

# UNIVERSAL ERROR PROPAGATION LAW

Xiaoyong CHEN<sup>a</sup> and Shunji MURAI<sup>b</sup>

<sup>a</sup> Asian Institute of Technology, P.O. Box 4, Klong Luang, Pathumthani 12120, Thailand - xychen@ait.ac.th

<sup>b</sup> IIS, University of Tokyo, 4-6-1. Komaba, Meguro-ku, Tokyo 153-8505, Japan – shi939murai@nifty.com

Commission II, SS-15

**KEY WORDS:** Accuracy, Analysis, Modelling, Spatial Information Science, Statistics

## ABSTRACT:

As an ubiquitous statistical theory, Gaussian Distribution (GD) or Gaussian Error Propagation Law (GEPL) has been widely used for modelling random errors in many engineering and application fields since 1809. In recent years, this theory has been extended to handle the uncertainties of spatial data in GIS, such as positional error modelling. But most of the results for spatial error modelling based on GD are contradictory with common senses and natural laws, such as energy law and Tobler's First Law (TFL) in geography. This paper presents a novel statistical approach for rigorous modelling of positional errors of geometric features in spatial databases. Based on Generalized Gaussian Distribution (GGD) and using errors in local points as the fundamental building blocks, a new spatial statistical theory – Universal Error Propagation Law (UEPL) is presented to handle global error propagations for spatial random sets (or objects). Practical examples and simulations are given to illustrate the error propagations based on UEPL for various spatial objects. Finally, the relationships between UEPL and Newtown's Universal Gravitation Law (NUGL) and TFL have been successfully established, which shows that UEPL is a new discovered natural law for spatial information field.

## 1. INTRODUCTION

As one of the three most important emerging and evolving fields along with nanotechnology and biotechnology in this century (Gewin, 2004, Nature), Geo-spatial Information System (GIS) plays increasingly important roles in decision-making processes in many disciplines that involve planning, research and management by using spatial data at the different spatial levels. However, more effective use of GIS requires explicit knowledge of the uncertainty inherent in the spatial data. Therefore, the quality of GIS application is strongly dependent on the quality of spatial data which needs a formal theory for handling spatial errors in GISs (Goodchild, 1989).

De Morgan (1838) asked, "what do we mean by a *law of error*?" in *Essay on Probabilities* and went on to describe "the standard law of facility of error". This law had been used by Gauss (1809) in his first theory of least squares and is called Gaussian Distribution (GD) or Gaussian Error Propagation Law (GEPL) today. As in many other fields of science, GD has played a predominant role in surveying data processing and error modelling (Mikhail and Ackermann, 1976). In the vast majority of these applications, it has been assumed that the point-based observation error under investigation is distributed with GD. This popularity of the Gaussian assumption has been motivated mainly by the theoretical appeal of GD due to the *central limit theorem* and equally by the desirable analytical properties of Gaussian Probability Density Function (PDF) which generally leads to linear equations. A Gaussian PDF is also the maximum-entropy density (Papoulis, 1991) when only the first two moments of a process are known.

In a vector-based GIS, spatial features are defined on the basis of point features, e.g. a line is defined as a sequence of digitized points connected by line segments, and a polygon is defined as the interior of its boundary delimited by a closed line. Then, the positional uncertainty of points should form the basis for

uncertainty analysis of all spatial objects. From this point of view, GD and GEPL have been extended for modelling spatial

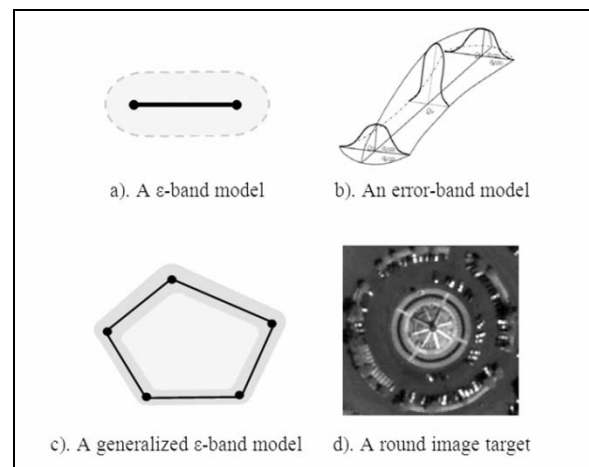


Figure 1. Different uncertainty models.

