American Journal of Theoretical and Applied Statistics
2015; 4(4): 305-311
Published online July 15, 2015 (http://www.sciencepublishinggroup.com/j/ajtas)
doi: 10.11648/j.ajtas.20150404.21
ISSN: 2326-8999 (Print); ISSN: 2326-9006 (Online)

A Simple Conditional Approach for Generating Spatial Correlated Binary Data

Renhao Jin, Tao Liu, Fang Yan, Jie Zhu

School of Information, Beijing Wuzi University, Beijing, China

Email address:
Renhao.jin@outlook.com (Renhao Jin)

To cite this article:
Renhao Jin, Tao Liu, Fang Yan, Jie Zhu. A Simple Conditional Approach for Generating Spatial Correlated Binary Data. *American Journal of Theoretical and Applied Statistics*. Vol. 4, No. 4, 2015, pp. 305-311. doi: 10.11648/j.ajtas.20150404.21

Abstract: Generating a spatial random field in which the observations are binary random variables with a particular covariance function may be impossible, because there are restrictions on the parameters of Bernoulli variables. This paper develops a conditional method based from spatial GLMM for generating spatial correlated binary data, which can generate spatial correlated binary data, with the variograms of the simulated data are similar to the variograms of the corresponding latent Gaussian random field. However, the closed form for their spatial correlation is not available specifically.

Keywords: Spatial Binary Data, Generalized Linear Mixed Model, Variogram

1. Introduction

The main goals of this paper are to offer a method to generate spatially correlated binary data, named as conditional method, which is based on spatial generalized linear mixed model (GLMM). Simulating spatial data is very important on theory research, as the worth of a spatial statistical method can be established convincingly only if the method proved to be long-run satisfactory. In many cases, the assessments of the spatial models are mainly based on simulated data. In this paper, the authors only focus on spatially correlated binary data, which are encountered in many applications ranging from epidemiology to forestry. Infectious disease data often have spatially clustered observations. In forestry binary responses, for example, the presence or absence of some disease is often observed.

Generating a spatial random field is not a simple task unless it is a Gaussian random field (GRF). However, generating a random field in which the observations are binary random variables with a particular covariance function may be impossible, because there are restrictions on the parameters of Bernoulli variables. What can be done is to generate random deviates whose marginal moments (mean and variance) "behave like" those of binary variables (Schabenberger and Gotway (2005), Chapter 7). Schabenberger and Gotway (2005) suggested the convolution representation method to generate spatially correlated binary data. However, their method can only simulate second-order stationary data, i.e., constant mean and constant variance for all observations.

Several authors have proposed different methods for generating correlated binary data. A study of their methods was performed and it was tried to extend their methods to spatially correlated binary data. However, the majority of these methods have limitations with respect to generating spatially correlated binary data with non-constant mean. For example, Lunn and Davies (1998) showed a method of generating correlated binary variables with a very simple correlation structure, which is suitable for generating variables with correlation structures which are exchangeable, and is easily extended to cater for correlation structures which are autoregressive or stationary M-dependent. However it is impossible to extend their method to general spatial correlation structures and also their method only generates binary data with constant means.

Park et al. (1996) developed a method for generating spatial binary data based on generating correlated Poisson random variables which are then recoded as zero or one. The approach by Park et al. relies on the property that any Poisson random variable can be expressed as a convolution of several other independent Poisson random variables. The binary variables have desired correlations by sharing common independent Poisson variables. The authors used this property for generating correlated Poisson variates, which are used in turn for generating correlated binary

