

# RANDOM FIELDS AND (MARKED) POINT PROCESSES: A PRACTICAL COMPARISON OF TWO STOCHASTIC MODELS

Arie CROITORU and Yerach DOYTSHER

Technion – Israel Institute of Technology  
Faculty of Civil Engineering, Division of Geodetic Engineering  
Technion City, Haifa 32000, Israel

ariec@tx.technion.ac.il, doytsher@geodesy.technion.ac.il

**Key words:** vector data quality, random field, marked point process

## ABSTARCT

Reliable detailed information on the positional accuracy of spatial data is essential to end-users. In order to accommodate this two essential tools, namely reporting tools (“Metadata”) and modeling tools (a mathematical framework describing the positional accuracy) have been developed in recent years. One such modeling tool is the random field stochastic model, in which the discrepancies between two data sets are treated as a random process (a "signal") over a two or three-dimensional domain. The characteristics of a random field may be attained using well-known geostatistical estimators, such as the variogram and the correlelogram. The foremost advantage of this approach is that the errors are considered as a spatial phenomenon, in which correlations are accounted for. This contribution discusses the link between the random field model and two other stochastic tools, Least Squares Collocation (LSC), and point process analysis. It is argued that this linkage is required due to the inability of the random field model to carry out signal filtering or prediction, and its lack of sensitivity to the spatial distribution of the data points in the domain that is being evaluated. Each of these tools is described, and its usage in the context of spatial data is discussed.

## INTRODUCTION

Recent advances in information technology as well as in spatial data capturing and processing techniques provide a wealth of spatial information. Such information can be easily and quickly downloaded from various data providers via the internet or other broad-band communication facilities, and its usage is almost immediate as it can be downloaded in a variety of commonly used formats. This so-called "plug-and-play" approach is new to spatial data providers, as well as to data end users ("users"). Unlike the situation where data is collected by an organization according to its particular specifications and for its exclusive usage, data is no longer “tailor-made” and direct communication between the data provider and the user is no longer assured. This may result in data misuse, forcing the user to check the applicability of the data. Data providers may also be affected by this new situation, as there is no guaranty of proper usage of the data, and as the proficiency of the user is unknown. This may expose data providers to undesirable liability and other legal complications.

It is possible to resolve some of these difficulties by providing additional descriptive information regarding the data set at hand. Such additional information may be delivered to the user by adding *metadata* to the data set, in which various aspects of the data (such as data quality information, spatial reference information, or temporal information) are detailed. Yet for many applications the available

metadata is not detailed enough, and its implications for a specific usage are unclear (for example, DTM based applications). Consequently, users may still face uncertainty regarding the applicability of the data, even where metadata is available.

While the metadata approach leaves the user passive, users can take an "active" approach by assessing the "fitness for use" of data according to their specific needs (Agumaya and Hunter, 1999). Agumaya and Hunter (1999) suggest that this assessment should be based on an estimation of the uncertainty in the end product, caused by errors and uncertainties in a given data set. This end-product uncertainty can then be used for assessing the total *risk* the user faces. Although a risk-based analysis may provide an appropriate quality evaluation, its implementation is not straightforward since it requires a proper mathematical error propagation model, as well as a detailed description of the error behavior in the source data set. Without such information the risk of using a particular data set can not be estimated and a decision on the fitness for use can not be made. Thus the error characteristics and the error propagation model are interdependent.

These various difficulties give emphasis to the need for a more systematic and comprehensive approach to data quality issues. Particularly, the geometric accuracy of the data seems to be one of the most essential and challenging aspects of data quality. In view of this, it is suggested that a comprehensive approach to the geometric accuracy issues should be comprised of these three fundamental elements:

**Modeling** – the geometric accuracy of spatial data is a spatial phenomenon by itself. Modeling would require the ability to express accuracy in terms of spatial variability. A key point in achieving this is providing a proper stochastic model that can support a spatial phenomenon.

