Science Research
2015; 3(4): 206-212
Published online July 23, 2015 (http://www.sciencepublishinggroup.com/j/sr)
doi: 10.11648/j.sr.20150304.18
ISSN: 2329-0935 (Print); ISSN: 2329-0927 (Online)

Generating Spatial Correlated Binary Data Through a Copulas Method

Renhao Jin, Sha Wang, 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, Sha Wang, Fang Yan, Jie Zhu. Generating Spatial Correlated Binary Data Through a Copulas Method. *Science Research*. Vol. 3, No. 4, 2015, pp. 206-212. doi: 10.11648/j.sr.20150304.18

Abstract: Simulating spatial correlated binary data is very important on many cases, but it is not easily to accomplish, as there are restrictions on the parameters of Bernoulli variables. This paper develops a copulas method to generate spatial correlated binary data. The spatial binary data generated by this method has an inverse spatial pattern comparing with the latent Gaussian random field data, however they have similar empirical variograms, although the closed form for the spatial correlation is not available specifically.

Keywords: Spatial Binary Data, Copulas, Simulation, Variogram

1. Introduction

The main goals of this paper are to offer a method to generate spatially correlated binary data through a copulas method. In probability theory and statistics, a copula is a multivariate probability distribution for which the marginal probability distribution of each variable is uniform. Copulas are used to describe the dependence between random variables. They are named for their resemblance to grammatical copulas in linguistics. Sklar's Theorem states that any multivariate joint distribution can be written in terms of univariate marginal distribution functions and a copula which describes the dependence structure between the variables. Copulas are popular in high-dimensional statistical applications as they allow one to easily model and estimate the distribution of random vectors by estimating marginal and copulas separately. There are many parametric copula families available, which usually have parameters that control the strength of dependence.

Simulating spatial data is very important on many cases. The absence of replication in most spatial data sets requires repeated observation of a phenomenon to obtain empirical estimates of mean, variation, and covariation. 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.

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. Al-osh and Lee (2001) introduced a simpler approach than that of Park et.al (1996) for generating non-negatively correlated binary data. Their proposed method only uses properties of binary random variates that eliminates the need for the intermediate step of using correlated Poisson variates as in Park et.al (1996). The key idea lies in the fact that any Bernoulli random variable can be expressed as a convolution of other

independent Bernoulli random variables and that correlation among the binary observations can arise as a result of their sharing some common elements that induce such correlation. The algorithms of Al-osh and Lee (2001) are almost the same as Park et al. (1996), and their demonstrations and results on simulating 3-dimensional binary vectors are similar, and only differ in whether the convolution is of Poisson variables or Bernoulli variables. In comparing these two methods' efficiency, Al-osh and Lee's method is expected to be more efficient since it generates binary variables directly without any intermediate step as in Park et al.'s method.

However, Al-osh and Lee (2001) did not discuss the restrictions of their method very much, and just stated that their algorithm should work for most practical cases in generating a vector of binary variates with nonnegative correlation structure. Their method is not as powerful as claimed. The restrictions on the possible combinations of mean and correlation structure required in their algorithm are such that no single simulation method can handle moderate to large sample sizes easily together with the restrictions. Unlike Park et al. (1996) and Al-osh and Lee (2001), Qaqish (2003) introduced a family of multivariate binary distributions with a certain conditional linear property. This family is particularly useful for efficient and easy simulation of correlated binary variables with a given marginal mean vector μ and correlation matrix R. His method can be used to generate spatially correlated binary data with non-constant mean, but this method also had restrictions. Qaqish (2003) stated a Lemma giving restrictions on $\{\mu, R\}$, for which his method is available. For certain patterned correlation matrices, such as exchangeable, AR(1), and MA(1) correlations, the algebraic inverse forms of their correlation matrices are available and there are simple rules to decide beforehand on the parameters in $\{\mu, R\}$ to satisfy the Lemma. However, for many other correlation matrices, such as spatial correlation matrices, it is difficult to obtain the algebraic inverse form of the correlation matrix. An example for which it works is a binary process regularly spaced on a 1-dimensional transect with exponential correlation. Since the exponential correlation now is actually AR (1) correlation and it is easy to obtain the inverse of AR (1) correlation matrices, this can be simulated by the method of Qaqish (2003). However for a binary process regularly spaced on a 2-dimensional grid, no simple rules exist for the algebraic inverse of their correlation matrices even for exponential correlation. For a general R, Qaqish (2003) then suggested trying permutations of $\{\mu, R\}$ and computing numerical inverse of the permutated R and checking those that satisified the conditions of the Lemma, but he noted that this was not a practical approach for actual simulation work as even for a small sample size 50, the number of possible permutations $50!$ is a huge figure ($50! = 3.04 \times 10^{64}$). For sample size 100, 10000 permutations for each of several (μ, R) with spatial correlation matrices were done here and none of the 10000 permutations met the conditions of the Lemma.