variates. Their method allows unequal means and only positive correlations, and thus may be extended to generate spatially correlated binary data. Park et al. (1996) did discuss some restrictions of their method. Firstly, for Bernoulli data, there is a natural restriction on the correlation coefficient ρ_{ij} between two binary variates Z_i and Z_j. Note that $E(Z_iZ_j) \leq \min\{P(Z_i = 1), P(Z_j = 1)\} = \min\{p_i, p_j\}$. Therefore $\text{cov}(Z_i, Z_j) \leq \min\{p_iq_j, p_jq_i\}$, where $q_i = 1 - p_i$ and $q_j = 1 - p_j$. So ρ_{ij} is not free on $[0,1]$ but is constrained by $\rho_{ij} \leq \min\left\{[p_iq_j/(p_jq_i)]^{1/2}, [p_jq_i/(p_iq_j)]^{1/2}\right\}$. Based on this natural restriction, if $\{p_i\}$ varies a lot, all the $\{\rho_{ij}\}$ will be much smaller than 1. Then a spatial correlation structure that satisfies this restriction is difficult to find, because the spatial correlations ρ_{ij} should decrease from 1 to 0 as distances increase. Park et al. (1996) did not spell out the restrictions of their method but they gave three conditions that if they were held, their method would succeed in generating correlated binary as desired. However, to generate spatially correlated binary data, even assuming they have a constant mean, these three conditions are still not easily to satisfy in a simulation algorithm.

In this paper, a conditional method based from spatial GLMM for generating spatially correlated binary variables are developed that do not have the shortcomings of the methods above. The conditional approach listed here is similar to the simulation method in Crainiceanu, Diggle and Rowlingson (2008).

2. Methods

2.1. Spatial GLMM

To better explain the conditional method for generating spatial binary data, the spatial GLMM model is firstly described in detail. For the spatial GLMM model, the spatial data are assumed conditionally dependent on an underlying, smooth, spatial process $\{S(s): s \in D\}$. Given S(s), Z(s) has a Bernoulli distribution given by

$$Z(s_i)|\, S(s_i) \sim \text{Bernoulli}(\mu(s_i));$$

$$S(s) \sim G(0, \sigma_s^2 R_s(\alpha_s));$$

$$E[Z(s_i)|\, S(s_i)] = \mu(s_i);\ \text{Var}[Z(s)|S(s)] = \sigma^2 V_{\mu(s)};$$

$$\text{logit}\{\mu(s_i)\} = X(s_i)'\beta + S(s_i).$$

Here $\{S(s)\}$ is a Gaussian random field with mean 0 and covariance function $\sigma_S^2\rho_S(s_i - s_j; \alpha_s)$. Thus, the assumption of conditional independence defers treatment of spatial autocorrelation to the $\{S(s)\}$ process. $V_{\mu(s)}$ is a diagonal matrix with $\mu(s_i)(1 - \mu(s_i))$ in the diagonal. σ^2 is the parameter for modeling the over-dispersion in the data. As explained in the Introduction 3.1, in theory a conditional model has a marginal formulation, but the closed marginal form of $E[Z]$ and $\text{Var}[Z]$ is unavailable

The marginal mean of $Z(s_i)$ for this model is

$$E[Z(s_i)] = E_s\left[E\left(Z(s_i)\middle|S(s_i)\right)\right] = \int \frac{\exp(X(s_i)'\beta + S(s_i))}{1+\exp(X(s_i)'\beta + S(s_i))} dF_{S(s_i)} \quad (1)$$

$F_{S(s_i)}$ is the probability-distribution function of $S(s_i)$, so $F_{S(s_i)}$ is a Gaussian probability-distribution function with mean 0 and variance σ_s^2, i.e. $N(0, \sigma_s^2)$. It is difficult to obtain a theoretical expression for $E[Z(s_i)]$, but its numerical value can be easily calculated using Riemann summation. For a continuous function $f(x)$ on $[a, b]$, $\int_a^b f(x)\,dx$ always exists and can be computed by Riemann summation as

$$\int_a^b f(x)\,dx = \lim_{n\to\infty} \sum_{i=1}^{n} f(x_i^*)\Delta x_i$$

for any choice of x_i^* in $[x_{i-1}, x_i]$ with $\Delta x_i = x_i - x_{i-1}$, $\sum_{i=1}^{n} \Delta x_i = b - a$ and $\Delta x_i \to 0$.