errors in GIS, such as *ε-band model* (Chrisman, 1982), *error-band model* (Shi, 1994), and *generalized ε-band model* (Leung and Yan, 1998) [Figure 1 (a)-(c)] and etc. In the same way, the spatial features in an image database are defined based on the pixels. Error propagation among the neighbouring pixels plays an important role for analyzing image accuracies. As the example illustrated in Figure 1 (d), the center point  $(x_c, y_c)$  of a round-target can be estimated by its boundary pixels as  $x_c = \frac{1}{n} \sum_{i=1}^n x_i$  and  $y_c = \frac{1}{n} \sum_{i=1}^n y_i$ . Then, its positional accuracies can be estimated by  $\sigma_{x_c} = \sigma_{x_0} / \sqrt{n}$  and  $\sigma_{y_c} = \sigma_{y_0} / \sqrt{n}$  based on GEPL, where  $\sigma_{x_0}$  and  $\sigma_{y_0}$  are the standard deviations of a pixel in  $x$  and  $y$  directions.

However, a large class of errors encountered in many real-world problems can be characterized as non-Gaussian and frequently as the distributions with heavy tails. It is a common experience that conventional least-squares estimation techniques based on GD perform very poorly in removing these real-world errors. Similarly, spatial data are always autocorrelated and dependent on each other (Cressie, 1993). Modeling spatial errors based on GD/GEPL always cause the contradictions with common senses and natural laws, such as energy law and Tobler's First Law (TFL) in geography (Tobler, 1970). For example, in the error-band model shown in Figure 1 (b), the accuracy of a distant point is always higher than the accuracy of a starting control-point. Likewise, in Figure 1 (d), when  $n$  is becoming very large,  $\hat{\sigma}_{xc}$  and  $\hat{\sigma}_{yc}$  will be extremely small. This is far below the limitation of empirical tests.

In the past, various models for processing spatial errors have been developed from both *local* and *global* levels. In the local level, Robust Estimation has been developed based on non-Gaussian error distributions (Huber, 1981; Hampel and et al, 1986), in which the distant (or spatial) errors are treated as the outliers that are going to be eliminated from the final estimation result. In the global level, Geo-statistics has been developed by Matheron (1963), in which spatial errors are presented as *stationary random processes*. The *kriging* methods have also been used for handling spatial autocorrelations in Geo-statistics (Cressie, 1993).

The objective of this paper is to develop a novel statistical approach to rigorously model the positional errors of geometric features in various spatial databases. Based on Generalized Gaussian Distribution (GGD) and using errors in local points as the fundamental building blocks, a new spatial statistical theory –Universal Error Propagation Law (UEPL) is presented to handle global error propagations for spatial random sets (or objects).

## 2. GENERALIZED GAUSSIAN DISTRIBUTION

### 2.1 Probability Distribution Function of GGD

A random variable is distributed as Generalized Gaussian Distribution if its PDF is given by (Müller, 1993)

$$f(x|\mu, \sigma, p) = \frac{p}{2\Gamma(1/p)\pi(p)\sigma} e^{-\frac{|x-\mu|^p}{\sigma^p}} \quad (1)$$

where  $\Gamma(\cdot)$  is the gamma function,  $\mu$  is the mean,  $\sigma^2$  is the variance,  $\pi(p) = [\Gamma(1/p)/\Gamma(3/p)]^{1/2}$  is a scaling factor, and  $p$  is a positive shape parameter that describes the overall structure of GGD. With  $p=2$ , GGD reduces to a standard GD, with  $p=1$  to a Laplacian distribution. Whereas in the limiting case  $p \rightarrow +\infty$ , GGD converges to a uniform distribution in  $(\mu - \sqrt{3}\sigma, \mu + \sqrt{3}\sigma)$ , and when  $p \rightarrow 0+$  to a degenerate one in  $x = \mu$ . When  $p < 2$ ,  $p = 2$  and  $p > 2$ , GGD is classified into super-GD, GD and sub-GD respectively. The notation  $x \sim N^*(\mu, \sigma, p)$  donates that  $x$  is a random variable with PDF as in Equation (1), and  $N^*(\sigma, p) = N^*(0, \sigma, p)$ .

### 2.2 Parameter Estimation of GGD

The Maximum Likelihood (ML) estimation based on GGD is equivalent to  $L_p$ -norm estimation (Müller, 1993), i.e.