In this paper, a method based from copulas for generating spatially correlated binary variables are developed that do not have the shortcomings of the methods above. This copulas method is simple but are totally new and not found elsewhere.

2. Methods

2.1. Generating Spatial Binary Data Through Copulas

Copulas method is a simulation method, and it is easy to understand and manipulate. This method is wildly used to mathematical experiments, and its procedure is explained in this section.

Assume that K random variables $\{V(s_i)\}$ are K-variate normally distributed and the cumulative distribution function of each $V(s_i)$ is $F_{s_i}(.), i = 1,2,\cdots,K$. The copulas method first transforms $V(s_i)$ to $U(s_i)$ by $U(s_i) = F_{s_i}(V(s_i))$, and now $U(s_i)$ is uniform distributed in $[0,1]$. Then random variables are then generated as required based on $\{U(s_i)\}$.

Here spatially correlated binary data $\{Z(s_i)\}$ are generated based on $\{V(s_i)\}$. Let $\{V(s_i)\}$ be spatially correlated, and let $\rho\left(V(s_i), V(s_j)\right) = \rho_{ij}$. For $V(s_i)$ with arbitrary mean and variance, $U(s_i) = F_{s_i}(V(s_i))$ is always uniformly distributed in $[0,1]$. For simplicity it is assumed $V(s_i) \sim N(0,1)$. The spatially correlated binary data $\{Z(s_i)\}$ are generated as $Z(s_i) = I\{U(s_i) < EZ(s_i)\}$.

To generate spatially correlated binary data $\{Z(s_i)\}$, with $EZ(s_i) = p(s_i)$ and $\rho\left(Z(s_i), Z(s_j)\right) = \varphi_{ij}$, the procedure is as follows:

Step1. Generate spatially correlated $\{V(s_i)\}$, $V(s_i) \sim N(0,1)$ for all s_i and $\rho\left(V(s_i), V(s_j)\right) = \rho_{ij}$.

Step2. Obtain $\{U(s_i)\}$ by $U(s_i) = F_{s_i}(V(s_i))$ for all s_i.

Step3. Generate $Z(s_i)$ by $Z(s_i) = I\{U(s_i) < p(s_i)\}$ for all s_i. Now $\{Z(s_i)\}$ has $EZ(s_i) = p(s_i)$ and $\rho\left(Z(s_i), Z(s_j)\right) = \varphi_{ij}$. There is no closed form relationship between ρ_{ij} and φ_{ij}, but $\{Z(s_i)\}$ are spatially correlated based on the spatial correlation between $\{V(s_i)\}$. The nature of the correlation is investigated here through the relationship between the processes.

For an isotropic second-order stationary spatial process, the variogram function $\gamma(h)$ is expected to increase as h increases. At the range h^*, the $\gamma(h)$ achieve its sill σ^2, i.e. $\sigma^2 = \gamma(h^*)$. If the variogram achieves the sill only asymptotically, then the practical range is defined as the lag distance at which the variogram achieves 95% of the sill. For the spatial process $\{V_i\}$, if its variogram achieves the sill at h^*, for arbitrary s_i, s_j if for $d(s_i, s_j) > h^*$, $\rho\left(V(s_i), V(s_j)\right) = 0$. By the copulas algorithm, it is clear the corresponding $Z(s_i)$ and $Z(s_j)$ also has $\rho\left(Z(s_i), Z(s_j)\right) = 0$. But for arbitrary s_i, s_j if $d(s_i, s_j) < h^*$, then $\rho\left(V(s_i), V(s_j)\right) > 0$, and the corresponding $Z(s_i)$ and $Z(s_j)$ also has $\rho\left(Z(s_i), Z(s_j)\right) > 0$. It can be concluded therefore, that the process $\{Z(s_i)\}$ has the same range as $\{V(s_i)\}$. But when the variogram of $\{V(s_i)\}$ achieves the sill only asymptotically, then for arbitrary $s_i, s_j, \rho\left(V(s_i), V(s_j)\right) > 0$. As the $d(s_i, s_j)$ increase,

