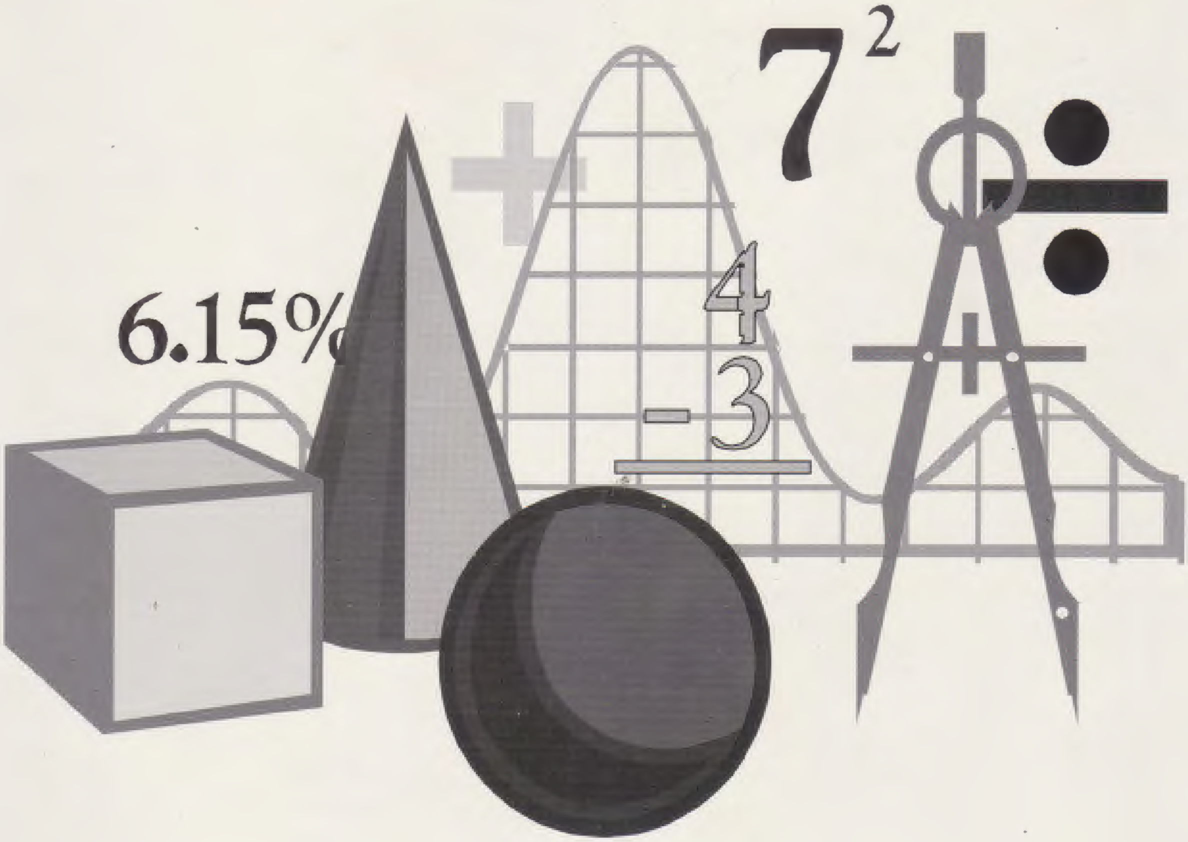
ISSN 1410-1335

Terakreditasi berdasarkan SK No. 52/DIKTI/Kep/2002

INTEGRAL

MAJALAH ILMIAH MATEMATIKA DAN ILMU PENGETAHUAN ALAM

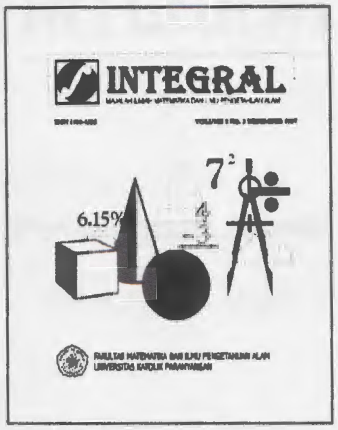
VOLUME 9, NO. 2, JULI 2004



FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS KATOLIK PARAHYANGAN

Integral	Vol. 9	No. 2	Halaman 58 - 89	Bandung Juli 2004	ISSN 1410-1335

daftar isi



INTEGRAL adalah majalah yang bertujuan menyediakan sarana komunikasi ilmiah untuk meningkatkan apresiasi terhadap perkembangan Matematika dan Ilmu Pengetahuan Alam serta penerapannya, khususnya untuk bidang Matematika, Fisika, Ilmu Komputer, dan bidang interdisiplin yang terkait.

INTEGRAL (ISSN 1410-1335)
Terakreditasi berdasarkan SK No. 52/ DIKTI/Kep/2002
diterbitkan tiga kali setahun : Maret, Juli dan November oleh Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Katolik Parahyangan.

Alamat redaksi
Fakultas MIPA
Universitas Katolik Parahyangan
Jl. Ciumbuleuit 94 Bandung 40141
Telp. (022) 204 1964 ext. 707 / ext. 701
Fax . 022 204 2141
E-mail : integral@home.unpar.ac.id
Website: http://home.unpar.ac.id/~integral/

INTEGRAL
MAJALAH ILMIAH MATEMATIKA DAN ILMU PENGETAHUAN ALAM

Editorial

Selamat datang dalam edisi kedua Volume 9 tahun 2004. Dalam edisi ini, INTEGRAL mempersembahkan 4 makalah yang terdiri dari 1 makalah dalam bidang matematika dan 2 makalah dalam bidang fisika dan 1 makalah dalam bidang ilmu komputer. Judul-judul serta abstrak makalah dalam edisi ini, serta makalah-makalah pada edisi sebelumnya juga dapat diakses melalui situs ***http://home.unpar.ac.id/~integral***.

Kontribusi dalam bidang matematika dipersembahkan oleh Gandhi Pawitan (Universitas Katolik Parahyangan – Bandung) dan David Steel (University of Wollonggong, NSW – Australia). Dalam makalah tersebut dibahas tentang penggunaan model spasial untuk membangkitkan proses acak.