The variance of $Z(s_i)$ and the covariance function between $\{Z(s_i)\}$ are as follows:

$$\text{Var}[Z(s_i)] = E[Z(s_i)^2] - [E[Z(s_i)]]^2, \quad (2)$$

$$\text{Cov}\left[Z(s_i), Z(s_j)\right] = E\left[Z(s_i)Z(s_j)\right] - E[Z(s_i)]E\left[Z(s_j)\right], \quad (3)$$

$$\text{Corr}\left[Z(s_i), Z(s_j)\right] = \frac{\text{Cov}[Z(s_i), Z(s_j)]}{\sqrt{\text{Var}[Z(s_i)]}\sqrt{\text{Var}[Z(s_j)]}}. \quad (4)$$

The numerical value of (2) can be calculated through the numerical value of (1). The numerical value of $E\left[Z(s_i)Z(s_j)\right]$ can also be calculated by a Riemann summation, thus the numerical values of (3) and (4) can be obtained. However, the theoretical mean and covariance of Z(s) are not available for binary data generated by this conditional method.

2.2. Algorithm of Conditional Method

Based on the definition of conditional GLMM above, the algorithm below generates spatially correlated binary data by a conditional method:

1. Generate $S(s)$, $S(s) \sim G(0, \sigma_s^2 R_s(\alpha_s))$,
2. Obtain $L(s_i)$ by $L(s_i) = x(s_i)'\beta + S(s_i)$,
3. Obtain $\mu(s_i)$ by $\mu(s_i) = \exp(L(s_i))/[1 + \exp(L(s_i))]$,
4. Generate $Z(s_i)$ using a random number generator from Bernouli $(\mu(s_i))$.

The algorithm above for simulating GLMM data is a new method but very similar to that of Crainiceanu, Diggle and Rowlingson (2008). In the simulation part of their paper, they simulated binomial data, and comparing with steps 1 and 2 in this algorithm they used random effects vector with a design matrix instead of a Gaussian random field $S(s)$.

2.3. Description of the Simulation Study

Spatial binary data $Z(s)$ with sample size 100 on a regular grid were generated. The grid is on $[0,40] \times [0,40]$ with intervals of 4 in both directions. The maximum distance

between the data points was 50.91 and a half of this was 25.46. $S(s)$ was a zero-mean intrinsically stationary Gaussian process whose variogram was continuous at the origin. The Gaussian, exponential and spherical variograms were considered. Gaussian and exponential variograms are from Matérn class of variogram functions with no nugget is given by

$$\gamma(h) = \sigma_0^2 - \sigma_0^2 \frac{1}{\Gamma(\nu)} \left(\frac{\theta h}{2}\right)^{\nu} 2K_{\nu}(\theta h)\ ) \ \nu > 0, \theta > 0.$$

The smoothness of the process increases with ν and among the most commonly used parametric variogram models are the Gaussian ($\nu = \infty$), Whittle ($\nu = 1$) and exponential ($\nu = 0.5$). The spherical variogram given by

$$\gamma(h) = \sigma_0^2 (\frac{3}{2}\frac{h}{\alpha} - \frac{1}{2}(\frac{h}{\alpha})^3)$$

is also commonly used. A nugget effect can be incorporated by adding a constant. Figure 1 gives an illustration. The spherical model attains its sill, but the Matérn models achieve their sill only asymptotically and thus their practical ranges are defined as where 95% of the sill is attained.

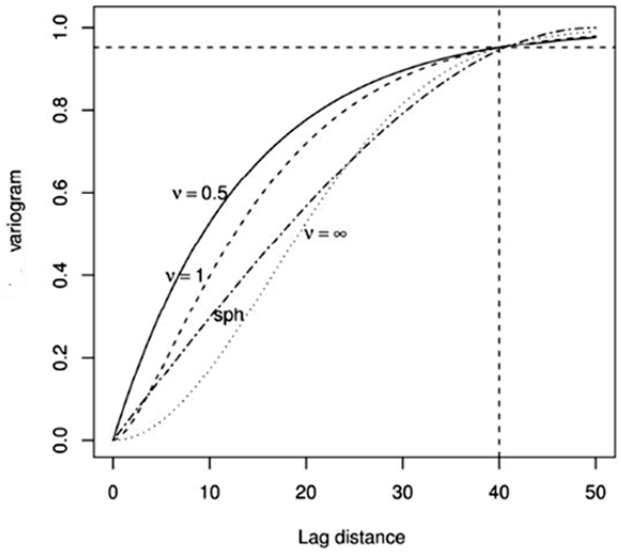


Figure 1. *Variograms for Gaussian, Whittle, exponential and spherical models with nugget* $c_0 = 0$, *sill* $c_0 + \sigma_0^2 = 1$ *and practical range 40 indicated by the vertical line. The horizontal line denotes 95% of the sill.*

