I(Z > R), \] 
(12)
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 its condition (e.g., $Z \leq R$) is met and 0 otherwise. This transforms the vector, $Z = (z_i)$ containing infinite support, to a new random variable $Y$ with support on $(0, R)$. This transformation is commonly known as Winsorization. Kaiser [19] states that the choice of $R$ does...
affect the MLE of the parameters. To eliminate large biases, a value of $R$ that is large relative to the data set is the best choice. Generally, it is sufficient to choose $R$ such that $R \geq 3\eta_m$, where $\eta_m = \max \{\eta_1, \ldots, \eta_n\}$. However, in practice since $\eta_m$ is unknown, Kaiser and Cressie [19] state that it is acceptable to choose a value of $R$ greater than the largest observation. The conditional probability mass function (pmf) of $Y$ is now

$$P(y_i | y_j : j \neq i) = \exp(-\eta_i)\eta_i^{yi} / y_i!, \quad (13)$$

where

$$\eta_i = \begin{cases} 
\exp(\mu + \sum_{j=1}^{i} \omega_{ij}y_j) & \text{if } y_i \leq R - 1, \\
\exp(\mu + \sum_{j=1}^{i} \omega_{ij} (y_j - \mu_j)) - \psi_i & \text{if } y_i = R,
\end{cases} \quad (14)$$

and

$$0 < \psi_i < \mu_i. \quad (15)$$

The term $\psi_i$ comes from the Taylor expansion of $\exp(\eta)$ (Kaiser [19]). Since $E[Z] \approx E[Y]$ for large $R$, it is sufficient to choose $\psi_i = 0$ (for $i = 1 \ldots n$) for modeling purposes. We still assume the condition $\omega_{ij} = \omega_{ji}$ and that $\omega_{ij} = 0$ if location $j$ is not a neighbor of location $i$ as in previous sections.

### 3.4. Generalized Linear Mixed Models

Generalized Linear Mixed Models (GLMM) (see Breslow and Clayton [4]) are similar to the simultaneously defined models for correlated normal data. Generalized linear models (McCullagh and Nelder [23]) 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. In GLMM we also incorporate a model for Poisson count data similar to equation (6). Namely, we can define $\eta_i$ as the mean of the $ith$ point on the lattice and

$$\log \eta_i = \mu_i + \epsilon_i, \quad (16)$$

where $\mu_i$ may be defined in terms of a limited number of covariates and $\epsilon_i$ is a normally distributed error term. A model for binomial data can be defined as

$$\log \frac{p_i}{1 - p_i} = \mu_i + \epsilon_i. \quad (17)$$

In both models consideration of spatially-related locations leads us to define $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$ as $N(0, \Sigma)$ where $\Sigma$ has a spatial covariance structure (Vanleeuwen [33]). Thus the correlated error term in this model replaces the $\theta_{ij}z_j$ terms in equation (6). For example, we may define the covariance between $\epsilon_i$ and $\epsilon_j$ as
a decreasing function of distance of these two points (e.g. \( \text{Cov}(\varepsilon_i, \varepsilon_j) = \sigma^2 \rho d_{ij} \)). Several such structures are discussed in the SAS PROC MIXED documentation [21] and by Zimmerman [34]. Some not so common correlation structures are shown by Opsomer et al [24] and Zimmerman and Nunez-Anton [34]. It should be noted that these models have typically been used for point data rather than data collected on a lattice.

### 4. Simulating Correlated Poisson Data Using Conditionally Specified Model

Our first model for generating data is based on the conditionally-specified Poisson model defined in Section 3.1. However, we use a Winsorized form of the model as described in Section 3.3 in order to allow for positive spatial dependence. We use a specific neighborhood structure and a proximity index defined from Ferrandiz, et al [11]. First, they define the proximity index between the \( i \)-th and \( j \)-th locations,

\[
a_{ij} = \frac{\sqrt{N_i N_j}}{d_{ij}},
\]

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 villages \( i \) and \( j \)), and \( d_{ij} \) represents the distance in miles between location \( i \) and \( j \). They now define location \( i \) as a neighbor of location \( j \) if \( a_{ij} > a \), where \( a \) is a threshold. When \( N_i = N \lor i \), this is equivalent to defining a neighbor as being within a prespecified distance. The term \( a_{ij} \) is also known as a "weight" (see Ord [25]). It is common to use weights that are based on distances between locations for irregular lattices (see Cliff and Ord [7]).