$\rho\left(V(s_i), V(s_j)\right)$ will be close to 0 but still bigger than 0. By the copulas algorithm, for the corresponding $Z(s_i)$ and $Z(s_j)$, $\rho\left(Z(s_i), Z(s_j)\right) > 0$, for arbitrary s_i, s_j also. So the $\{Z(s_i)\}$ also achieves its sill only asymptotically. However, $\{V(s_i)\}$ and $\{Z(s_i)\}$ have different scales, and there is no closed form for the relationship of the covariance functions between $\{V(s_i)\}$ and $\{Z(s_i)\}$. They will also have different practical ranges, but their practical ranges are close, as shown in the simulations section of this paper.

2.2. Description of the Simulation Study

Firstly spatially correlated normal data $\{V(s_i)\}$ with sample size 100 on a regular grid were generated. The grid chosen was 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. Gaussian, exponential and spherical variograms of $\{V(s_i)\}$ were generated. For each variogram type, a sill of 1, nuggets of 0, 1/3 and 2/3, and a practical range of 20 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

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

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.

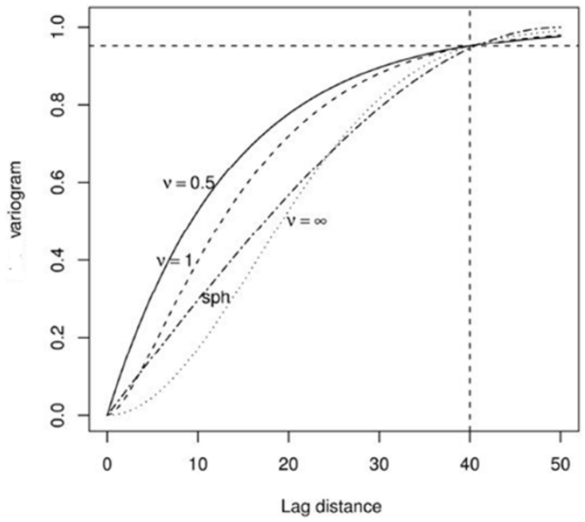


Figure 1. *Variograms for Gaussian, Whittle, exponential and spherical models with nugget $c_0 = 0$, sill $c_0 + \sigma_0^2 = 1$ and practical range 40 indicated by the vertical line. The horizontal line denotes 95% of the sill.*

Secondly spatial binary data $\{Z(s_i)\}$ with non-constant mean $\{p(s_i)\}$ were generated from $\{V(s_i)\}$ by the copulas method introduced in this paper. The definition of $p(s_i)$ is defined as

$$p(s_i) = \exp\left(L(s_i)\right) / \left[1 + \exp\left(L(s_i)\right)\right],$$

$$L(s_i) = -2 + x_1(s_i) \cdot 1,$$

where $x_1(s_i)$ is a random number from a uniform distribution on $[0.5,1.5]$. Thus the mean of the generated $Z(s)$ was around 0.27, since exp(-1)/(1+exp(-1))= 0.27.

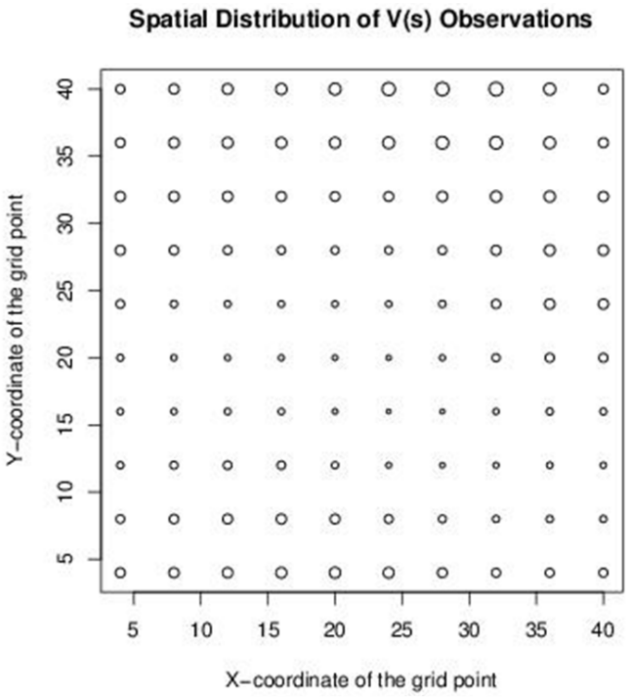
Data were simulated in this paper using SAS software (SAS® 9.2, SAS Institute Inc., Cary, N.C.). The spatial $S(s)$ in the conditional method were generated by the SAS SIM2D Procedure.