$\sum_{i=1}^n |x_i - \mu|^p = \min$ . Let  $X = (x_1, x_2, \dots, x_n)$  be a random vector with  $X \sim N^*(\mu, \sigma, p)$ , the ML estimated parameters  $\hat{\mu}$ ,  $\hat{\sigma}$  and  $\hat{p}$  can be derived as

$$\hat{\mu} = \frac{\sum_{i=1}^n \frac{|x_i - \mu|^{p-2}}{\sum_{j=1}^n |x_j - \mu|^{p-2}} x_i}{\sum_{i=1}^n \frac{|x_i - \mu|^{p-2}}{\sum_{j=1}^n |x_j - \mu|^{p-2}}} \quad (2)$$

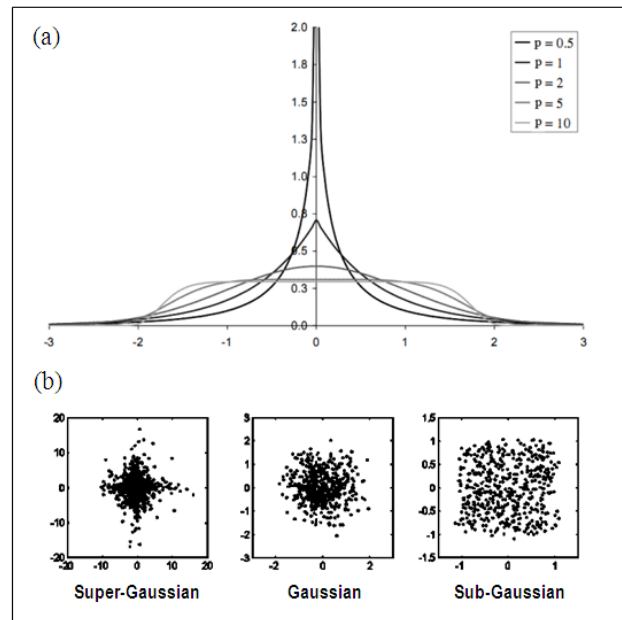


Figure 2. GGDs: a) PDFs; b) Sub and Super Gaussians.

$$\hat{\sigma} = \frac{1}{\pi(p)} \left( \frac{p}{n} \sum_{i=1}^n |x_i - \mu|^p \right)^{1/p} \quad (3)$$

$$\begin{aligned} \psi + \psi(1/p) + \log(p) + \log \left( \frac{1}{n} \sum_{i=1}^n |x_i - \mu|^p \right) \\ - \frac{\psi \sum_{i=1}^n |x_i - \mu|^p \log(|x_i - \mu|)}{\sum_{i=1}^n |x_i - \mu|^p} = 0 \end{aligned} \quad (4)$$

where  $\psi(x) = \Gamma'(x)/\Gamma(x)$ , is the digamma function.

Kuruoglu and et al (1998) have proven that the estimated parameter  $\hat{\mu}$  is the *mean* when  $1 < p \leq 2$  and the *median* when  $0 < p \leq 1$ . To get the solutions of Equation (2)-(4), they have also suggested an efficient way called Iteratively Reweighted Least-squares Algorithm (IRLA), which starts from the well-known least-squares solution and at each iteration solves a new least-squares problem by employing the weighted residuals from the previous iteration.

Another good initial guess for the shape parameter  $p$  can be found based on the matching moments of the data set with those of the assumed distribution (Sharifi and Leon-Garcia, 1995). Letting  $m_1 = \sum_{i=1}^n |x_i - \mu|$  and  $m_2 = \sum_{i=1}^n |x_i - \mu|^2$  which denote absolute moments,  $p$  is estimated by

$$\beta = M^{-1} \left( \frac{m_1}{\sqrt{m_2}} \right) \quad (5)$$

where

$$M(\varphi) = \frac{f(\varphi/\beta)}{\sqrt{f(\varphi/\beta)f(\varphi/\beta)}} \quad (6)$$

### 3. UNIVERSAL ERROR PROPAGATION LAW

#### 3.1 Different Random Data Sets

Generally, there are four types of common-used random data sets: *random variable*, *random process*, *random field*, and *random set* (Figure 3). Their key characteristics are briefly summarized as follows:

- A *random variable* is the mapping of the elements of a sampling space onto a set of *numerical values*. It is the basis for all point-based statistical analyses, and has been widely used for modelling random errors in classical probability theories and for handling both random errors and outliers in Robust Estimations.
- A *random process* is the mapping of the elements of a sampling space to a space of *continuous functions* of time  $t$ , where  $t$  is a distinct variable. It has been mainly used for *time series analysis* according to *autocorrelation functions* in different times.
- A *random field* is the simple extension of a random process from one-dimensional  $t$  to two-dimensional  $x$  and  $y$ , where distances  $|x|$  and  $|y|$  are still treated as distinct parameters even though both  $x$  and  $y$  are random variables. Geo-statistics is based on this random data type (Cressie, 1993).
- A *random set* (or *object*) is the mapping of a confidential region for an estimated parameter. Consider the family of all sets of interesting (usually closed sets). Equip it with an  $\sigma$ -algebra and then define a random set as a measurable map from a given probability space to this space of sets. It is the most generalized random data type that we would like to use for modelling spatial errors in this paper.

#### 3.2 Relativity of Spatial Probability Distribution

A random set is generally complicated. In a spatial random set, the probability distributions at different points may be heavily overlapped with each other or dramatically changed place to place. Up to the knowledge of the first two moments ( $\mu$  and  $\sigma^2$ ) and the shape parameter  $p$ , GGDs can be used for approximating the real PDFs at each point in the given random set. Because of the flexibility of GDD, it can adaptively account all types of errors, such as sub-Gaussian, Gaussian and super-Gaussian.

Let  $X = \{x_1, x_2, \dots, x_n\}$  be a set of positional random variables with  $x_i \sim N^*(\mu_i, \sigma_i, p_i)$ , and  $Y$  be a linear function of  $X$ , i.e.  $Y = x_1 + x_2 + \dots + x_n$ . It has been proven that  $Y$  is no longer a random variable of GGD except when  $p = 2$  (Schilder, 1970). This is because  $L_p$ -norm estimation based on fractional lower order moments will inevitably introduce nonlinearity to even linear problems. The linear space of GGD is a *Hilbert space* when  $p = 2$ , a *Banach space* when  $1 \leq p < 2$ , and only a *metric space* when  $0 < p < 1$ . Banach and metric spaces do not have as nice properties and structures as Hilbert spaces for linear estimation problems. Therefore, it is a serious problem to apply GGD for spatial error modeling.

The key problem that needs to be solved is how to separate the spatial-independent local observations (i.e. without outliers) from spatial-mixed global observations (i.e. with outliers). Independent Component Analysis (ICA) in signal processing can be applied for this purpose (Choi and et al, 1998). In ICA, a weight function is adapted in such a way to make the output observations as spatially independent as possible. This can be achieved by composing a valid objective function that attains its extrema when the output observations become spatially independent. Both Infomax and ML approaches lead to the best separation *nonlinearity* as

$$\varphi_1(x_i) = -\frac{\partial \log f_1(x_i)}{\partial x_i} = \frac{\dot{f}_1(x_i)}{f_1(x_i)} \quad (7)$$

where  $f_1(x_i)$  and  $\dot{f}_1(x_i)$  are the PDF and its derivative of the source signals  $x_i$ . According to Equation (1), the nonlinearity is

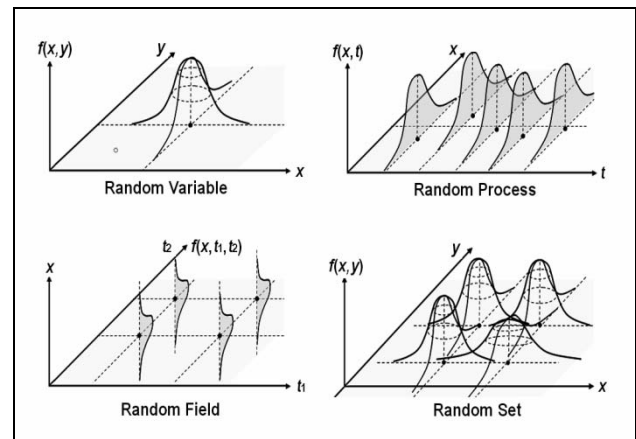


Figure 3. Different types of random data.

$$\varphi_1(x_i) = \frac{p_i}{\tau(\varphi_i)^{p_i}} \text{sgn}(v_i) \cdot |x_i|^{p_i-2} \quad (8)$$

where  $\text{sgn}(v_i) = x_i/|x_i|$ . In the following deductions, it can be found that the solution of Equation (8) is equal to the  $L_p$ -norm estimation of Equation (2).