Kontribusi dalam bidang fisika diawali oleh Mikrajuddin Abdullah (Institut Teknologi Bandung – Bandung) yang membahas penggunaan metoda *spraying* untuk membuat komposit ZnO dan silica dimana ZnO terperangkap dalam matriks silica. Dilanjutkan dengan kontribusi dari Sylvia Sutanto dan Paulus Tjiang (Universitas Katolik Parahyangan – Bandung) yang menguraikan hubungan kuantisasi pertama dalam fisika kuantum dengan variabel kanonik klasik melalui hubungan transformasi kanonik dengan aljabar Lie.

Kontribusi dalam bidang ilmu komputer dipersembahkan oleh Linda Gunawan (Universitas Katolik Parahyangan) yang mengulas tentang salah satu mekanisme pemetaan model XML ke skema XML, yaitu dengan menggunakan UML Profiles.

Akhirnya, semoga para pembaca memperoleh manfaat dari edisi kita kali ini.

Dewan Redaksi

GENERATING INTER-CORRELATED OBSERVATIONS UNDER A SPECIFIED SPATIAL MODEL

Gandhi Pawitan[1] and David G. Steel[2]
[1] Jurusan Administrasi Niaga, Fakultas Ilmu Sosial dan Ilmu Politik,
Universitas Katolik Parahyangan, Bandung, Indonesia
[2] School of Mathematics and Applied Statistics, University of Wollongong
New South Wales, Australia
e-mail : gandhi_p@home.unpar.ac.id

Abstract

The requirement to generate this random process needs only to define the variance-covariance matrix of the random process. Since the random process is defined in three dimensional space, then we can use a spatial model. One of the spatial model to define the random process is in the form of variogram, which is a function of distance between pairs of observations. The variance-covariance matrix may be determined in relation with two other properties, those are correlogram and covariogram.