3. Results

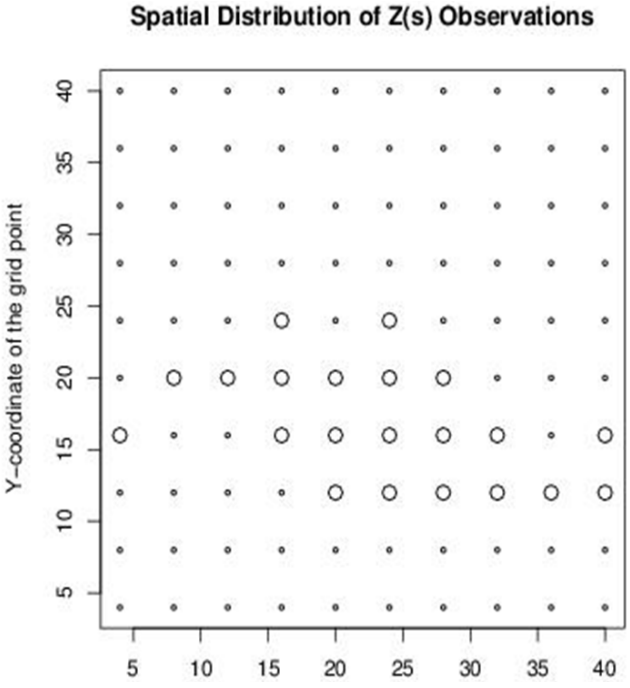
3.1. Simulations of the Data

Here a sample of 100 spatially correlated normal data $\{V(s_i)\}$ were first generated on the regular grid $[0,40] \times [0,40]$ with intervals of 4 in both directions. The Gaussian, exponential and spherical correlation types were considered for $\{V(s_i)\}$. For each variogram type, a sill of 1, nuggets of 0, 1/3 and 2/3, and a practical range of 20 were considered. The $\{V(s_i)\}$ were transformed to uniformly distributed data $\{U(s_i)\}$ by the transformation $U(s_i) = F_{s_i}\left(V(s_i)\right)$, where $F(.)$ was the distribution function of $V(s_i)$. i.e. the Gaussian. A sample of size 100 spatial binary data $\{Z(s_i)\}$ were then generated by the simple transformation, $Z(s_i) = I\{U(s_i) < EZ(s_i)\}$. In this section, the results of the analysis of the spatial binary data generated by copulas method are shown below.