Let  $v_i = x_i - \beta$  be a centralized observation, Equation (2) can be rewritten as

$$\sum_i^p \text{sgn}(v_i) |v_i|^{p-2} = 0 \quad (9)$$

The scale parameter  $p/\tau(\varphi_i)^{p_i}$  in Equation (8) is a constant for a single GGD like in Equation (9), so it can be omitted. Then, Equation (9) can be continually simplified to  $\sum_i^p \varphi_1(x_i) = 0$ , which is equivalent to a simple *arithmetic mean* derived by least-squares estimation.

Up to the knowledge of the first two moments ( $\mu$  and  $\sigma^2$ ), the PDF of  $\varphi_1(x_i)$  can be approximately treated as a GD. It is because the PDF of  $\varphi_1(x_i)$  is close to a GD for a linear mixing of independent variables due to *central limit theorem*, the difference between maximizing the distance to the observation

or to a GD does not matter in practice (Lee and et al, 1997). In addition, if one assigns the first two moments of a PDF to agree with such information, but has no further information and therefore imposes no further constraints, then a GD fit to those moments will, according to the principle of Maximum Entropy, represent most honestly his/her state of knowledge about the error. Eventually, through the Gaussian-liked observation  $\varphi_1(x_i)$ , GEPL can be conducted for modeling spatial errors.

In summary, the PDF at each location in a spatial random set is relative depending on the different level of viewpoints. At the global level, the PDF of  $x_i$  is approximated to a GGD, i.e.  $x_i|_g \sim N^*(\mu_0, \hat{\sigma}_0, \rho_0)$ , and at the local level, the PDF of  $x_i$  is approximated to a GD, i.e.  $x_i|_l \sim N(\mu_0, \sigma_0)$ . The relationship between  $\sigma_0$  and  $\hat{\sigma}_0$  can be proven (the detail deductions are omitted due to the limited length of this paper) as

$$\sigma_0 = \frac{\rho_0 \Gamma(2-1/\rho_0) \Gamma(3/\rho_0)}{\Gamma(1/\rho_0)} \cdot \hat{\sigma}_0^{\rho_0-1} \quad (10)$$

### 3.3 Universal Error Propagation Law

Let  $X = [x_1, x_2, \dots, x_n]^T$  be a random vector with  $x_i|_l \sim N(\mu_0, \sigma_0)$  and  $x_i|_g \sim N^*(\mu_0, \hat{\sigma}_0, \rho_0)$ . The covariance matrix of

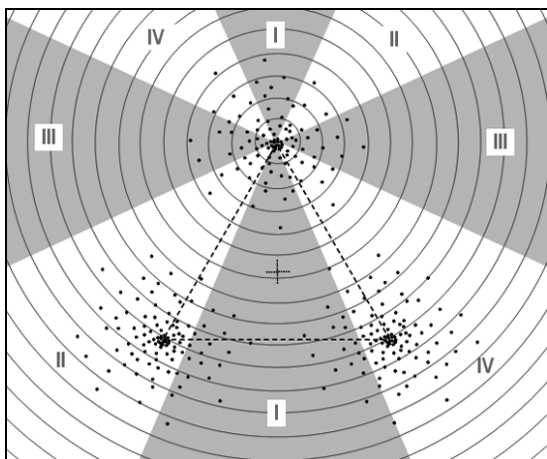


Figure 4. A simulated random set.

$X$  is known as

$$D_X = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12} & \dots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22}^2 & \dots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \dots & \sigma_{nn}^2 \end{bmatrix} \quad (11)$$

Let  $Y = f(X) = f(x_1, x_2, \dots, x_n)$  be a nonlinear function about  $X$ .  $Y$  can be approximately represented by its first-order Taylor series expansion about the approximate value of  $X$  at the point  $s$ ,  $X^s = [x_{1,s}, x_{2,s}, \dots, x_{n,s}]^T$ , as

$$Y = Y^s + \frac{\partial f}{\partial X} \Big|_{X^s} (X - X^s) \quad (12)$$

where  $Y^s = f(X^s) = f(x_{1,s}, x_{2,s}, \dots, x_{n,s})$ . Let  $dY^s = Y - Y^s$ ,  $dX^s = X - X^s$  and  $F_s = \frac{\partial f}{\partial X} \Big|_{X^s}$ . Since both  $X^s$  and  $Y^s$  are not random vectors,  $D_{dY^s} = D_X^s$  and  $D_{dY^s} = D_Y^s$ . From Equation (8), we have  $D_X^s = W_s^{-1} D_Y^s$ , where  $W_s$  is a weight matrix for handling the spatial dependences at the point  $s$ , and is defined as