**Reporting** – in view of the required modeling characteristics proper descriptors of the resulting accuracy model should be available to users. These descriptors should enable transferring the full extent of the stochastic model adopted so that explicit usage of it could be made. Such reporting tools should also incorporate proper visualization capabilities through which users can realize the extent of uncertainty in the data. An example to such an application via an animation based tool was recently suggested by Ehlschlaeger et al. (1997).

**Propagation** – even if a proper error model can be fully and explicitly delivered, the implications of such information and the impact of the geometric uncertainty must still be explored by the user, in view of the intended usage. In some simple applications the accuracy implications are straightforward, yet in others such implications are not clear nor can they be formulated. Consequently, other means (such as stochastic simulations) should be used. An important issue here would be the ability to fully exploit the reported stochastic model for this purpose.

This contribution addresses the first of these elements, namely stochastic modeling. In this context the term "error" or "accuracy" refers to the positional discrepancy or geometric accuracy of the data, respectively. According to Kyriakidis et al. (1999), when attempting to perform such an analysis, it is assumed that two data types are available to the user:

- *Hard data*  $\{X_i^H, Y_i^H, i=1...m\}$  – a data set ("data points") of high accuracy and low volume (sparse and spread throughout the interest area). It is assumed that this data can not be used directly for application purposes due to its sparsity.
- *Soft data*  $\{X_i^S, Y_i^S, i=1...n\}$  – a group of  $n$  data sets available to the user (via a clearinghouse, for example). These data sets are of unknown quality, but have high volume.

It is also assumed that  $m \gg n$ . A set of residuals may be computed for the homologous points of the two data sets by:

$$dx_i = X_i^S - X_i^H, \quad dy_i = Y_i^S - Y_i^H, \quad (\vec{d}_i = \{dx_i, dy_i\}) \quad (1)$$

Based on Equation (1), it is now necessary to extract the characteristics of each of the  $n$  soft data sets, which will then serve as a criterion in the prioritizing process. This requires the ability to model and quantify the errors, as will be detailed in the next section.

## 1. SPATIAL UNCERTAINTY CALCULATIONS

### 1.1 Summary statistics

One of the well-known techniques of describing the population of the quantities obtained from Equation (1) is by using scalar quantities such as the mean,  $(\overline{dx}, \overline{dy})$ , and the variance  $(s_{dx}, s_{dy})$ . The computation of these quantities, also known as *summary statistics*, may be accompanied by various statistical tests for significance assurance (for example, Barbato, 2000). Unfortunately summary statistics falls short of taking into account the *spatial relations* within the population. Consequently, a different framework for describing the population of Equation (1) should be employed.

### 1.2 Random process theory

The shortcomings of summary statistics necessitate an extension of the statistical framework for analyzing and description of errors. Such a framework can be formed by treating the errors as a *random process*. A random process  $\Omega$  is an extension of the random variable space, implemented by introducing a function that depends on a given space  $D$  (for example, time or location) over the random variable space  $s$  (Peebles, 2001):

$$\Omega = \{Z(s) : s \in D, D \subset \mathbb{R}^2\} \quad (2)$$

This indicates that although the values of  $s$  are random, the relationships between these values are described by the function over  $s$ . Hence the spatial relations between data elements are accounted for. As a result of this advantage, the random process framework was suggested by several authors for describing errors in spatial data (Goodchild et al. (1992); Ehlschlaeger and Goodchild (1994); Hunter and Goodchild (1996); and Church et al. (1998)). Since the domain of spatial data is 2D or 3D location, the term *random field* is frequently used to describe a random process over this space.

### 1.3 Describing the spatial behavior of a random field

The random process theory serves as a statistical framework, which must be accompanied by proper indices. These indices can no longer be single scalars as they must account for the spatial domain of the random field. Such indices are widely used in the field of Geostatistics.