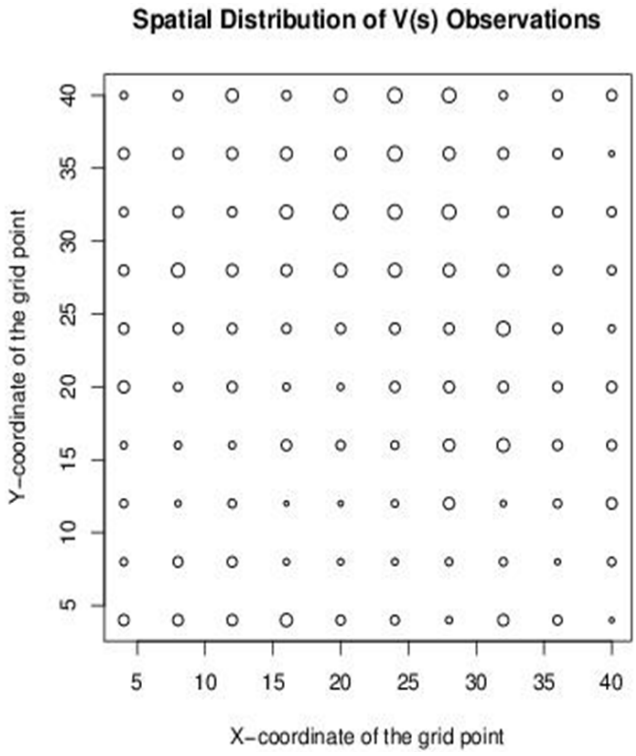
One simulation of a realized dataset of Gaussian random field data $\{V(s_i)\}$ and the corresponding spatial binary data generated by the copulas method are shown in Figure2. Here the sill and nugget of the variogram of the Gaussian random field were chosen to be 1 and 0 respectively. From the plots, it can be seen that the spatial patterns in the generated binary data were different from the spatial patterns in the corresponding Gaussian random field. The $\{V(s_i)\}$ were transformed to uniformly distributed data $\{U(s_i)\}$ by the transformation $U(s_i) = F_{s_i}\left(V(s_i)\right)$, but the $EZ(s_i)$ were determined without regard to $U(s_i)$ and $Z(s_i) = I\{U(s_i) < EZ(s_i)\}$ so it is expected that if U is large Z will be zero and U small Z will be 1. Thus the spatial patterns are the 'inverse' of each other. However, they should have similar variograms. Comparing the spatial patterns in $Z(s)$ generated by different variogram type, 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. 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 had a different spatial pattern from the other two, while the other two had similar spatial patterns.



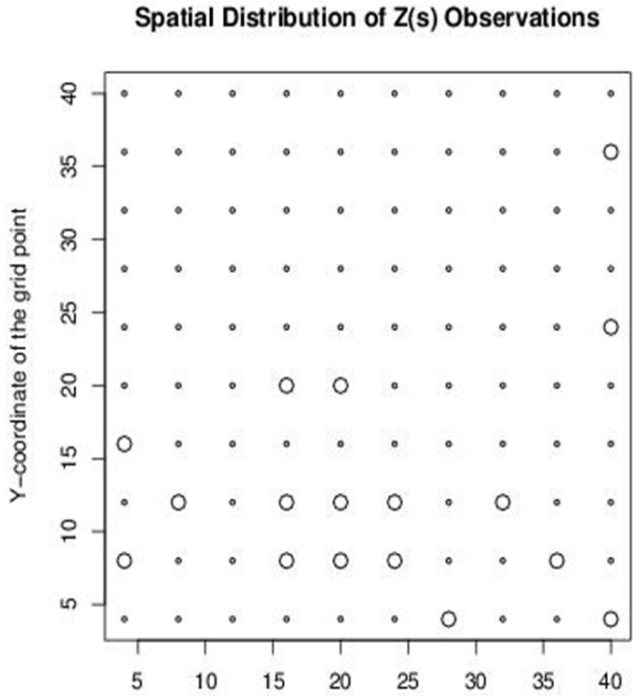
(a) A plot of V(s) with Gaussian variogram



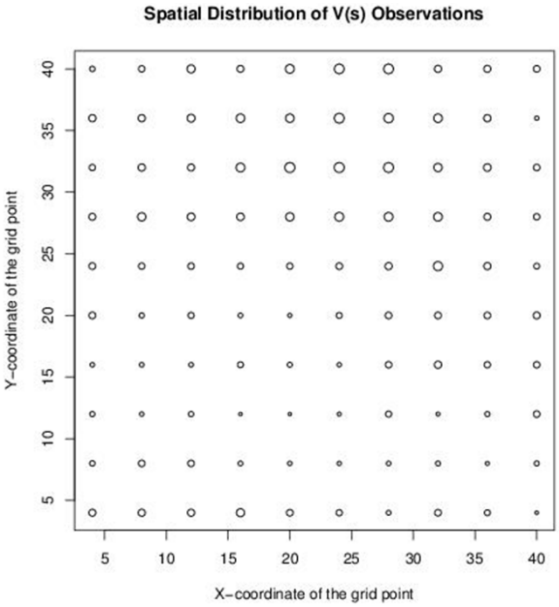
(b) The plot of generated spatial correlated binary data where V(s) has a Gaussian variogram