$$W_s = \begin{bmatrix} W_{1,s} & 0 & \dots & 0 \\ 0 & W_{2,s} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & W_{n,s} \end{bmatrix} \quad (13)$$

in which each weight factor  $W_{i,s}$  is represented as

$$W_{i,s} = \begin{cases} \frac{F_i}{r_i(\rho_i)^{\rho_i}} (dx_{i,s})^{\rho_i-2}, & \text{when } dx_{i,s} > 1 \\ 1, & \text{when } dx_{i,s} \leq 1 \end{cases} \quad (14)$$

According to GEPL, from the linear function  $dY^s = F_s dX^s$ , Universal Error Propagation Law (UEPL) can be derived as

$$D_Y^s = F_s D_X^s F_s^T = F_s W_s^{-1} D_Y^s F_s^T \quad (15)$$

### 3.4 Simulation

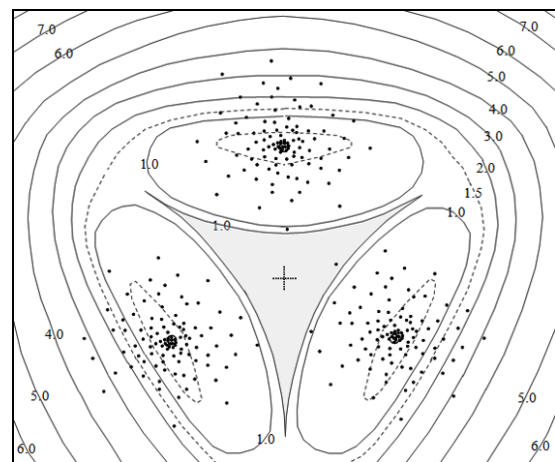


Figure 5. The accuracy counter-lines derived by UEPL.

In Figure 4-5, a Monte-Carlo simulation is accomplished to test UEPL for propagation of spatial errors, in which a triangle with equal side ( $d=5$ ) is put in the centre, and 100 random variables  $x_i \sim N(0, 1)$  ( $i=1, 2, \dots, n$ ) are simulated at each vertex. For calculation of  $\mu_1$  at each vertex, the data set is divided into four zones according to the different directions, and then the data in each zone are processed as a one-dimensional profile according to the Euclidean distances. The final  $\mu_1$  can be estimated by the simple arithmetic mean of the four zone results. Due to the symmetry of the vertexes, their shape parameters are the same and estimated as  $\rho \approx 1.333$ .

According to UPEL, the spatial accuracy  $\sigma(x, y)$  at each location can be estimated, and the derived accuracy counter-lines are shown in Figure 5. Meanwhile,  $\sigma(x, y)$  can be also



derived by using Equation (3) and Equation (10) based on the direct-estimated shape parameter  $p(x, y)$ . For example, at the centre point of the triangle,  $\sigma_c$  derived by UEPL is 0.982, and  $\sigma_c^1$  calculated based on  $p_c = 1.152$  is 0.993. This can be used for checking the estimation quality of UEPL.

#### 4. SPATIAL ERROR MODELLING

##### 4.1 Vector Lines

Examples of the error bands derived by vector lines based on UEPL (called as *UEPL-bands*) are shown in Figure 6. Comparing with  $\epsilon$ -bands and *error-bands*, UEPL-bands can automatically adapt to the intervals of sampling points. When the interval is small, the UEPL-band becomes smooth and approximates to the true line. Inversely, when the interval is large, the UEPL-band becomes abrupt and wider.

##### 4.2 Vector Polygons

There are two different ways to represent the error models for a vector polygon, i.e. *boundary-model* and *area-model*. A boundary-model is just the same as the error model of a closed boundary line [figure 7(a)]. But an area-model is the error propagation for the whole polygon [figure 7(b)]. Generally, the weight centre of a polygon is recognized as the position with the highest accuracy, and the gradient of the error propagation should be different comparing with the inside and the outside of a polygon. If  $p$  ( $0 < p \leq 2$ ) is the shape parameter for the outside of a polygon, then  $p' = 4 - p$  can be selected as the shape parameter for the inside of the same polygon.