Two primary geostatistical indices, namely the *variogram* and the *correlelogram*, are used to characterize a random field. The (experimental) variogram (Equation (3a)) describes the variation of the variance between elements in the field, while the correlelogram (Equation (3b)) describes the correlation between data elements (Cressie, 1993):

$$\hat{\gamma}(h) = \frac{1}{2|N(h)|} \sum_{N(h)} [Z(s_i) - Z(s_j)]^2 \quad (a)$$

$$\hat{C}(h) = \frac{1}{|N(h)|} \sum_{N(h)} [(Z(s_i) - \bar{Z})(Z(s_j) - \bar{Z})] \quad (b) \quad (3)$$

where:

$$\bar{Z} = \frac{\sum_{i=1}^m Z(s_i)}{m} \quad (4)$$

$|N(h)|$  is the number of data pairs  $h$  units apart:

$$|N(h)| = \left\{ (s_i, s_j) : \|s_i - s_j\| = h ; \quad i = 1 \dots m ; \quad j = 1 \dots m \right\}, \quad (5)$$

and  $m$  is the size of the data set. For the discrete case, both indices are computed by dividing the space ( $D$ ) into equally spaced *lags* ( $h$ ), where for each lag an average value is taken. As can be seen from Equation (3), the variogram and the covariogram are only a function of the distance between data points, and that these estimators can be applied to stationary isotropic random fields. Several authors have also attempted to address the problem of variogram estimation for stationary non-isotropic fields (Wingle and Poeter, 1998).

#### 1.4 Least Squares Collocation

The random field scheme and its descriptors, the variogram and covariogram, provide a stochastic framework that can successfully describe the geometric accuracy of spatial data. Yet, it lacks the functionality needed, namely the ability to:

- Account for systematic effects in the data (trend)
- Predict the random field between data points (interpolation)
- Estimate the random field at the data points (filtering)

One scheme successfully implemented for physical geodesy purposes is the *Least-Squares Collocation* (LSC) scheme. In LSC the residuals obtained from Equation (1) are considered as a combination of two random components, namely a *signal* ( $s$ ) component and a *noise* ( $n$ ) component, and a *trend* component ( $x$ ) resulting from the transformation between the data sets. By taking the residuals obtained from Equation (1) as observations ( $l = \{dx_1, dx_2, \dots, dx_m ; dy_1, dy_2, \dots, dy_m\}$ ), the following relation between ( $l$ ) and the mathematical model ( $Ax$ ) could be derived (Moritz, 1972):

$$v = l - Ax = s + n, \quad (6)$$

where the discrepancies  $v$  are decomposed into a *signal* component ( $s$ ) and a *noise* component ( $n$ ):

$$s \sim N(0, C_s) \quad , \quad n \sim N(0, C_n). \quad (7)$$

A least-squares solution of Equation (10) is obtained by (Moritz, 1972) ; (Cross, 1983):

$$\begin{aligned} \hat{x} &= \left( A^T (C_s + C_n)^{-1} A \right)^{-1} A^T (C_s + C_n)^{-1} l \\ \hat{s} &= C_s (C_s + C_n)^{-1} (l - A\hat{x}) \\ \hat{n} &= C_n (C_s + C_n)^{-1} (l - A\hat{x}) \end{aligned} \quad (8)$$

where  $C_s$  is the variance-covariance matrix of the signal (a full matrix), which can be obtained from Equation (3b), and  $C_n$  is the variance-covariance matrix of the noise (a diagonal matrix) that can be estimated using Equation (3a).

## 2. PRACTICAL CONSIDERATIONS

There are several practical considerations that should be taken into account when implementing the LSC computational scheme in the context of spatial data and coordinate transformations in the presence of signals, of which two will be discussed here.

### 2.1 The effect of an error in the covariance function

The first consideration, the accuracy of the covariance function and its impact on the LSC results, is particularly essential to LSC due to the centrality of the covariance matrices in the estimation of  $x$ ,  $s$  and  $n$  (Moritz, 1972); (Cross, 1984); (Xu, 1991). This can be seen from Equation (11), where a solution will always be obtained if given  $C_n$  and  $C_s$  that comply with the basic rules of a covariance function (Blais, 1984). This led to the development of "specialized" analytical covariance functions for fields in which LSC is commonly implemented, such as the study of the gravity field (for example, Moritz (1980)) or photogrammetry (Ebner, 1976); (Rampal, 1976).

