Spatial Simulation and Estimation of Generalized Linear Mixed Models with Non-Normal Data

Lynette Smith[1], Aimee Schwab[2,*] and David Marx[2]

[1] *College of Public Health, University of Nebraska Medical Center, Omaha, NE, USA 68198-4375; lmsmith@unmc.edu*
[2] *340 Hardin Hall, University of Nebraska-Lincoln, Lincoln, NE, USA 68583-0963; schwab.aimee@huskers.unl.edu, david.marx@unl.edu*
[*] *Corresponding author*

Abstract. *It is often of interest to predict spatially correlated discrete data, such as counts arising from disease incidence or mortality rates, or indicator variables arising from population thresholds or measuring presence or absence of a given phenomenon. A generalized linear mixed models (GLMM) approach to prediction using Poisson and Bernoulli response variables conditional on the spatial location is simulated using G-side models. We simulated data from a Poisson and Bernoulli distribution with spherical correlation structure, and separately simulated covariates correlated with the original variable from Gaussian, Binomial, and Beta distributions. This was accomplished using NORTA (Normal to Anything) after simulating a spatial Gaussian structure. We then compared prediction of unobserved spatial locations under various conditions: with the entire response variable (Poisson or Bernoulli) available or various fractions of it missing, and with the entire covariate variable (Gaussian, Binomial, Beta) or some of it missing. We also fit a multivariate GLMM with both the response variable and the covariate as outcome variables to compare its prediction with the other scenarios as described. We found, as expected, the addition of a covariate improved prediction in the GLMM models. However, the comparison of interest is looking at the effect of the various covariate distributions. Moreover, spatial modeling of non-normal data with GLMM presents some unique challenges, and should not be pursued without prior understanding.*

Keywords. *Spatial prediction; Poisson; Bernoulli, NORTA, GLMM.*

1 Introduction

In medical science and epidemiology, disease mapping is an important part of identifying trends, determining causes of disease and predicting where new case will occur. Often the number of cases at several locations, or the presence or absence of cases, are determined and presented for analysis to make

predictions or determine geographical correlations with risk factors in the environment.

Gotway and Wolfinger discuss several methods of spatial prediction for disease counts and rates. They compare kriging, the traditional method of prediction in spatial analysis, to a generalized linear mixed model (GLMM) analysis method using the SAS macro GLIMMIX fitting both conditional and marginal models (Gotway, 2003). Gotway and Wolfinger simulated count data using a Gaussian random field, which is not ideal for discrete data, whether recorded in counts (Poisson) or as indicators (Bernoulli).

Yahav and Shmueli describe a method of generating multivariate Poisson data called NORTA (Normal To Anything) based on simulating data from a multivariate Normal distribution and converting it to a continuous distribution using the inverse cumulative distribution function (Yahav and Shmueli, 2012). Prates further applied this method of simulation with a spatial correlation structure in a Bayesian context (Prates, 2012).

Oliver describes a method of co-simulating spatial random fields that allows for different auto-covariance models for the two fields. In this way one can simulate an outcome variable with a particular spatial structure (such as a spherical auto-covariance) along with a covariate for the outcome variable with a completely different spatial structure (such as an exponential auto-covariance) and yet maintain a correlation between the two spatial random fields.

In this paper we examine the predictive ability of GLMM (generalized linear mixed models) for unobserved spatial locations under various conditions: with the entire response variable (Poisson or Bernoulli) available or various fractions of it missing and with the entire covariate variable (Gaussian, Binomial or Beta) or some of it missing. We also fit a multivariate GLMM with both the response variable and the covariate as outcome variables to compare its prediction with the other scenarios as described.

2 Simulation Methods and Results

To model the Poisson spatial data, we chose a conditional framework for the GLMM, or the G-side model.

Let $Y(s_1), Y(s_2), \ldots, Y(s_n)$ be data that we have observed at spatial locations $s_1, s_2, \ldots, s_n$. Then $Y|u \sim Poisson(\lambda)$, where u is the random spatial effect. We know

$$E(Y|u) = \lambda$$

$$Var(Y|u) = \lambda$$

The traditional link function of count data is the *log* link, $\eta = g(\mu|u) = log(\lambda)$.

We can model the Bernoulli spatial data in a similar framework. For observed spatial data $Y(s_1), Y(s_2), \ldots, Y(s_n)$, $Y|u \sim Bernoulli(\pi)$. Then

$$E(Y|u) = \pi$$

$$Var(Y|u) = \pi(1-\pi)$$

For Bernoulli data, the traditional link function is the logit, $\eta = g(\mu|u) = logit(\pi) = log(\frac{\pi}{1-\pi})$.