##### 4.3 Relativity of Distance

When the additional information is given, the results of spatial error propagations based on UEPL can be adjusted according to the determined relative distances. In this case, Equation (14) can be adapted to

$$w = \begin{cases} \frac{p}{r(p)^2} (k \cdot d)^{p-2} & \text{when } d > 1 \\ 1 & \text{when } d \leq 1 \end{cases} \quad (16)$$

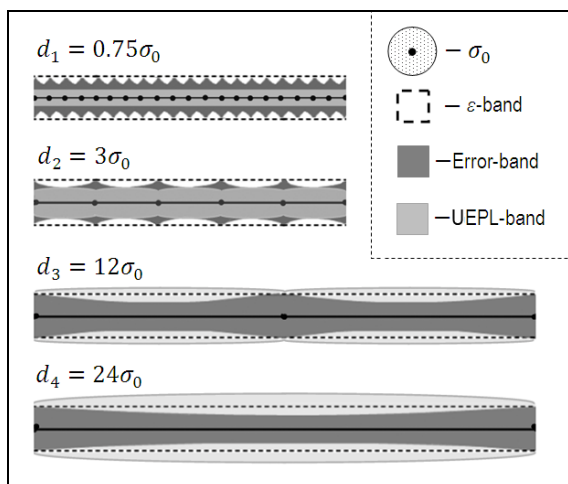


Figure 6. The different error models for vector lines.

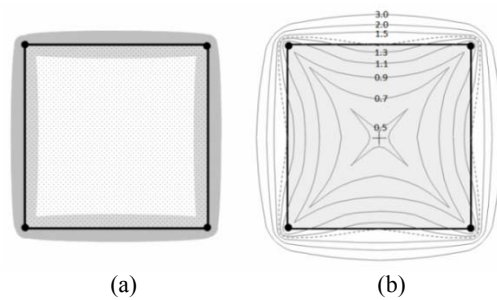


Figure 7. The error models for a polygon.

where  $k$  is the constant for adjusting the real distances. The example is shown in Figure 8: (a) is the result based on the real sampling points; (b) is the result derived by the compressed points; (c) is the result of the adjusted relative distances.

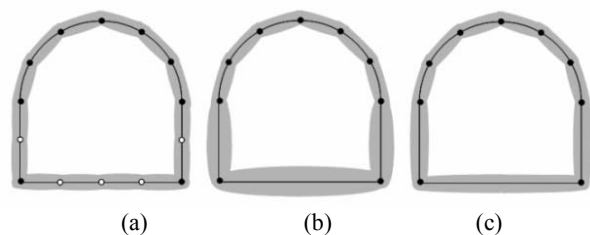


Figure 8. The relative error models for vector lines.

##### 4.4 Raster Objects

According to UEPL, the accuracy of the centre point  $(x_c, y_c)$  of a round-target in Figure 1 (d) can be estimated as

$$\sigma_c = \sqrt{\frac{k(p)}{n}} \sigma_0 r^{p-2} = \sqrt{\frac{k(p)}{2\pi}} \sigma_0 r^{\frac{1-p}{2}} \quad (17)$$

where  $\sigma_0$  is the standard deviation of a pixel,  $r$  is the radius of the round-target,  $n = 2\pi r$  is the number of boundary pixels, and  $p$  is the shape parameter of each boundary-pixel. For example,  $\sigma_c = 0.22$  ( $\sigma_0 = 1, p = 2, r = 3$ );  $\sigma_c = 0.17$  ( $\sigma_0 = 1, p = 1.6, r = 50$ );  $\sigma_c = 0.12$  ( $\sigma_0 = 1, p = 1.5, r = 100$ ) and  $\sigma_c = 0.48$  ( $\sigma_0 = 1, p = 1, r = 2000$ ). In summary, the accuracy can be achieved to  $0.1\sigma_0 \leq \sigma_c \leq 0.5\sigma_0$ , which is very close to the results of the empirical experiments.

#### 5. THEORITICAL ANALYSIS

##### 5.1 Comparing with GEPL

When  $p_i = 2$  ( $i=1, 2, \dots, n$ ), Equation (11) is simplified to  $\sigma_y = \sqrt{D_x} \sigma^2$ , which shows that UEPL is a natural extension of GEPL for modeling spatial errors. Figure 9 shows the universal error propagations for the different shape parameters: (1) when  $0 < p < 1$ , UEPL is the extrapolation of the spatial errors far outside the data group; (2) when  $1 < p < 2$ , UEPL is the interpolation of the spatial errors among the sampling points within the data group; (3) when  $p = 2$ , UEPL is the same as

GEPL for estimation of the local errors only; (4) when  $p > 2$ , UEPL is the inverse extrapolation of the spatial errors inside the spatial objects, such as polygons.