Unfortunately it can not be assured that a covariance function can be analytically derived when dealing with the general case of an existing spatial data set, from which Equation (1) is estimated. Since a spatial data set (used here as the soft set) may be the product of numerous operations (such as local geometric transformations or updating procedures) and can be composed of various sources, its geometric accuracy is not homogeneous nor does it comply with any definite physical rules. Thus, deriving an empirical covariance function directly from the data set is needed. Using the geostatistical estimators defined in Equation (3) is beneficial for this purpose. Yet the question of their accuracy and the impact of an erroneous covariance function should be addressed.

The influence of the error in the covariance matrix (which is based on the covariance function) was addressed by several authors. Moritz (1976) presented an expression for the accuracy of the estimated signal based on the true covariance function, and showed that an error in the covariance function will result in a non-optimal LSC solution. Furthermore, Moritz concluded that the accuracy of an LSC with an erroneous covariance function will always produce a less accurate signal estimation. A more practical analysis was derived by Xu (1991), who assessed the impact of an error in the covariance (or weight) matrix for the case of a unified LSC scheme. Krakowsky and Biacs (1990) discussed various aspects of statistical testing for LSC and indicated that a preliminary test for assessing the suitability of the mathematical model or the covariance matrices used is based on a statistical test of the estimated variance factor

### 2.2 The spatial distribution of data points – the point process scheme

It is well known that the distribution of the data points is essential in any coordinate transformation carried out. Well distributed data points will assure the quality of the transformation and eliminate singular cases, such as a set of data points located on a straight line or "leverage points" (Kampmann, 1996). For this reason it is a common practice to have data points on the circumference of the interest area. In the case of filtering or prediction the importance of the data point spatial distribution is even greater. This is because data points can be seen as "emitters of information", where the extent to which information is emitted is determined by the correlation function (Halmos et al., 1974). Thus an area without data points or with distant data points will suffer from low "information emittance", and the reliability of prediction or filtering may be doubted. It is therefore essential to evaluate the distribution of data points prior to the implementation of the LSC scheme in order to avoid intricate data point configurations. Unfortunately such an assessment can not be carried out directly with geostatistical estimators such as Equation (3).

A different approach, in which the data points themselves are considered a stochastic phenomenon, is the *point process* scheme. Unlike Equation (2), where the domain  $D$  is a fixed continuous domain, the domain of a point process is a random set by itself (Cressie, 1993). This approach has been widely used in field Biology, where events such as plants or bird nests are studied. Assuming no systematic data collection is applied (for example, in the form of a grid in the case of a DTM), a similar approach can be adopted here for the analysis of the hard data set. As the location of the hard data points is commonly determined by an operator or by a field surveyor in accordance to practical considerations, the distribution of the hard data set will frequently be a stochastic process as well. When each point is also described by an attribute (a *mark*), the stochastic framework can be extended to a *marked point process*, where the location as well as the marks are considered as stochastic processes. Additionally, a distinction between *mapped* point processes (all the events in the study area were collected) and *sampled* point processes (sample events were collected) is usually made. In our case the hard data is considered a mapped point process since all "events" are taken into account.

The basic question addressed when analyzing a point process is whether the events (points) are distributed in complete randomness (CSR – complete spatial randomness) or whether there are spatial patterns or characteristics that can be identified (Diggle, 1983). This distinction is elementary to the analysis since if the point process is considered as completely random there is no point in further searching for a distinct spatial pattern. If CSR was not established the point pattern can be either clustered or regular. A bench mark for checking CSR for a given point process  $N$  is the basic Poisson point process model (Ripley, 1981); (Cressie, 1993):

$$P(N(D) = n) = \frac{\mu(D)^n e^{-\mu(D)}}{n!}, \quad n = 0, 1, 2, \dots \quad (11)$$

Where (Diggle, 1983):

$$\mu(D) = E[N(D)] = \int_D \lambda(x) dx \quad (12)$$

$\lambda$  is therefore the *intensity* of the process, defined as (Diggle, 1983):

$$\lambda(x) = \lim_{|dx| \rightarrow 0} \left\{ \frac{E[N(dx)]}{|dx|} \right\} \quad (13)$$

Based on the Poisson point process, a test for CSR can be based either on a comparison of the quadrant count behavior or on a distance based analysis. Quadrant count tests are carried out by dividing  $D$  into a set of spatial cells ("quadrants") and counting the number of events in each cell. This count is then compared to the expected number of events resulting from a Poisson point process. The obvious difficulty in this estimation scheme is that it depends heavily on the type and size of cells used, hence its robustness is not assured (Ripley, 1981).

