Geo-spatial Information Science 11(4): 243-246
DOI 10.1007/s11806-008-0129-1

Volume 11, Issue 4
December 2008

Article ID: 1009-5020(2008)04-243-04

Document code: A

Indicator and Multivariate Geostatistics for Spatial Prediction

ZHANG Jingxiong YAO Na

Abstract There are various occasions where simple, ordinary, and universal kriging techniques may find themselves incapable of performing spatial prediction directly or efficiently. One type of application concerns quantification of cumulative distribution function (CDF) or probability of occurrences of categorical variables over space. The other is related to optimal use of co-variation inherent to multiple regionalized variables as well as spatial correlation in spatial prediction. This paper extends geostatistics from the realm of kriging with uni-variate and continuous regionalized variables to the territory of indicator and multivariate kriging, where it is of ultimate importance to perform non-parametric estimation of probability distributions and spatial prediction based on co-regionalization and multiple data sources, respectively.

Keywords auto- and cross-covariance; indicator kriging; co-kriging; data support; block

CLC number P208

Introduction

In various environmental studies, it is often necessary to evaluate probabilistically the risk of contamination exceeding certain thresholds[1-3]. Such events can be conceived of as indicators of success or failure, positive or negative when tested against some criteria. Indicator geostatistics, particularly indicator kriging, offers techniques for prediction concerning occurrences of indicators of discrete nature, which makes no assumption of normality and is essentially a non-parametric counterpart to kriging with continuous variables. Instead of assuming a normal distribution at each estimate location, indicator kriging builds the cumulative distribution function (CDF) at each point based on the behavior and correlation structure of indicator transformed data points in the neighborhood. To achieve this, indicator kriging needs a series of threshold values between the smallest and largest data values in the set. These threshold values, referred to here as *cutoffs,* are used to numerically build the CDF at the estimation point. For each cutoff, data in the neighborhood are transformed into 0s and 1s: 0s if the data are greater than the threshold, and 1s if they are less. The probability that the estimation point is less than the threshold value, given this neighborhood of transformed data and a model of the cutoff correlation structure, is estimated by indicator kriging. Performing this operation for each cutoff across the range of data approximates the CDF at the estimation point. After the CDF is built, it must be post processed to produce probability maps and *E*-Type values for estimation maps and risk maps.

In spatial applications, there exist various sources of indirect information (*secondary data*) about the *primary variable* to be estimated or simulated. For instance, topography indirectly determines precipita-

Received on August 10, 2008.

Supported by the National 973 Program of China (No. 2007CB714402-5).

ZHANG Jingxiong, School of Remote Sensing and Information Engineering, Wuhan University,129 Luoyu Road, Wuhan 430079, China.

E-mail: jxzhang@whu.edu.cn

tion. The traditional, non-geostatistical approach to using multiple attributes for prediction is to determine a possibly non-linear regression relation between the primary variable Z_1 and any set of secondary attributes, $Z_i(i=2,\cdots,B.)$ A function for the relationship defined by $Z_1=\varphi(Z_2,Z_3,\cdots,Z_B)$ can be calibrated, for example, to model porosity from any set of co-located seismic or well-log attributes, which can then predict $Z_1(x)$ anywhere in the reservoir where $Z_2,Z_3,\cdots,Z_B$ data are available at *x*. This procedure ignores the spatial variability of the original $Z_1(x)$ variable, and does not provide a measure of spatial uncertainty about the unknown $Z_1(x)$ values. Multivariate geostatistics provides a sound basis for modeling and predicting regionalized variables in the light of all covariates, which are themselves regionalized variables.

This paper will first discuss indicator geostatistics for non-parametric quantification of CDF for originally continuous variables, which may also be used for prediction of probabilistic occurrences of discretized interval classes or originally categorical variables. This is followed by a treatment of multivariate geostatistics, which is advantageous for combined use of multi-source data in spatial prediction, especially when the primary variable is sparsely sampled as opposed to the secondary variable that is densely sampled. Lastly, the conclusion section wraps up the paper with some remarks as to topics for further investigation.

1 Indicator geostatistics

For regionalized variables *C*(*x*) of originally discrete nature, it is possible to define the indicator transform:

$$I(x,k)=\begin{cases}1, & \text{if } C(x)=k\\ 0, & \text{else}\end{cases} \tag{1}$$

where *k* is the class label taking values in the set $\{1,2,\cdots,K\}$, with *K* being the total number of classes.