We will place one covariate in the model corresponding to the column coordinate or "longitude". When the slope of this covariate is positive, it will cause an east-west trend in the data (i.e., the means will be larger as one travels from west to east). The conditional mean for the \( i \)-th region given the realization of the neighbors is defined by

\[
\log \eta_i = \alpha + \log N_i + \beta x_i + \sum_{j=1}^{n} \omega_{ij} z_j,
\]

where \( \alpha \) is the intercept, \( x_i \) is column coordinate of region \( i \), \( N_i = \text{size (e.g. population at risk)} \) 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
SIMULATING DISCRETE SPATIALLY CORRELATED...

(i.e. the means get larger as you travel to the east) while values of \( \beta = 0 \) imply no trend. As a result, large trend effects are easily spotted through visually looking at the data while smaller effects require use of testing methods, e.g. the slope test. We define \( \omega_{ij} \) as

\[
\omega_{ij} = \gamma a_{ij},
\]

where \( \gamma \) is an interaction parameter that allows for the spatial dependence to be proportional to the proximity index \( a_{ij} \). This means (equation (19)) will be used in the conditionally-specified Poisson model.

The random variables in the conditionally-specified model are Winsorized (as described in Section 3.3) in the following manner: Winsorization is performed whenever \( z_i > 3 \times \eta_m \) (where \( \eta_m = \max \{ \eta_1 \ldots \eta_n \} \)). Kaiser [19] shows that \( 3 \times \eta \) is an acceptable Winsorization value. The \( \eta_i \) values are not fixed. They are generated at each iteration using equation (19). At each iteration, if the Poisson count generated is greater than \( \eta_m \) Winsorization is applied.

4.1. The Algorithm

The algorithm that we are using is similar to the Gibbs sampling algorithm which is a special case of the Metropolis-Hastings algorithm (MH) (Gilks [13]). Geman and Geman [12] were among the first to find practical uses of the Gibbs algorithm. The MH algorithm is a generalization of the Metropolis algorithm (see Gilks [13, p. 5]). Suppose one is interested in generating random variables \( z_1, z_2, \ldots, z_n \) such that \( z_t \) depends only on \( z_{t-1} \), where \( t \) represents a specific time in this case. This procedure is known as a Markov chain. The MH algorithm uses Markov chains to generate data from a multivariate distribution. A common use of the MH and Gibbs sampler is to approximate integrals corresponding to posterior distributions derived in Bayesian applications. At a time \( t \), the next state \( z_{t+1} \) is chosen by sampling a point from a proposed distribution (which could depend on \( z_t \)). The Gibbs sampler works similarly. Gilks [13] discusses using the Gibbs sampler to generate data for the conditionally-specified Poisson. Kaiser and Cressie [19] also use the Gibbs sampler for their Winsorized conditionally-specified Poisson model. Casella and George [6] and Ritter and Tanner [30] give an explanation of the properties that allow the Gibbs sampler to work. Other uses of the Gibbs sampler are shown by Gilks et al [14] and Gilks and Wild [15].

To obtain a sample from a joint distribution \( p(z_1, z_2, \ldots, z_d) \), the algorithm is implemented in the following iterative manner, where \( z_i^d \) refers to random variable \( z_i \) at iteration \( d = 1 \ldots H \) and we denote the neighbors of a location
by \( j \):

**Step 1:** Sample \( z_1^{d+1} \) from \( p(z_1 | z_2^d, \ldots, z_n^d) \).

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

\[ \vdots \]

**Step \( g \):** Sample \( z_n^{d+1} \) from \( p(z_n | z_1^{d+1}, \ldots, z_{n-1}^{d+1}) \).

The process is completed after a predetermined number of steps \( g \) have been completed for each of \( H \) iterations. More specifically, for the conditionally specified distribution we have the following steps, where \( \sum_{\{j: j \neq i\}} \) means summation over all subscripts except \( j = i \):