A more general scheme is based on *distances* instead of counts. In this scheme two types of points are considered in measuring distances, namely *Event points* (locations in which the process has occurred, for example tree locations) and *Sample points* (locations determined arbitrarily).

Based on this, two primary distance metrics could be derived:

- *Nearest neighbor empirical distribution function (G)* – this metric describes the distance distribution between a given event to the nearest event (Cressie, 1993):

$$\hat{G}(d) = \frac{\sum_{i=1}^N I(d_i < d)}{n}, \quad (14)$$

where  $n$  is a set of randomly chosen points from  $N(D)$  and  $I$  is an indicator function that gives 1 if the distance condition is fulfilled and 0 for all other cases. The empirical distribution can be compared to the expected distribution:

$$G(d) = 1 - e^{-\lambda\pi d^2} \quad (15)$$

- *Nearest event empirical distribution function ( $F$ )* – this metric describes the distance distribution between a given sample point and an event. For this metric the empirical and the theoretic distributions are identical to Equation (14) and Equation (15) respectively.

For our purposes, these two metrics can be utilized to estimate the distribution of data points for prediction as well as for filtering. For prediction purposes  $F$  can be used, as it estimates the probability of obtaining at least one data point within a given distance  $d$  from a sample point for which prediction is required (in fact  $F$  is also referred to as the "empty spaces" function (Diggle, 1983)). The  $G$  estimator can be beneficial for filtering purposes, since it can evaluate the data point configuration. For example, a high concentration of small distances would imply clustering, while a "step function" would imply a grid-like configuration.

### 3. CONCLUSION AND FUTURE WORK

Geometric accuracy modeling and analysis for spatial data is a demanding task and is of great importance for various activities, such as data exchange, data updating, or data analysis. The geometric accuracy of a spatial data set is a spatial stochastic phenomenon by its own right. Consequently proper spatial analysis frameworks should be employed if prediction and filtering are necessary. Although the Random field model can successfully serve as such a framework it still falls short of accounting for two fundamental aspects of the data, namely trends in orientation and the spatial configuration of the data points. The orientation problem can be partially addressed by a computation of a set of covariograms in different directions, if independence between  $x$  and  $y$  is assumed. A second shortcoming originates from the inability of Equation (3) to account for the spatial distribution of the data points. For this purpose point process analysis can serve as an indicator to the overall distribution of the data points. In addition, such a tool could also serve for designing the configuration of the hard data set *prior* to its collection in the field.

In view of the capabilities of such stochastic geostatistical tools further research is needed. In the field of computational frameworks a development of a discontinuity detection schemes is needed since there is no guarantee that the whole data set would be homogeneous. In fact, a more realistic approach might assume that the data set is a combination of patches from different sources. Furthermore, the development of generic spatial correlation functions is required, since in many cases a predefined correlation model can not be assumed or verified. This would also require proper accuracy estimators for the correlation functions themselves.

In addition, the development of reporting tools and error propagation schemes is needed for various GIS applications. This would require the development of protocols or standards, through which detailed accuracy information (such as variogram, covariogram,  $F$  or  $G$  functions, etc.) could be transferred to users or be utilized directly by a software module.

### 4. PRESENTATION

Due to limited length of the paper, a demonstration as well as examples will be given in a short oral presentation, in which the usage of the various tools will be discussed and analyzed.