Indicator random variables may also be defined for originally continuous random variables, as briefly mentioned in the first paper of this sequel. For various thresholds $z_k,k=1,2,\cdots,K,$ one defines an indicator variable below:

$$I(x,z_k)=\begin{cases}1, & \text{if } Z(x)\leqslant z_k\\ 0, & \text{else}\end{cases} \tag{2}$$

Each of the *K* indicator variables allows us then to focus on various ranges of the variable. Taking z_k equal to the median allows quantifying the connectivity of average *z*-values, while taking z_k equal to some extreme quantiles allows quantifying connectivity of extremes. A spatial measure of connectivity between any two points in space separated by a lag-distance *h* is now the indicator variogram,

$$\gamma_I(h,z_k)=\text{var}(I(x,z_k)-I(x+h,z_k)) \tag{3}$$

which can be modeled from the sample data by transforming each sample into a vector of size *K* of indicator data (zeros and ones)[4].

Once the *K* indicator variograms are determined, one can perform indicator kriging to determine local probability models that do not rely on any Gaussian assumption. Indeed, in indicator method, one relies on the following simple rule:

$$\begin{aligned}&E[I(x,z_k)]\\&=\Pr(I(x,z_k)=1)\times 1+\Pr(I(x,z_k)=0)\times 0\\&=\Pr(I(x,z_k)=1)=\Pr(Z(x)\leqslant z)\end{aligned} \tag{4}$$

Hence, any estimation of $I(x,z_k)$ is also an estimation of $\Pr(Z(x)\leqslant z)$, which means that kriging of an indicator variable is nothing more than determining the local uncertainty about the original *Z*-variable:

$$E[I(x,z_k)\,|\,(n)]=\Pr(Z(x)\leqslant z\,|\,(n)) \tag{5}$$

For example, to determine the probability of porosity at un-sampled location *x* to be above z_k, given the local well data, one only needs to krige (estimate) the indicator variable using indicator kriging, then that kriged value is also a probability. The indicator kriging estimate is written as:

$$\begin{aligned}&[I(x,z_k)]^*-p\\&=\sum_{\alpha=1}^{n}\lambda_\alpha(x)(i(x_\alpha,z_k)-p)\\&=[\Pr(Z(x)\leqslant z_k\,|\,(n))]^*\end{aligned} \tag{6}$$

where *p* is the estimated proportion of data above the cutoff z_k. The indicator kriging weights are determined by solving the indicator kriging equations:

$$\begin{aligned}&\sum_{\beta=1}^{n}\lambda_\alpha(x)\,\text{cov}(I(x_\alpha,z_k),I(x_\beta,z_k))\\&=\text{cov}(I(x_\alpha,z_k),I(x,z_k))\qquad\forall\alpha=1,\cdots,n\end{aligned} \tag{7}$$

By estimating $\Pr(Z(x)\leqslant z_k\,|\,(n))$ for various

thresholds z_k, indicator kriging provides the uncertainty about the property at an un-sampled location *x*, given any sample data (*n*). Note that no Gaussian or any parametric distribution model is assumed, the uncertainty is quantified through a series of cumulative probabilities at given thresholds z_k.

2 Multivariate geostatistics

Geostatistics allows various approaches for integrating secondary variable in spatial prediction. A simple but often effective way to introduce secondary data in either kriging or simulation algorithms is using secondary information as a trend. In *kriging with locally varying mean*, the primary variable is decomposed into a mean and residual component

$$Z_1(x) = m(x) + R(x) \tag{8}$$

Since secondary data is often smooth in nature, the mean-component can be used to model the trend

$$E[Z_1(x)] = m(x) = \varphi(z_2(x)) \tag{9}$$

where the function needs to be calibrated from data. This requires that the secondary data is available everywhere and $Z_1(x)$ needs to be determined.