**Step 1:** Sample \( z_1^{d+1} \) from

\[
P(z_1 | \{z_j^d : j \neq 1\}) = \exp(-\eta_1) \frac{\eta_1^{z_1}}{z_1!},
\]

where

\[
\eta_1 = \exp(\mu_1 + \sum_{j=2}^{n} \omega_{1j} z_j^d).
\]

**Step 2:** Sample \( z_2^{d+1} \) from

\[
P(z_2 | \{z_j^d : j \neq 2\}) = \exp(-\eta_2) \frac{\eta_2^{z_2}}{z_2!},
\]

where

\[
\eta_2 = \exp(\mu_2 + \sum_{\{j: j \neq 2\}} \omega_{2j} z_j^d).
\]

\[ \vdots \]

**Step \( g \):** Sample \( z_n^{d+1} \) from

\[
P(z_n | \{z_j^d : j \neq n\}) \exp(-\eta_n) \frac{\eta_n^{z_n}}{z_n!},
\]

where

\[
\eta_n = \exp(\mu_n + \sum_{j=1}^{n-1} \omega_{nj} z_j^d).
\]

### 4.1.1. Convergence and Initialization

The choice of starting values for the algorithm is not an extreme concern but may impact the number of iterations in the burn-in period. It is well established (Gilks et al [13]) that any Monte Carlo Markov Chain (MCMC) sampler will "forget" its starting values once the algorithm has run long enough, also known
as a “burn-in” period (see Gilks [13, p. 29]). The length of the burn-in period is a factor of the initial starting state and the rate of convergence to the desired distribution (see Gilks [13, p. 14]). For this simulation, we run the algorithm for different burn-in periods to determine the necessary time to obtain adequate parameter estimates. However, it is possible to obtain convergence much sooner with more appropriately determined initial starting values. Our starting means, $\eta^0$, are defined as

$$\log \eta_i^0 = \alpha + \beta x_i + \gamma \sum_{j=1}^{n} a_{ij} z_j^0,$$  \hspace{1cm} (21)

where $\alpha = -\log(\max \{N_1, \ldots, N_n\})$, $\beta x_i$ is as previously defined, and $z_j^0$ is a Poisson random variable. Also $N_i$ refers to the population at risk at location $i$.

5. Simulating Correlated Poisson Data Using Simultaneously-Specified Models

We now describe the method we used to generate data under GLMM for the Poisson distribution. This method is known as the “simultaneous” method while the previous method was the “conditional” method. The difference is that the correlation structure is defined in the error term for the simultaneous models and defined in the mean for the conditional models.

5.1. Defining the Models

We first consider the GLMM for the Poisson distribution as defined in Section 3.4. We recall from equation (16) that we defined the mean response at a location $i$ as

$$\log \eta_i = \mu_i + \epsilon_i.$$

(22)

We now define $\mu_i$ as the first part of equation (19), namely

$$\mu_i = \alpha + \log N_i + \sum_{s=1}^{k} \beta_s x_i.$$  \hspace{1cm} (23)

It is now through the $n \times 1$ vector, $\epsilon=(\epsilon_i)_{n \times 1}$, that we define the spatial dependence structure of the data. First we define

$$\epsilon \sim \text{Gau}(0, \Sigma).$$  \hspace{1cm} (24)

We then illustrate the steps necessary to define a correlation structure that is similar to the spatial dependence structure (the $\omega_{ij} z_j$ term) defined in the latter
part of equation (19). First, we define the $n \times n$ variance-covariance matrix, $\Sigma$, as

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

(25)

where $\gamma$ and $\sigma$ are scalar parameters, $D$ is a diagonal $n \times n$ matrix with elements $1/N_i$ (recall $N_i$ is the population at risk at location $i$), and $A$ is an $n \times n$ matrix with elements $a_{ij}$ as defined in equation (18); see Kaluzny et al [20, p. 128].