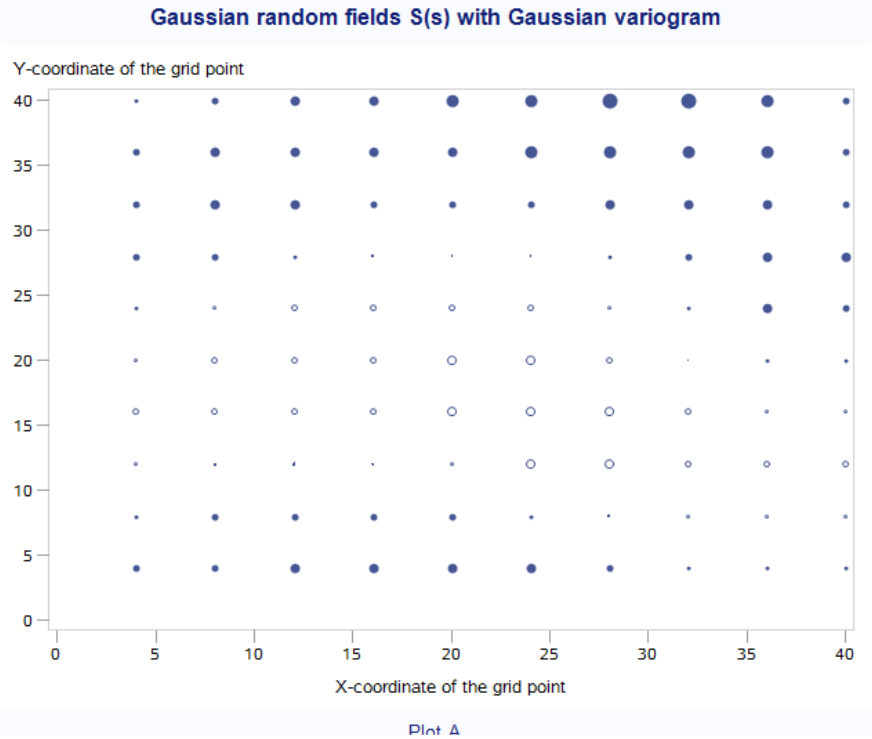
The sill of $S(s)$ was 1 with nugget 0 and its practical range was 20 for each of the three variograms. The spherical variogram attains its sill at the range, and its range is 24.65 corresponding to a practical range of 20. So now the practical ranges of the three variograms were close to one half of maximum distance between the data points, and the range of the spherical variogram was less than this distance. In the equation for the conditional mean $L(s_i) = X(s_i)'\beta + S(s_i)$, $X(s_i)'\beta$ is defined as $-2 + x_1(s_i) \cdot 1$, where $x_1(s_i)$ is a random number from a uniform distribution on $[0.5,1.5]$. This choice of $x_1(s_i)$ in $L(s_i)$ was made so that $S(s_i)$ is an important part of the model, since exp(-1)/(1+exp(-1))= 0.27, and to make the mean of the generated $Z(s)$ to be around 0.3. When a uniform random number was generated, it was kept the same for all simulations. Data were simulated by the conditional method using SAS software (SAS® 9.2, SAS Institute Inc., Cary, N.C.). The spatial $S(s)$ in the conditional method were generated by the SAS SIM2D Procedure.