*The simulation process was started by generating a random points within a particular shape of region. The locations are uniformly distributed within the region. Lets V is a variance-covariance matrix of the random process **Y[L]**. The random process **Y[L]** may be defined by the semivariogram model $\gamma(d_{ij})$. The d_{ij} is a Cartesian distance between two different individual within domain $\mathcal{D}$ of boundary $\mathcal{B}$. The distribution-based approaches can be applied to generate random observations using Choleski decomposition.*

Keywords : spatial data, random generation, semivariogram, Choleski decomposition

Intisari

Untuk membangkitkan suatu proses acak hanya membutuhkan pendefinisian matrik varian ko-varian dari proses acak tersebut. Selama proses acak tersebut didefinisikan dalam ruang dimensi tiga, maka dapat digunakan model spasial. Salah satu model spasial untuk menjelaskan proses acak adalah dalam bentuk variogram, yaitu sebuah fungsi dari jarak antar pasangan observasi. Matriks varian ko-varian dapat ditentukan berdasarkan dua properti lainnya dari proses acak yaitu korelogram dan kovariogram.

Proses simulasi dimulai dengan membangkitkan titik-titik acak dalam suatu wilayah. Lokasi dari titik tersebut diasumsikan menyebar seragam. Bila ditetapkan bahwa V adalah matrik varian ko-varian dari suatu proses acak Y(L). Proses acak Y(L) tersebut didefinisikan dengan model variogram $\gamma(d_{ij})$, dimana d_{ij} adalah jarak dalam sistem Cartesian dari dua individual dalam *domain wilayah $\mathcal{D}$ dan dalam batasan $\mathcal{B}$. Pendekatan distribusi dapat diterapkan untuk membangkitkan observasi acak melalui dekomposisi Choleski.*

Kata kunci : data spasial, pembangkitan acak, semivariogram, dekomposisi Cholesk

Received : March 10, 2004
Accepted for publication : June 15, 2004

1. Introduction

We may have interest in looking at variability of the observations in the space. In this case spatial variability may be specified as target of inference of the study. One way to observe spatial variability could be done by measuring spatial autocorrelation [1].

Pawitan [2] investigated aggregation bias in variogram analysis, but the limitation is the availability of data at different level, such as individual level and aggregated level data. One possible solution is done by generating individual level data and its aggregated level data.

This paper will discuss a method of generating spatial series data, in such a way it will have a specified inter-dependent among observations according to a variogram model. The method will be based on Choleski decomposition, univariate case will be considered.

2. Literature review

Arbia [3] discussed two methods of simulation of spatial data, that is distribution-based methods and model-based approaches. The distribution-based method will be applied in this paper. The distribution-based approaches generate random observations according to distribution of the process under study, which is defined through a variance-covariance matrix.

Haining, Griffith, and Bennet [4] presented a framework for the generation of surfaces that posses the property of spatial autocorrelation. They stated two objectives of this simulation, the first is to generate spatial data with known, specific, and limited spatial's characteristics. The second is to obtain realization of a spatial process in order to identify properties of the process. Meanwhile, Goodchild [5] proposed an algorithm to generate data to be considered at aggregated level, which took account of the spatial autocorrelation factor.

3. Definition of population

Define $\mathcal{U}$ is a finite population of individual in a particular region $\mathcal{D}$. The boundary of the region $\mathcal{D}$ is defined explicitly. A random process Y is considered with the elements $Y[\ell_i]$, which indicates a characteristic value of an object located at li. Assume the random process Y have the following moment structure $E(Y[\ell_i]) = \mu(li)$ and $Cov(Y_i[\ell_i]; Y_j[\ell_j]) = \Delta_{ij}(\ell_i;\ell_j)$ with $l_i, l_j \in \mathcal{D} \subset R^2$ and $i,j \in \mathcal{U}$. The $Cov(Y_i[\ell_i]; Y_j[\ell_j]) = \Sigma(\ell_i)$ if $i = j$, that is a population variance.