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

$$
\Sigma = LL^T,
$$

(26)

where $L$ is a lower triangular matrix. Now $\varepsilon$ can be generated from $L$ by

$$
\varepsilon = L\xi,
$$

(27)

where

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

By definition,

$$
E[\varepsilon] = E[L\xi] = LE[\xi] = 0
$$

and

$$
\text{Var}[\varepsilon] = \text{Var}[L\xi] = L\text{Var}[\xi]L^T = \Sigma,
$$

thus giving us our desired distribution. The precise coding of this algorithm is shown by Kaluzny [20, p. 144]. The computing cost of using this algorithm is on the order of $(n \times n)^3$.

5.2. The Algorithm

The algorithm for generating the GLMM is as follows, using a log 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 (23).

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

It should be noted that this approach can be generalized to one which considers overdispersion by use of a negative binomial and a formulation of the negative binomial given in McCullagh and Nelder [23, p. 132].
5.3. Replication

The number of replications of the simulation is an important part of the verification and validation process. In a real spatial study, the researcher will generally not have replications of a data set. However, it is important to replicate sufficiently in order to validate inferences. Suppose we would like to obtain a confidence interval for a parameter such that $\Pi = 95$ percent of the time the null hypothesis is not rejected. To assure that the coverage probability is correct, we solve the following equation for $m$ to obtain the total number of replications needed. $\Pi$ represents coverage probability, with

\[
\text{Standard error} = \sqrt{\frac{\Pi(1-\Pi)}{m}}.
\]  

(28)

For our simulations, we wish to obtain a standard error $< .005$. Solving equation (28) for $m$ we find that 1900 replications are necessary to obtain our desired confidence limit. To obtain each replication we ran the simulation algorithm for 190,000 iterations and used each 100-th iteration as a replication.

6. Winsorization and Data Simulations

In Section 3.1 we described a method, called Winsorization, which allows for positive spatial dependence in the conditionally-specified Poisson model. We also referenced Kaiser [19], who showed that a Winsorized conditionally-specified Poisson model random variable has a mean very close to the conditionally-specified Poisson model as long as $R \geq 3 \times \eta_m$, where $\eta_m = \max \{\eta_1, \ldots, \eta_m\}$. This condition proved not to be sufficient for generating spatial data sets with low counts. A problem arises during iterations of the Gibbs sampler when small counts are generated intentionally. Consider the case, where we use a small value for $\gamma$. Ferrandiz et al [11] estimated $\gamma$ as $-2^{-10}$. Also, place a slope effect just as small. During the first few iterations of the algorithm to generate data we repeatedly came across small counts, where a few of the data points generated were equal to $3 \times \eta_m$. In fact in a $10 \times 10$ dataset under these condition we were able to generate a dataset in which 5% of the data was $3 \times \eta_m$. Kaiser argues that the probability of that event happening is negligible. However, the probability of such an event obviously depends heavily on $\eta$. If the maximum mean generated is small (as with rare disease data sets) then you have a much greater chance of obtaining too many Winsorized data points. To illustrate this consider the following:
Let $z_i, i = 1, \ldots, n$ be independent Poisson ($\nu_i$) random variables, respectively. Also, let $\nu_m = \max_{i=1, \ldots, n} \nu_i$. Then,
\begin{align*}
P(\{z_1 \cup \ldots \cup z_n\} \geq 3\nu_m) &= P(\text{at least one of } \{z_1, \ldots, z_n\} \geq 3\nu_m) \quad (29) \\
&= 1 - P(\text{none of } \{z_1, \ldots, z_n\} \geq 3\nu_m). \quad (30)
\end{align*}

Now, before we can finish the calculation we need to define the following: Let
\[ x_i = \begin{cases} 1 & \text{if } z_i \geq 3\nu_m \\
0 & \text{otherwise.} \end{cases} \]

It is clear that $x_i \sim \text{Bernoulli}(p_i)$ where,
\begin{flalign*}
p_i &= P(z_i \geq 3\nu_m) \quad (31) \\
&= 1 - P(z_i < 3\nu_m) \quad (32) \\
&= 1 - P(z_i \leq [3\nu_m - 1]) \quad (33) \\
&= 1 - \sum_{z=0}^{[3\nu_m-1]} \frac{e^{-\nu_i}\nu_i^z}{z!}. \quad (34)
\end{flalign*}