3. Results

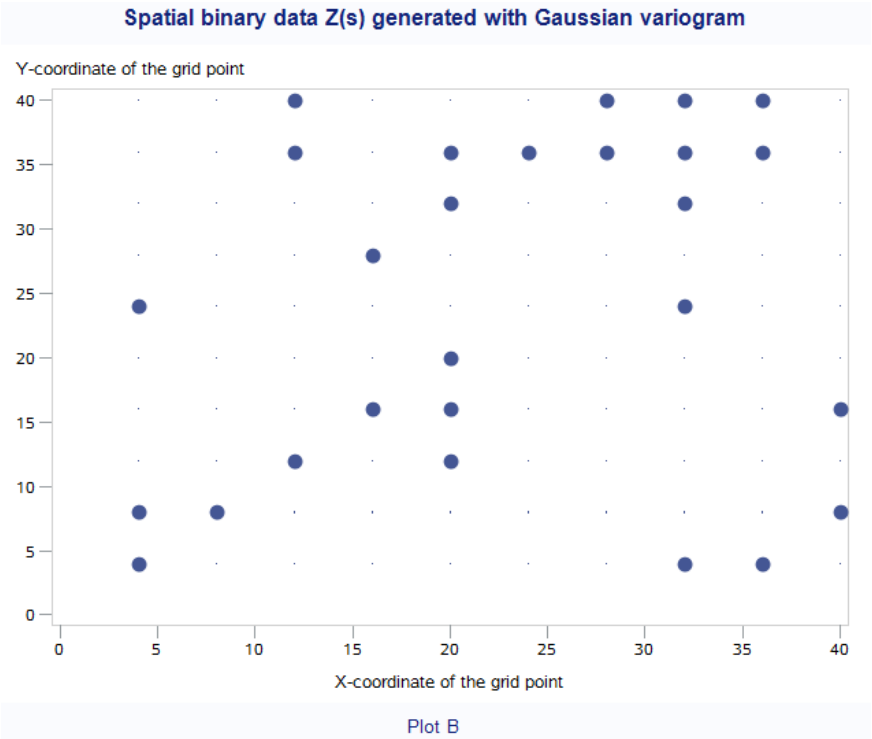
In this section, spatial binary data were simulated by the procedure described in Method section, and the binary data $Z(s)$ was generated on a regular grid and in the model $\text{logit}\{Z(s_i)|\ S(s_i)\} = x(s_i)'\beta + S(s_i)$, $x(s_i)'\beta$ was same for all simulations but the variogram of $S(s)$ varied in different simulations. Three spatial binary datasets of $Z(s)$ were generated conditionally with the variogram of $S(s)$, Gaussian, exponential and spherical respectively.

A typical realized dataset from one simulation of a Gaussian random field $S(s)$ and the corresponding spatial binary data generated by the conditional method is shown in Figure 2. From the plots, it can be seen that the spatial patterns in the generated binary data are similar to the spatial patterns in the corresponding latent Gaussian random field. Recall the conditional method procedure in Method section, where a large value of $S(s_i)$ may lead to a large $\mu(s_i)$, the mean of the $Z(s_i)$, and thus $Z(s_i)$ is likely to be 1. Comparing the spatial patterns in $Z(s)$ generated by different variogram types, little difference was found between the binary data generated by exponential and spherical variograms. However, the spatial binary data generated by Gaussian variogram had a different spatial pattern from the data by the other variogram types, being more smooth. The reason can be found from their corresponding realizations of Gaussian random fields. As shown in (a), (c), (e) of Figure 2, the Gaussian random field with Gaussian variogram is more smooth than the other two.