4. Variogram

Variogram is defined by considering two assumptions, those are the intrinsically stationary and second order stationary.

Accepting the first assumption then the intrinsically stationary assumption will follow, that is

$$(i)\quad E(Z(l+d)-Z(l))=0; \qquad \forall \quad l \in \mathcal{D}$$
$$(ii)\quad V(Z(l+d)-Z(l))=2\gamma(d); \qquad \forall \quad l \in \mathcal{D} \qquad (1)$$

where, l indicates a particular location, and d indicates a distance apart from two different location. Second order stationary is defined by following two conditions, $E(\mathbf{Y}[\mathbf{L}]) = \mu$ and $Cov(Y_i\,[\ell_i\,];\ Y_j\,[\ell_j\,]) = C(l_i - l_j\,)$. Where $C(.)$is a covariance function or covariogram, which is defined as a function of distance. The distance may be defined depend on a direction, but an isotropic condition will be considered instead. The isotropic condition stated that $C(\ell_i - \ell_j\,)$ is a function only of $||\ell i - \ell j||$ [6,7]. In this case, the $\ell_i - \ell_j$ is a distance between location l_i and l_j in any directions.

The expression of $2\gamma(d)$ is defined as a variogram. Onehalf of the variogram or $\gamma(d)$ is defined as a semivariogram. Semivariogram is widely used in practice, therefore this term will be used in subsequent discussion. Relation between variogram and covariogram can be defined as

$$C(d) = C(0) - \gamma(d) \qquad (2)$$

Semivariogram is a statistics which can be modeled through a distance function. Cressie [8] presented some models of semivariogram, such as exponential, spherical, Gaussian, linear, and others. For example, the exponential model of the semivariogram is,

$$\gamma(d_{ij}) = \mathsf{n} + (\mathsf{s}-\mathsf{n})(1-exp(\frac{-3d_{ij}}{\mathsf{r}})), \qquad d_{ij} \geq 0 \qquad (3)$$

The theoretical semivariogram models commonly contain three parameters, those are nugget, sill, and range (**n**,**s**,**r**). These parameters are

- nugget effect (**n**) : is the value of semivariogram at distance equal to 0. Theoretically $\gamma(0) = 0$ but in practice $\gamma(0) > 0$.
- sill (**s**) : is the value of semivariogram when the distance is approaching an infinity, that is $\lim_{d_{ij}\to\infty} \gamma\left(d_{ij}\right) = s$. Theoretically it will equal to σ^2 or $C(0)$.
- range (**r**) : is a distance d_0 for $d_0 > 0$ such that the value of $\gamma(d_0)$ is turning closely to **s** or $C(0)$. This distance indicates a situation of $\rho(d) = 0$, that is when individuals are independent with each other.

5. Generating observations

The framework is developed in general but for a practical reason the exponential semivariogram model of equation (3) will be considered.

The process is initiated by generating a uniform individual locations in two dimensional Eucledian space, $\mathcal{D}$, within the boundary $\mathcal{B}$. All pairs of distance are calculated, and then the variance-covariance matrix is computed by relation (2). Then applying Choleski decomposition into the variance-covaraince matrix to generate the individual data values.

5.1 Generating uniform individual locations

The individual locations are randomly generated following a uniform distribution within the range of [X-min;X-max] and [Y-min;Y-max]. One realization is shown in Figure (1). The points are generated between [20,90] in x axes and between [10,70] in y axes.

5.2 Choleski Decomposition

Some of decomposition algorithm may destroy the symmetry of the symmetric matrix, such that the original matrix is broken down into **L** and **U** matrix components. The **L** component indicates the lower triangular matrix, and the **U** indicates the upper triangular matrix. Hence it calls as LU decomposition.