Now, continuing from equation (30), we have that
\begin{flalign*}
= 1 - P(\text{none of } \{z_1, \ldots, z_n\} \geq 3\nu_m) \quad (35) \\
= 1 - P(x_1 = 0, x_2 = 0, \ldots, x_n = 0). \quad (36)
\end{flalign*}

The $x_i$'s are independent of each other since the correlated means that are produced during the Gibbs sampler in Step 1 in Section 4.1 have no bearing on any of the individual locations. For example, suppose a mean, $\nu_1$, of 2 is produced for location 1 and a mean, $\nu_2$, of say 3 is produced for location 2. Then, the probability the count generated for location 1 is three times greater than $\nu_m = 3$, has no bearing on the probability that the count generated for location 2 is three times greater than, $\nu_m = 3$.

Thus,
\begin{flalign*}
= 1 - P(x_1 = 0, x_2 = 0, \ldots, x_n = 0) \quad (37) \\
= 1 - [P(x_1 = 0) \times P(x_2 = 0) \times \ldots \times P(x_n = 0)] \quad (38) \\
= 1 - \prod_{i=1}^{n} (1 - p_i). \quad (39)
\end{flalign*}

Now, to illustrate the problems encountered when generating Winsorized data sets with small counts, we consider the following $99 \times 1$ mean vector,
\[ V = (1 \ 2 \ 3 \ 1 \ 2 \ 3 \ \cdots \ 1 \ 2 \ 3)^T; \]
from here we see $\nu_m = 3$. Computing the calculation in equation (39) we get .037. Thus, almost a 4% chance of generating a random variable that was
<table>
<thead>
<tr>
<th>$x$</th>
<th>$P(Z \geq x \times v_m)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>.72</td>
</tr>
<tr>
<td>3</td>
<td>.037</td>
</tr>
<tr>
<td>4</td>
<td>.0005</td>
</tr>
<tr>
<td>5</td>
<td>$4.11 \times 10^{-6}$</td>
</tr>
<tr>
<td>6</td>
<td>$1.8 \times 10^{-8}$</td>
</tr>
<tr>
<td>7</td>
<td>$5.57 \times 10^{-11}$</td>
</tr>
<tr>
<td>8</td>
<td>$1.02 \times 10^{-13}$</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 1: Probability a random variable, $z_i$, is greater than $x \times v_m$ where $v_m = 3$ is the maximum mean of random variables from a $99 \times 1$ vector, $Z$, with elements $z_i$.

Winsorized to 9 (i.e. $3v_m$) at each iteration. Computing equation (39) with a vector, $V$ with larger counts and a larger variance does yield a probability close to zero for generating a random variable Winsorized. This does confirm with Kaiser’s [19] results. However, when the counts are small it became necessary to Winsorize at a value greater than or equal to $9 \times v_m$ to achieve the same results Kaiser obtained. A small *Splus* program was written to compute the calculation in equation (39) and some results are shown in the Table 6.

7. Conclusions and Future Work

In this research, we simulated data using algorithms for discrete data that exhibit spatial correlation to varying degrees. We found a limitation in the work of Kaiser and Cressie [19] on Winsorization. We showed that Winsorizing a value at three times the maximum mean may not work well for data sets with very low counts. This result was summarized in Table 6.

There are many directions for future research on this problem. We have focused on the case of discrete data on regularly spaced lattices. A natural direction would be to look at irregularly spaced lattices. Jones and Vecchia [17] have looked at irregularly spaced lattices for continuous data. We could apply their methods to data that has been transformed to approximately continuous data. Besag [3] shows how the simultaneously-specified and conditionally-specified
models are applied to non-lattice data. Irregularly spaced lattices occur predominantly in ecology studies as well as with social, economic and health data aggregated by political units such as counties or states (see Skellam [32] and Diggle [9]).

References