## REFERENCES

- Agumaya A. and Hunter G.J., 1999. "Assessing 'Fitness for Use' of Geographic Information: What Risk Are We Prepared to Accept in Our Decisions?". "Spatial Accuracy Assessment – Land Information Uncertainty in Natural Resources", Lowell K. and Jaton A. (Editors), Ann Arbor Press, Chelsea Michigan. Pp. 35-43.
- Barbato F.D., 2000. "Accuracy Parameters Determination for GIS Base Map". In the proceedings of "Accuracy 2000", the 4<sup>th</sup> International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences", Amsterdam, July 2000. pp. 35-38.
- Blais J. A. R., 1984. "Generalized Covariance Functions And Their Applications In Estimation". *Manuscripta Geodetica*, Vol. 19. pp. 307-322
- Church R., Curtin K., Fohl P., Funk C., and Goodchild M. F., 1998. "Positional Distortions in Geographic Data Sets as a Barrier to Interoperation". ACSM Annual Convention, pp. 377-387.
- Cross P.A., 1983. "Advanced Least-Squares Applied to Position Fixing". Working paper No. 6, Northeast London Polytechnic, Department of Land Surveying, 205 pages.
- Diggle P. J., 1983. "Statistical Analysis of Spatial Point Patterns". Academic Press, New-York, USA. 148 pages.
- Ebner H., 1976. "Self Calibrating Block Adjustment". *Bildmessung und Luftbildwesen* 4/76. pp. 128-139.
- Ehlschlaeger C. R., Ashton M. Shortridge, and Michael F. Goodchild, 1997. "Visualizing Spatial Data Uncertainty Using Animation". *Computers in GeoSciences* Vol. 23, No 4
- Ehlschlaeger C.R. and Goodchild M.F., 1994. "Uncertainty in Spatial Data: Defining, Visualizing, and Managing Data Errors". *Proceedings of LIS/GIS '94 Annual Conference*, pp. 246-253.
- Goodchild M.F., Guoqing S. and Shiren Y., 1992. "Development and Test of an Error Model for Categorical Data". *Int. J. Geographic Information Systems*, Vol. 6 (2), pp. 87-104.
- Halmos F., Kadar I., and Karsay, F., 1974. "Local Adjustment By Least Squares Filtering". *Bulliten Geodesique*, 111 (1974). pp. 21-51.
- Hunter G.J. and Goodchild M.F., 1996. "A New Model for Handling Vector Data Uncertainty in Geographic Information Systems". *URISA Journal*, Vol. 8 (1), pp. 51-57.
- Kampmann G., 1996. "New Adjustment Techniques For The Determination Of Transformation Parameters For Cadastral And Engineering Purposes". *Geomatica*, Vol. 50(1). Pp. 27-34.
- Krakiwsky E. J., and Biacs Z. F., 1990. "Least Squares Collocation And Statistical Testing". *Bulletin Geodesique*, Vol. 64(1). pp. 73-87.
- Kyrakidis P.C., Shortridg A.M. and Goodchild M.F., 1999. "Geostatistics for Conflation and Accuracy Assessment of Digital Elevation Models". *Int. J. of Geographical Information Science*, Vol. 13, No. 7. pp. 677-707.
- Mikhail E.M., 1976. "Observations and Least-Squares (with contributions by F. Ackermann)". IEP-Dun Donnelley, New-York, 497 pages.
- Moritz H., 1972. "Advanced Least Squares Methods". *Reports of the Department of Geodetic Science*, Report No. 175. 129 pages.

- Peebles P.Z., 2001. "Probability, Random Variables and Random Signal Processing, 4<sup>th</sup> edition". McGraw-Hill Series in Electrical Engineering and Computer Engineering. 462 pages.
- Rampal K. K., 1976. "Least Squares Collocation In Photogrammetry". Photogrammetric Engineering and Remote Sensing, Vol. 42(5). pp. 659-669.
- Repley B. D., 1981. "Spatial Statistics". John Willy & Sons Press, New-York, USA. 252 pages.
- Wingle, W.L., and Poeter E.P., 1998. "Directional Semivariograms: Kriging Anisotropy Without Anisotropy Factors" Advances in Geostatistics, 1998 AAPG Annual Meeting, Salt Lake City, Utah, May 17-20, 1998.
- Xu V. P., 1991. "Least Squares Collocation With Incorrect Prior Information". ZfV 6/1991. pp. 266-273.