In kriging with locally varying mean, the spatial variability of the soft data is largely neglected. Often, the secondary data has its own particular spatial continuity, or shows distinct patterns of spatial correlation with the primary variable. In such a case, one would like to use the full spatial correlation of the secondary data expressed in the correlogram $\rho_{Z_2}(h)$ and the spatial correlation between secondary and primary expressed through the cross-covariance

$$\rho_{Z_1Z_2}(h) = \frac{\mathrm{cov}(Z_1(x), Z_2(x+h))}{\sqrt{\mathrm{var}(Z_1)\mathrm{var}(Z_2)}} \tag{10}$$

Consider a regionalization characterized by a set of *K* spatially intercorrelated random variables $\{Z_k(x), k = 1 \text{ to } K\}$. The first- and second-order moments of these variables, assuming stationarity, are

$$E(Z_k(x)) = m_k \quad \forall x \tag{11}$$

$$\mathrm{cov}_{kk'}(h) = E((Z_k(x_{s1}) - m_k)(Z_{k'}(x_{s2}) - m_{k'})) \quad \text{for } x_{s1} - x_{s2} = h \tag{12}$$

Many spatial data, such as reflectance and land cover, are actually defined and measured on supports of finite size, such as pixels and parcels, which are assumed to be generated from finite-support random variables that originate from point-support variables *Z*(*x*). Denote such a block-support variable $Z(v_x), v_x$ standing for a finite support centered at *x*. This notation applies also to point-support variables, when a block v_x is reduced to a point *x*. Thus, block-support variables and their corresponding data are used as a general setting in the following.

It is usually assumed that the block value (the coarse scale information) is a linear average of all the fine scale values within that block. Therefore, the data on *N* block supports, i.e. $s\{z(v_\alpha), \alpha = 1, \cdots, N_v\}$, can be considered as integrals of point support values within their respective supports.

Supposing it is required to estimate the arithmetic average of variable Z_k over a block $v_k(x)$ centered at location $x(v_k(x)$ can also be shortly denoted as v_x without causing confusion), which is discretized into an array of n_v points, as:

$$z_k(v_x) = \frac{1}{n_v}\sum_{k_v=1}^{n_v} z_k(x_{k_v}) \tag{13}$$

The available data $\{z_k(x_{k_s}), k_s = 1, 2, \cdots, n_k; k = 1, 2, \cdots, K\}$ are defined on the support $\{v_k(x_{k_s}), k = 1, 2, \cdots, K\}$, which may be points or areas centered at points $\{x_{k_s}, k_s = 1, 2, \cdots, n_k\}$ [5]. The estimation of Z_j over a block $v_i(x)$ centered at location *x* is provided by a linear combination:

$$\hat{z}_j(v_j(x)) = m_j + \sum_{k=1}^{K}\sum_{k_s=1}^{n_k} \lambda_{k_s}(z_k(v_{k_s}) - m_k) \tag{14}$$

The weights in Eq.14 are derived by the solution of the following system of simple co-kriging:

$$\begin{aligned} &\sum_{k'=1}^{K}\sum_{k'_s=1}^{n_{k'}} \lambda_{k'_s}\bar{c}_{k'k}(v_{k'}(x_{k'_s}), v_k(x_{k_s})) \\ &= \bar{c}_{jk}(v_j(x), v_k(x_{k_s})) \\ &\forall k_s = 1, 2, \cdots, n_k, k = 1, 2, \cdots, K \end{aligned} \tag{15}$$

where $\bar{c}_{k'k}$ and $\bar{c}_{jk}$ stand for the average covariance between pairs of locations within data supports $v_{k'}$ and v_k, and the average covariance between locations discretizing the support to be estimated (i.e., $v_j(x)$) and those within data support v_k, respectively. For instance, $\bar{c}_{k'k}$ is calculated by:

$$\bar{c}_{k'k}(v_{k'}(x_{k'_S}), v_k(x_{k_S}))$$

$$= \frac{1}{Nv_{k'} Nv_k} \sum_{k'_v=1}^{Nv_{k'}} \sum_{k_v=1}^{Nv_k} \mathrm{cov}_{k'k}(x_{k'_v} - x_{k_v}) \qquad (16)$$

where $x_{k'_v}$ and x_{k_v} denote locations falling within data support $v_{k'}(x_{k'_s})$ and $v_k(x_{k_s})$, respectively.