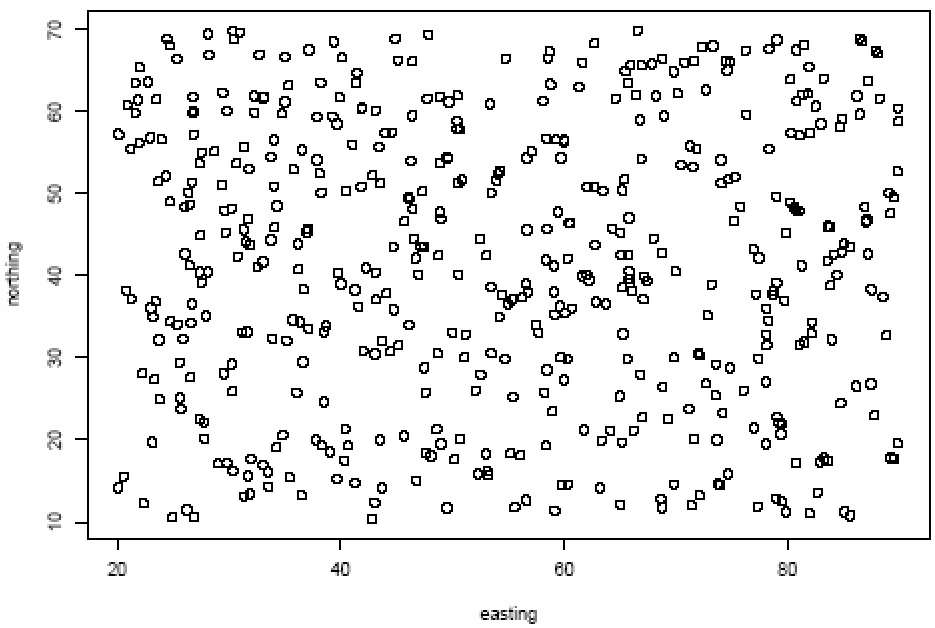


Figure 1: Random individual location

The Choleski decomposition may preserved the symmetry of the symmetric matrix. This decomposition is derived from the LU decomposition. Let consider **A** is a symmetric matrix of dimension $n \times n$, then the **A** can be decomposed into,

$$\mathbf{A} = \mathbf{L}\mathbf{D}\mathbf{L}^{\mathsf{T}} \quad (4)$$

where **A** is a symmetric matrix and **D** is a diagonal matrix.

If the matrix **A** is a symmetric and also positive definite then the elements of the matrix **D** are positive. Hence $\sqrt{\mathbf{D}}$ could be defined as the matrix with each elements are the square root of the elements of matrix **D**. Then we may have $\tilde{\mathbf{L}} = \mathbf{L}\sqrt{\mathbf{D}}$.

That will give,

$$\mathbf{A} = \mathbf{L}\sqrt{\mathbf{D}}\sqrt{\mathbf{D}}\mathbf{L}^{\mathsf{T}} = \tilde{\mathbf{L}}\tilde{\mathbf{L}}^{\mathsf{T}} \quad (5)$$

The decomposition of matrix **A** into $\tilde{\mathbf{L}}\tilde{\mathbf{L}}^{\mathsf{T}}$ is attributed to Choleski decomposition. The general formulae for Choleski decomposition could be defined as follow, [9]

$$l_{ii} = \left(a_{ii} - \sum_{k=1}^{i-1} l_{ik}^2 \right)^{\frac{1}{2}}$$

$$l_{ij} = \frac{a_{ij} - \sum_{k=1}^{j-1} l_{ik} l_{jk}}{l_{jj}} \quad ;\text{for } j < i \quad (6)$$

where l and a are the element of matrix **L** and **A**, respectively.

5.3 Generating data values

Lets consider that **V** is a matrix of

variance-covariance of the individuals data values, **Z**. If **Z** is a vector of individuals data values with dimension $n \times 1$ then the **V** will have $n \times n$ dimension.

Arbia [3] noted a method of generating such individuals data, that is applying Choleski decomposition procedure. This method is called as distribution-base approach. Using relation (5), we get the **L** from the **V**, that is $\mathbf{V} = \tilde{\mathbf{L}}\tilde{\mathbf{L}}^{\mathsf{T}}$.