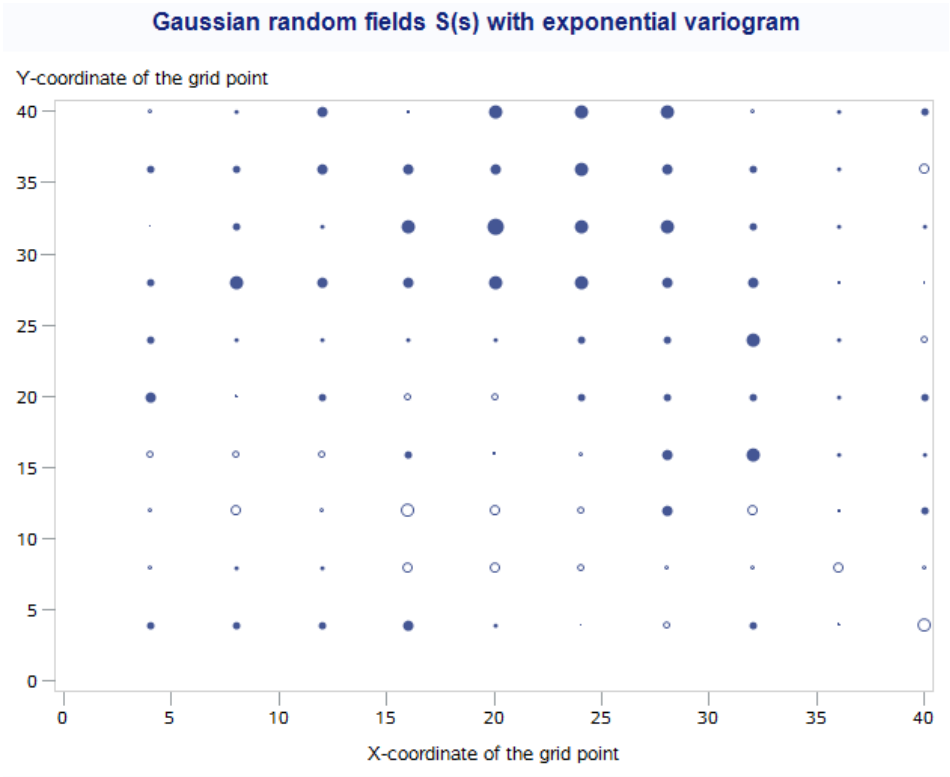
Above all, the Algorithm of conditional method in this paper can generate spatial correlated binary data, with the variograms of the simulated data are similar to the variograms of the corresponding latent Gaussian random field. However, the theoretical variogram of the binary data thus generated is still unavailable. Further work is needed to find good approximations to the correlation function of the data generated by the conditional method.


Gaussian random fields S(s) with Gaussian variogram
Y-coordinate of the grid point
0
5
10
15
20
25
30
35
40
X-coordinate of the grid point

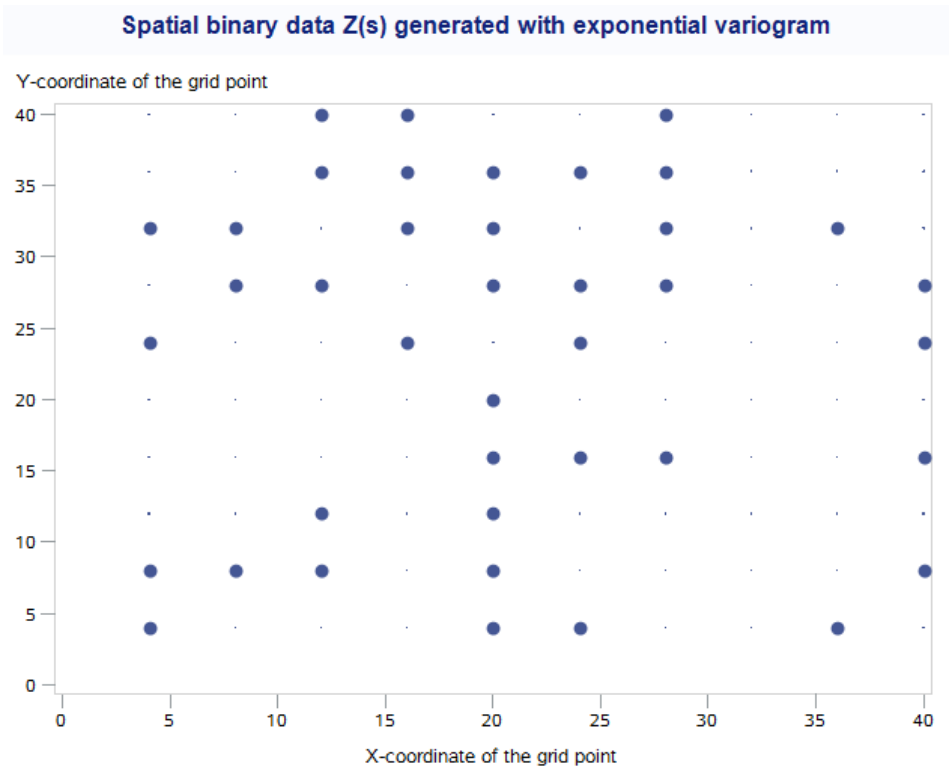
Plot A


Spatial binary data Z(s) generated with Gaussian variogram
Y-coordinate of the grid point
0
5
10
15
20
25
30
35
40
X-coordinate of the grid point