Gotway and Wolfinger simulated count data using a Gaussian random field; however this is not ideal for the simulation of discrete data. Yahav and Shmueli describe a method of generating multivariate Poisson data that can arise from a spatial correlation structure (Yahav and Shmueli, 2012). Normal to Anything (NORTA) is an approach to generate data from a multivariate distribution with a given univariate marginal and pre-specified covariance structure.

2.1 Generating Poisson Data

A ρ-vector can be generated from a multivariate normal distribution with covariance structure R_N and then transformed to a Poisson distribution using the inverse cumulative distribution function. The algorithm is as follows.

1. Generate a ρ-dimensional Normal vector Y_N with mean vector $\mu = 0$ and variance vector $\sigma = 1$, and covariance matrix R_N.
2. For each value Y_{N_i}, $i \in 1, 2, \ldots, \rho$, calculate the Normal CDF.
3. For each $\Phi(Y_{N_i})$, calculate the Poisson inverse CDF (quantile) with mean λ_i.

$$Y_{Pois_i} = \Xi^{-1}(\Phi(Y_{N_i}))$$

where

$$\Phi(Y_{N_i}) = \int_{-\infty}^{y} \frac{1}{\sqrt{2\sigma^2}} e^{-u^2 2} du$$

$$\Xi(y) = \sum_{i=0}^{y} \frac{e^{-\lambda}\lambda^i}{i!}$$

$Y_{Pois_i}$is a ρ-dimensional Poisson vector with covariance matrix R_{Pois} and mean vector Λ. If the elements of Λ are sufficiently large, the poisson distribution is known to be asymptotically Normal and $R_{Pois} = R_N$. If one of more of the means λ are small, then $R_{Pois} \neq R_N$. However, a formula to adjust the covariance matrix is provided in Yahav and Shmueli (Yahav and Shmueli, 2012).

2.2 Generating Bernoulli Data

A similar algorithm was developed to generate Bernoulli data from a multivariate normal distribution with covaiance structure R_N. However an additional step is necessary to get the parameter π.

1. Generate a ρ-dimensional Normal vector Y_N with mean vector $\mu = 0$ and variance vector $\sigma = 1$, and covariance matrix R_N.
2. For each value Y_{N_i}, $i \in 1, 2, \ldots, \rho$, calculate the Normal CDF.
3. For each $\Phi(Y_{N_i})$, calculate the inverse CDF of a $Beta(\alpha, \beta)$ distribution, Y_{Beta_i}.
4. Let each $Y_{Beta_i} = \pi_i$, and randomly draw a $Bernoulli(Y_{Beta_i} = \pi_i)$ observation, Y_{Bern_i}.

Then Y_{Bern_i} is a ρ-dimensional Bernoulli vector with covariance matrix R_{Bern} and mean vector Π.

2.3 Preliminary Results

We tested the NORTA method by simulating a Gaussian random field on a 30×30 grid with a spherical covariance structure. The ranges simulated were 2.5, 5, and 15, assuming a sill of 1 and a nugget effect of 0. After simulating the three Gaussian random fields, the distribution was transformed to four Poisson cases and six Beta cases, while preserving the spherical covariance structure. After performing the NORTA transformation, the range, sill, and nugget, as well as the ratio of the sill to the nugget were estimated for each using ArcGIS. Maps of the data were also created for visual comparison.

Results presented will include estimated model parameters for Poisson and Bernoulli data using both ArcGIS and GLMMs implemented using PROC GLIMMIX (SAS 9.3). Results also will include addition of covariates into the model following the method outline by Oliver (2003), and a comparison of unobserved spatial locations under a variety of conditions: with the entire response variable available or fractions of it missing and with the entire covariate variable (Gaussian, Binomial, or Beta) or some of it missing.

References

[1] Gotway, C. A. and Wolfinger, R. D. (2003), Spatial prediction of counts and rates. Statistics in Medicine, 22: 1415-1432.

[2] Yahav, I. and Shmueli, G. (2012), On generating multivariate Poisson data in management science applications. Applied Stochastic Models in Business and Industry, 28: 91-102.

[3] Oliver, D. S. (2003), Gaussian cosimulation: modelling of the cross-covariance. Mathematical Geology, 35-6: 681-698.