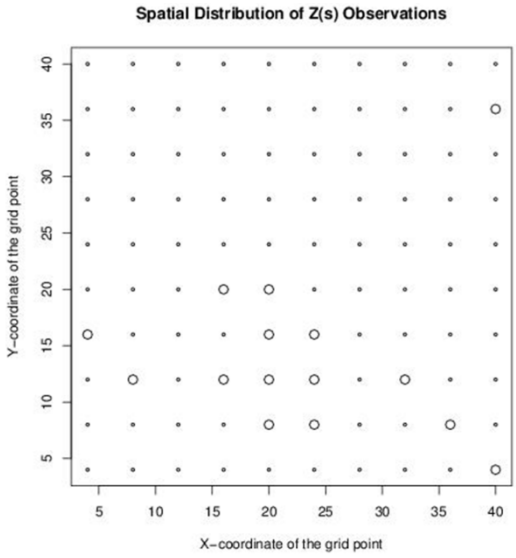
(c) A plot of V(s) with an exponential variogram



(d) The plot of generated spatial correlated binary data where V(s) has an exponential variogram



(e) A plot of V(s) with spherical variogram



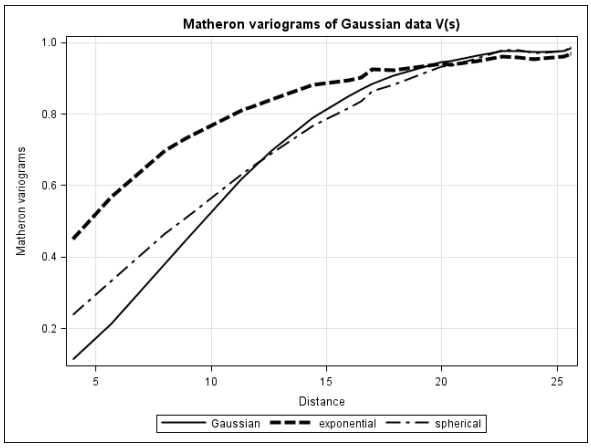
(f) The plot of generated spatial correlated binary data where V(s) has spherical variogram

Figure 2. *The Gaussian random field data* $\{V(s_i)\}$ *with Gaussian, exponential and spherical variograms generated on the grid* $[0,40] \times [0,40]$ *with intervals of 4 in both directions and shown in (a), (c), (e) respectively. In all plots, the size of circles are proportional to their numeric values. The (b), (d), (f) were the corresponding generated spatial binary data* $\boldsymbol{Z}(\boldsymbol{s})$ *by the copulas method. Here the sill and nugget of the variogram of the Gaussian random field data were chosen to be 1 and 0 respectively.*

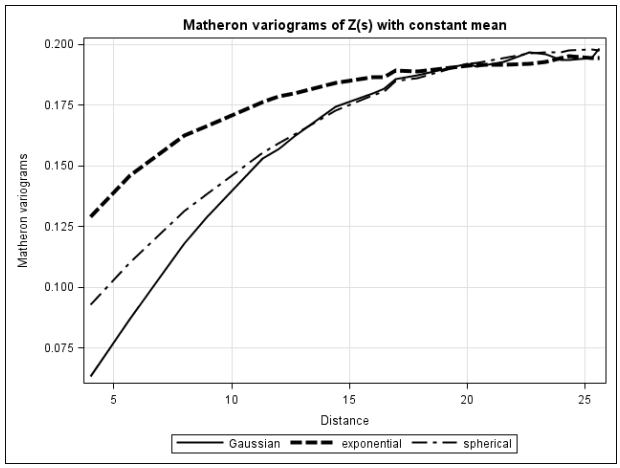
3.2. Variogram Plots

In this section, to examine the relationship of the variogram between the data sets $\{V(s_i)\}$ and $\{Z(s_i)\}$, 500 simulations were done. For the data $\{Z(s_i)\}$, constant and non-constant means were considered. For the non-constant mean case, the means of $\{Z(s_i)\}$ were set as in Method section. For the constant mean case, the mean of $\{Z(s_i)\}$ were simply taken to be 0.27. In each simulation, the Matheron estimators of the variograms of $\{V(s_i)\}$ and $\{Z(s_i)\}$were calculated. In order to accurately estimate the variograms, only lags which had more than 30 data pairs were kept. As the $\{V(s_i)\}$ and $\{Z(s_i)\}$ were generated on a regular grid, there were sufficient lags that had more than 30 data pairs, so the data were not binned. To show the result of 500 simulations, the Matheron estimator at each lag is the average value of Matheron estimators at that lag over 500 simulations.