Plot B


Gaussian random fields S(s) with exponential variogram
Y-coordinate of the grid point
0
5
10
15
20
25
30
35
40
X-coordinate of the grid point

Plot C


Spatial binary data Z(s) generated with exponential variogram
Y-coordinate of the grid point
0
5
10
15
20
25
30
35
40
X-coordinate of the grid point

Plot D

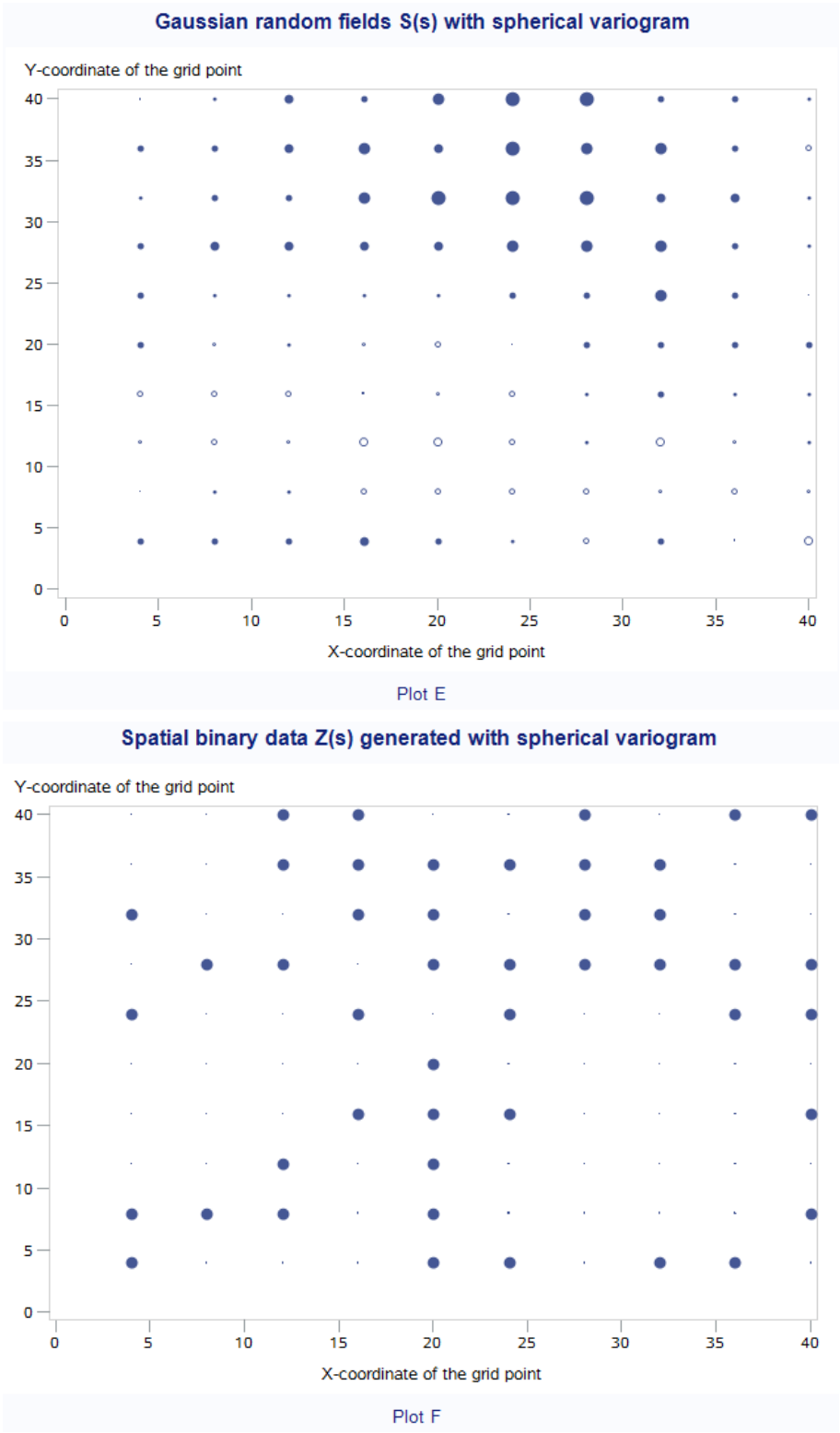


Figure 2. *The Gaussian random fields* $S(s)$ *with Gaussian, exponential and spherical variograms were generated on the grid* $[0,40] \times [0,40]$ *with intervals of 4 in both directions and shown in (a), (c), (e) respectively. Plots (b), (d), (f) are for the corresponding spatial binary data* $Z(s)$ *generated by the conditional method.*

Acknowledgements

This paper is funded by the project of National Natural Science Fund, Logistics distribution of artificial order picking random process model analysis and research (Project number: 71371033); and funded by intelligent logistics system Beijing Key Laboratory (No.BZ0211); and funded by scientific-research bases---Science & Technology Innovation

Platform---Modern logistics information and control technology research (Project number: PXM2015_014214_000001); University Cultivation Fund Project of 2014-Research on Congestion Model and algorithm of picking system in distribution center (0541502703).

References

[1] Al Osh, M. A., & Lee, S. J. (2001). A simple approach for generating correlated binary variates. Journal of Statistical Computation and Simulation, 70(3), 231-255.

[2] Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. Journal of the American Statistical Association, 88(421), 9-25.

[3] Crainiceanu, C. M., Diggle, P. J., & Rowlingson, B. (2008). Bivariate binomial spatial modeling of Loa loa prevalence in tropical Africa. Journal of the American Statistical Association, 103(481), 21-37.

[4] Cox, D. R., & Wermuth, N. (1991). A simple approximation for bivariate and trivariate normal integrals. International Statistical Review/Revue Internationale de Statistique, 59(2), 263-269.