Lets assume that **e** is a vector of independent identically distributed of standard normal (N(0,1)) random variable. Then the random process **Z** may be generated from **e** by using relation (7).

$$\mathbf{Z} = \tilde{\mathbf{L}}\mathbf{e} \tag{7}$$

5.4 Implementation

Let defines a random process Z_l, where l is defined in $\mathcal{D}$, which is a subset of Euclidean space, $\mathcal{R}^2$. Let assumes that the individual locations are distributed uniformly within the region $\mathcal{D}$. And consider the region $\mathcal{D}$ is a planar region with rectangle shape. The boundaries of the region $\mathcal{D}$ are defined within the following range values, [20,80] easting and [10,90] northing. The individuals locations are generated and distances (d_{ij}) of all pairs of points can be calculated. Let considers the exponential semivariogram model with the following parameters, **n** = 5, **s** = 20, and **r** = 15. The variance-covariance matrix, **V**, can be defined by applying equation (2) into calculated distances based on a specified semivariogram model. The matrix **V** is defined explicitly by the following,

$$\begin{pmatrix} C(0) & C(d_{12}) & \cdots & C(d_{1j}) & \cdots & C(d_{1N}) \\ C(d_{21}) & C(0) & \cdots & \cdots & \cdots & C(d_{2N}) \\ \vdots & \cdots & C(0) & C(d_{ij}) & \cdots & \vdots \\ C(d_{j1}) & \cdots & C(d_{ji}) & C(0) & \cdots & C(d_{jN}) \\ \vdots & \cdots & \cdots & \cdots & C(0) & \vdots \\ C(d_{N1}) & \cdots & \cdots & C(d_{Nj}) & \cdots & C(0) \end{pmatrix} \tag{8}$$

where the diagonals' elements are equal to C(0) or equal to the sill (**s**). Meanwhile, the other elements are defined by equation (2). The $\tilde{\mathbf{L}}$ matrix is derived from the **V** by the Choleski decomposition. The vector **e** of standard normal random variable of size N is generated. Therefore, the individual data values are generated by applying equation (7). Result of the simulation is shown on the following Figure (2), with 1500 numbers of observations,

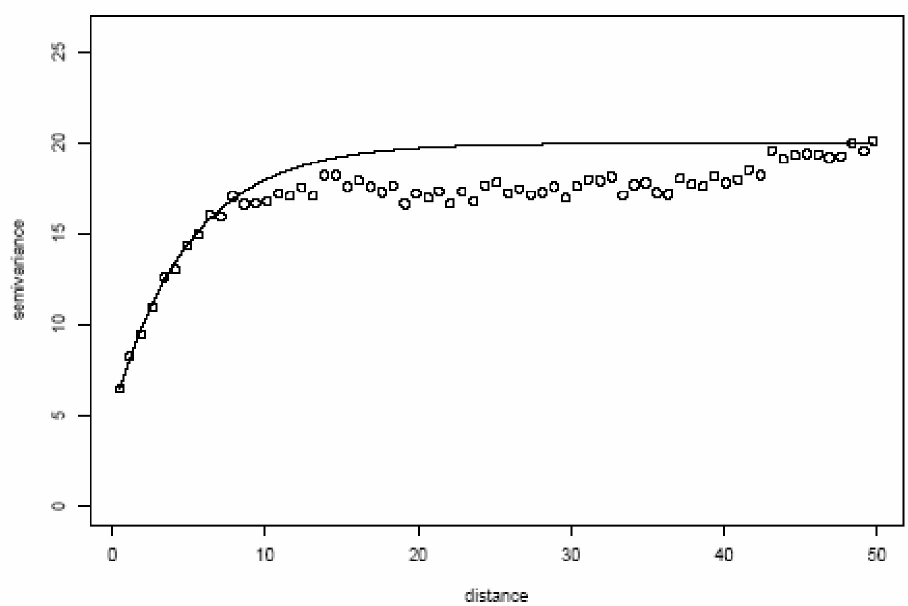


Figure 2:
Result of the simulation. The solid line is a theoretical semivariogram and the points are a simulation result.