In order to develop ordinary co-kriging equations (), denote k_0 to be the descriptor for the variable being analyzed, $k_0 \in \{1,2,\cdots,K\}$ with K being the total number of regionalized variables operating in the problem domain. It is possible to rewrite the simple co-kriging estimator for $Z_{k_0}(v_x)$ over a block v_x as:

$$\begin{aligned} \hat{z}_{k_0}(v_{k_0}(x)) &= \sum_{k=1}^{K} \sum_{\alpha_k=1}^{n_k} \lambda_{\alpha_k} z_{\alpha_k} \\ &+ m_{k_0}(1 - \sum_{k_0=1}^{n_{k_0}} \lambda_{k_0}) - \sum_{k \neq k_0} (\sum_{k_s=1}^{n_k} \lambda_{k_s}) m_k \end{aligned} \qquad (17)$$

In order to cancel out the effects of unknown means of co-regionalized variables, it is necessary to impose the following constraints:

$$\begin{cases} \sum_{\alpha_{k_0}=1}^{n_{k_0}} \lambda_{\alpha_{k_0}} = 1 \\ \sum_{\alpha_k=1}^{n_k} \lambda_{\alpha_k} = 0 \quad \forall \quad k \neq k_0 \end{cases} \qquad (18)$$

This leads to the so-called ordinary co-kriging estimator below:

$$\hat{Z}_{k_0}(v_x) = \sum_{k=1}^{K} \sum_{\alpha_k=1}^{n_k} \lambda_{\alpha_k} z_{\alpha_k} \qquad (19)$$

where $z_{\alpha_k}(\alpha_k = 1,2,\cdots,n_k)$ stands for data defined on the set of supports $\{v_{\alpha_k}\}$, while λ_{α_k} for the weight assigned to v_{α_k}, which is obtained by solving the following equations:

$$\begin{aligned} &\sum_{k'=1}^{K} \sum_{\beta k'=1}^{n_{k'}} \lambda_{\beta_{k'}} \overline{C}_{k'k}(v_{\beta_{k'}}, v_{\alpha_k}) - \mu_k = \overline{C}_{k_0 k}(V_{k_0}, v_{\alpha_k}) \\ &\qquad \forall \quad \alpha_k = 1,2,\cdots,n_k \quad k = 1,2,\cdots,K \\ &\sum_{\alpha_{k_0}=1}^{n_{k_0}} \lambda_{\alpha_{k_0}} = 1 \\ &\sum_{\alpha_k=1}^{n_k} \lambda_{\alpha_k} = 0 \quad \forall \quad k \neq k_0 \end{aligned} \qquad (20)$$

with co-kriging variance evaluated as:

$$\begin{aligned} \sigma^2_{V_{k_0}} &= \overline{C}_{k_0 k_0}(V_{k_0}, V_{k_0}) + \mu_{k_0} \\ &- \sum_{k=1}^{K} \sum_{\alpha_k=1}^{n_k} \boldsymbol{\lambda}_{\alpha_k} \overline{C}_{k_0 k}(V_{k_0}, v_{\alpha_k}) \end{aligned} \qquad (21)$$

3 Conclusion

This paper has discussed indicator geostatistics as providing non-parametric solution to probabilistic mapping and CDF modeling techniques for categorical variables and continuous variables, respectively. Indicator geostatistics comes with costs. This concerns the considerable difficulty in generating indicator variograms for all indicator cut-off levels. Conditional simulation may be better pursued for probabilistic mapping, as will be discussed in the third paper of this sequel[1].

The other major task of this paper is about how multivariate geostatistics may be pursued for the solution to spatial prediction with co-regionalized variables[6]. Other than simple and ordinary co-kriging, other implementations of co-kriging should also be dealt with, such as co-kriging that can handle non-stationarity in co-regionalization. Only then, there will be a systematic treatment for co-kriging, just like simple, ordinary, and universal variants have been discussed in the first paper of this sequel.

References

[1] Journel A G, Huijbregts C (1978) Mining geostatistics [M]. Boston:Academic Press

[2] Journel A G (1989) Fundamentals of geostatistics in five lessons[C]. Short Course in Geology, American Geophysical Union, Washington D.C.

[3] Goovaerts P (1997) Geostatistics for natural resources estimation[M]. Oxford: Oxford University Press

[4] Rossi R E, Mulla D J, Journel A G, et al. (1992) Geostatistical tools for modeling and interpreting ecological spatial dependence[J]. *Ecological Monographs*, 62(2): 277-314

[5] Kyriakidis P C (2004) A geostatistical framework for area-to-point spatial interpolation[J]. *Geographical Analysis* 36(3): 259-289

[6] Zhou Yueqin, Stein A, Molenaa M (2003) Integrating interferometric SAR data with leveling measurements of land subsidence using geostatistics[J]. *International Journal of Remote Sensing*, 24 (18): 3 547-3 563