[5] Engel, B. and Keen, A. (1992). A simple approach for the analysis of generalized linear mixed models. LWA-92-6, Agricultural Mathematics Group (GLW-DLO), Wageningen, The Netherlands.

[6] Gotway, C. A., & Stroup, W. W. (1997). A generalized linear model approach to spatial data analysis and prediction. Journal of Agricultural, Biological, and Environmental Statistics, 2(2), 157-178.

[7] Lunn, A. D., & Davies, S. J. (1998). A note on generating correlated binary variables. Biometrika, 85(2), 487-490.

[8] Liang, K. Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. Biometrika, 73(1), 13-22.

[9] Park, C. G., Park, T., & Shin, D. W. (1996). A simple method for generating correlated binary variates. The American Statistician, 50(4), 306-310.

[10] Qaqish, B. F. (2003). A family of multivariate binary distributions for simulating correlated binary variables with specified marginal means and correlations.Biometrika, 90(2), 455-463.

[11] SAS Institute Inc, (2008). SAS/STAT® 9.2 User's Guide: The GLIMMIX Procedure (Book Excerpt). NC: SAS Institute Inc, Cary.

[12] SAS Institute Inc, (2008). SAS/STAT® 9.2 User's Guide: The SIM2D Procedure (Book Excerpt). NC: SAS Institute Inc, Cary.

[13] Schabenberger, O. and Gotway, C. A. (2005). Statistical methods for spatial data analysis, Chapman & Hall/CRC, Boca Raton.

[14] Stiratelli, R., Laird, N., & Ware, J. H. (1984). Random-effects models for serial observations with binary response. Biometrics, 961-971.

[15] Waclawiw, M. A. and Liang, K. Y. (1993). Prediction of random effects in the generalized linear model. Journal of American Statistical Association 88, 171-8.

[16] Wolfinger, R., & O'connell, M. (1993). Generalized linear mixed models a pseudo-likelihood approach. Journal of statistical Computation and Simulation,48(3-4), 233-243.

[17] Zeger, S. L., & Liang, K. Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. Biometrics, 42(1), 121-130.

[18] Zeger, S. L., Liang, K. Y., & Albert, P. S. (1988). Models for longitudinal data: a generalized estimating equation approach. Biometrics, 1049-1060.