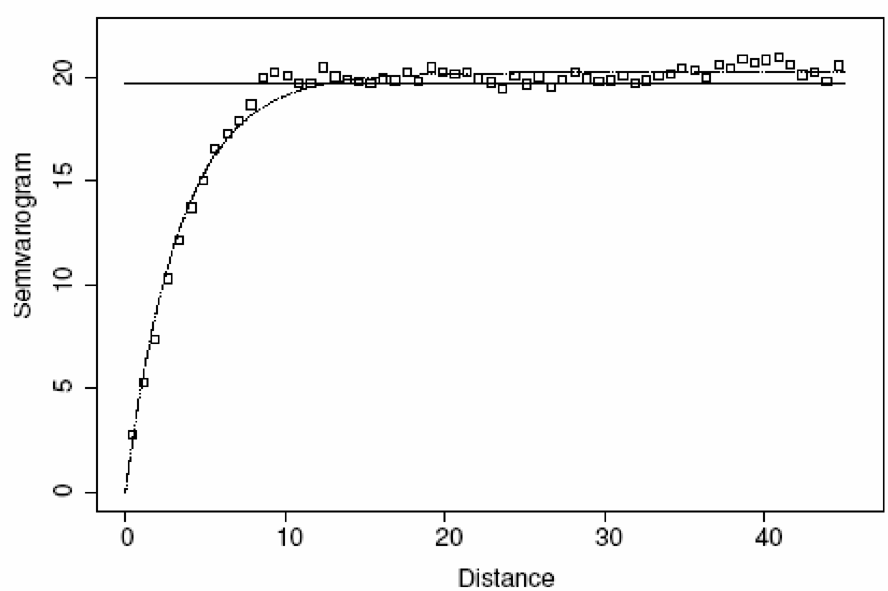


Figure 3:
One simulation result of the exponential semivariogram model (**n** = 0, **s** = 20, **r** = 10).

6. Application

Lets consider exponential semivariogram model, with the following parameters **n** = 0, **s** = 20, and **r** = 10. Hence the **V** matrix is defined as $C(0) = \mathbf{s} = 20$ and $C(d_{ij}) = 20 - \gamma(d_{ij})$.

Lets consider that *li* is uniformly distributed over the domain region $\mathcal{D}$. The region $\mathcal{D}$ is defined as rectangular with (20,10) lower left coordinates and (90,80) upper right coordinates. The simulation generate 1500 numbers of observations, and the result is in Figure (3).

The solid line in Figure (3) indicates a theoretical semivariogram of individual level data. The square dots ( ) indicate empirical semivariogram. The solid line starts from zero (nugget = 0), going up as a distance increase, and then turn into a constant line at distance ± 10 (range = 10) when the semivariogram value is around 20 (sill = 20). The horizontal straight line indicates variance of the individual data.