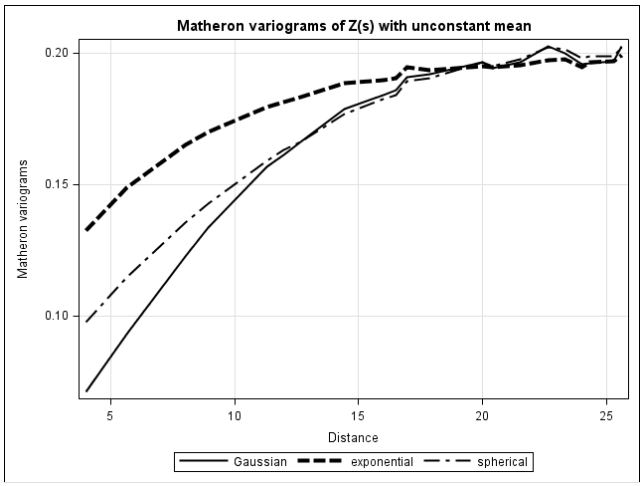
Figure 3 shows the results only for a nugget of 0 in each variogram type of $\{V(s_i)\}$. For the case of nuggets 1/3 and 2/3, similar results to that of Figure 3 were seen. As can be seen in Figure 3, clearly there is spatial association between the $\{Z(s_i)\}$. More importantly, the generated $\{Z(s_i)\}$ were found to have a similar spatial correlation type as the $\{V(s_i)\}$. Note that the variograms of $\{V(s_i)\}$ and $\{Z(s_i)\}$ have similar practical range, which was as expected in Method section. There is no closed form for the connection between the correlation functions of $\{V(s_i)\}$ and $\{Z(s_i)\}$. However, from the plot, it can be concluded that the copulas method kept a similar correlation type. From this same simulation, note that the nugget and sill of the variogram of $\{Z(s_i)\}$ are different from that of $\{V(s_i)\}$. A comparison of the variograms in Figures 3 (b) and (c), shows the two plots are very similar. This may be because the non-constant means of $\{Z(s_i)\}$ had a small variation around 0.27 (the value of the constant mean). However, in the Figure 3 (c) the non-constant mean variogram had a little more fluctuation at the big lags than Figure 3 (b).



(a)



(b)



(c)

Figure 3. *Matheron variogram plots of* $\{V(s_i)\}$ *and* $\{Z(s_i)\}$. *In the three plots, the Matheron estimator at each lag was the average value of Matheron estimators at that lag in 500 simulations. Figures (a), (b) and (c) are the Matheron variogram of the normal data* $\{V(s_i)\}$, *the spatial binary data* $\{Z(s_i)\}$ *with constant mean, and the spatial binary data* $\{Z(s_i)\}$ *with non-constant mean respectively. In Figure (a), the exp, Gaussian and sph denote the* $\{V(s_i)\}$ *generated with exponential, Gaussian and spherical correlation types respectively, while in Figures (b) and (c), exp, Gaussian and sph denote the* $\{Z(s_i)\}$ *generated from the corresponding* $\{V(s_i)\}$ *in Figure (a). This Figure shows results only for a nugget of 0 in each variogram of* $\{V(s_i)\}$.

4. Conclusion

Simulating spatial correlated binary data is very important on many cases, but it is not easily to accomplish, as there are restrictions on the parameters of Bernoulli variables. 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. This paper develops a copulas method to generate spatial correlated binary data. Copulas method is a simulation method, and it is easy to understand and manipulate. This method is wildly used to mathematical experiments, and its procedure is explained in this section. The spatial binary data generated by this method has an inverse spatial pattern comparing with the latent Gaussian random field data, however they have similar empirical variograms. The limitation of this copulas method is that the closed form for the spatial correlation is not available specifically. However, in many applications, the main requirements on simulation is to hold the designed variograms, and from this point the method proposed in this paper is delighted.

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] Jin, R. & Kelly G.E. (2015): Comparison of Sampling Grids, Cut Off Distance and Type of Residuals in Parametric Variogram Estimation. Communications in Statistics - Simulation and Computation, DOI: 10.1080/03610918.2015.1011785

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

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

[10] Nelsen, Roger B. (1999), An Introduction to Copulas, New York: Springer, ISBN 0-387-98623-5

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

[12] 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.

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

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

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

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

[17] 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.

[18] 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.

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

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