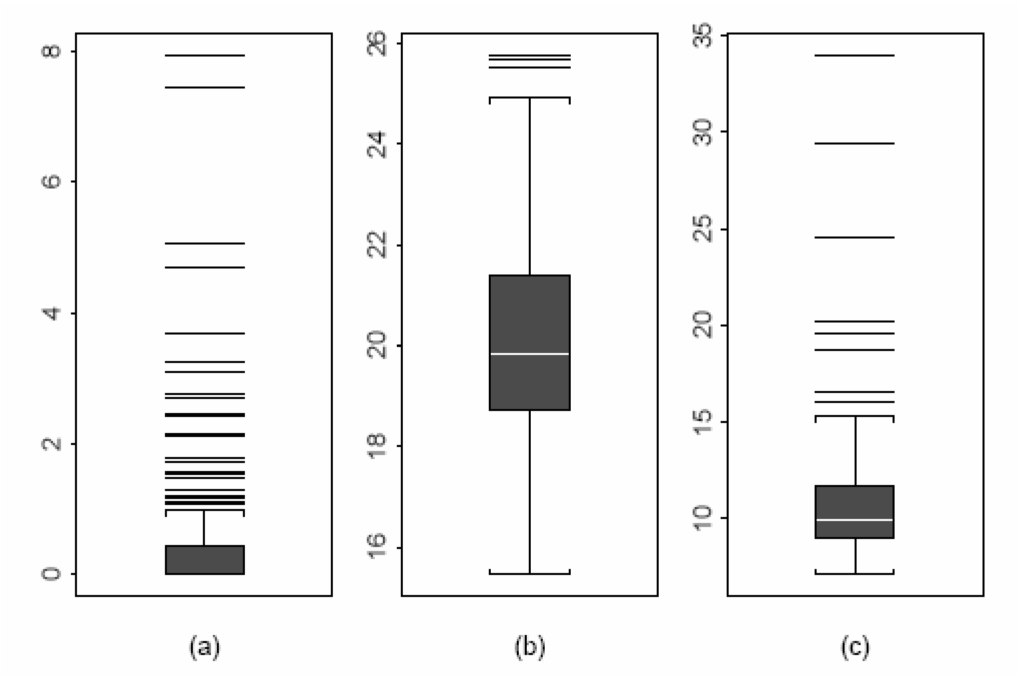


Figure 4:
Distribution of parameters; (a) nugget, (b) sill, and (c) range, from 200 simulations

Figure (4) shows distribution of the estimated parameters of the model from 200 simulations. The distribution exhibit that the generated value lay within the expected value (**n** = 0, **s** = 20, and **r** = 10). But there are some outlier in estimating the parameters of the model, since a divergence in iteration of non-linear least square procedure, which is resulted a very big values. One reason was identified as an initial value problem in non-linear estimation procedure.

7. Conclusion

The simulation process was started by generating a random points within a particular shape of region. The points locations are uniformly distributed within the region. The distribution based methods under a specified spatial model is an effective approaches to generate inter-dependent observations. The variogram can be used to specify inter-dependency observations spatially. Although in this paper is used the exponential model, other models of variogram can be applied in the same manner. There are some notes in the estimated parameters of the models from the generated data.

The nugget distribution shows some very large value of the nugget estimator. This is caused by non convergence estimator of the fitting procedure. And in general, the simulation could show the true parameter value of the nugget. The range distribution may have a better result in term of its predefined model. It seems that the range estimator is approaching the true parameter value, that is 10. It is caused by the non convergence estimator of the fitting procedure.

8. References

1. Cliff, A. D., & Ord, J. K., "*Spatial Processes Models and Applications*". London: Pion, 1981.
2. Pawitan, G., "*Analysis of Aggregated Spatial Social Data*". Unpublished Ph.D dissertation, School of Mathematics and Applied Statistics, University of Wollongong, 2001.
3. Arbia, G., "*Spatial Data Configuration in Statistical Analysis of Regional Economic and Related Problems*". Dordrecht: Kluwer, 1989.
4. Haining, R., Griffith, D. A., & Bennet, R., "*Simulating two-dimensional auto-correlated surfaces*". Geographical analysis, **15**(3), 247–255, 1983.
5. Goodchild, M. F., Algorithm 9: "*Simulation of autocorrelation for aggregated data*". Environment and Planning **A**, *12*, 1073–81, 1980.
6. Christensen, R., "*Linear Models for Multivariate, Time Series, and Spatial Data*". New York: Springer-Verlag, 1980.
7. Grondona, M. O., & Cressie, N., "*Using Spatial Considerations in the Analysis of Experiments*". Technometrics, **33**(4), 381–392, 1991.
8. Cressie, N., "*Spatial Statistics*". New York: Jon Wiley & Sons, 1991.
9. Jennings, A., "*Matrix computation for engineers and scientists*". New York: John Wiley & Sons, 1977.