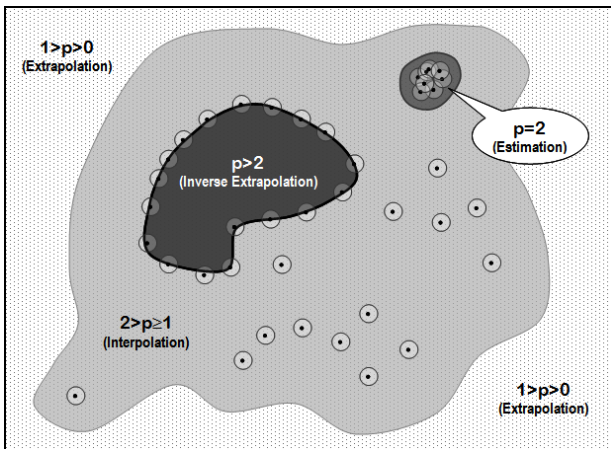


Figure 9. Universal error propagations with different  $p$ .

### 5.2 Comparing with Newton’s Universal Gravitation Law

Let  $k(r) = \gamma / r^2$  be the scale function of  $p$ , and  $r$  be the distance between the objects  $A$  and  $B$ . Since the mass of  $A$  is equal to the mass of  $B$ , i.e.  $m_A = m_B = \int_{-\infty}^{\infty} f(x) dx = 1$ , Equation (14), when  $r > 1$ , can be changed to

$$w = k(r) \frac{m_A m_B}{r^{2-p}} \quad (18)$$

which shows that the spatial-dependent factor  $w$  in UEPL is just the same as Newton’s Universal Gravitation Law (NUGL) (Obanian and Ruffini, 1994), i.e.  $w$  can be treated as a type of gravity in the information field. Here, the reason why the shape parameter  $p = 0$  is because the two PDFs of the objects  $A$  and  $B$  are overlapped in their tails. Generally, there are four different cases for the shape parameters: (1) when  $p = 0$  (i.e. spatial data only), the two PDFs are completely separated, and Equation (18) is just the same as NUGL; (2) when  $0 < p < 2$  (i.e. mixed spatial and local data), the two PDFs are overlapped in different degrees; (3) when  $p = 2$  (i.e. local data only), the difference between two PDFs is within  $\pm 3\sigma_p$ ; (4) when  $p > 2$ ,  $B$  is inside the object  $A$ , and since  $w$  may converge to infinite when  $r$  is extremely large, it may cause the special-affected areas which are just similar to the “black holes” in Cosmology (Obanian and Ruffini, 1994; Zwiebach, 2004).

### 5.3 Mathematical Representation of TFL

As Tobler’s First Law in geography mentioned, “Everything is related to everything else, but near things are more related than distant things” (Tobler, 1970). Unfortunately, this law is just an approximated description based on plain words, but it is not a rigorous mathematical representation. For example, it does not tell the people how to conceptualize *nearness* or which proximity measure to quantify *relatedness*. In this case, UEPL can be contributed to fill this gap. According to Mikhail and Ackermann (1976), a *weight* can be represented as  $w_i = \sigma_o^2 / \sigma_i^2$ . On the other hand, the Fisher’s information value  $I$  (Frieden, 2004) can be represented as

$$I = E \left( \frac{\partial \log f(x)}{\partial x} \right)^2 \quad (19)$$

When  $f(x)$  is a Gaussian PDF,  $I_o = 1/\sigma_o^2$  and  $I_i = 1/\sigma_i^2$ . For the reason, we can define  $\rho(I_i) = I_o/I_i = w_i$  as the measure of *similarity* of the Fisher’s information values between the points  $o$  and  $i$  with the distance  $r_i$ . From Equation (18), when  $r_i > 1$ , we have

$$\rho(I_i) = \frac{k(r)}{r_i^{2-p}} \quad (20)$$

which is a type of the mathematical representation of TFL. By using Equation (20), the *nearness* defined in TFL can be represented by the distance function  $r_i^{2-p}$  and the *relatedness* can be represented by the measure of *similarity*  $\rho(I_i)$ . Since  $\rho(I_i) > 0$ , the spatial-dependence represented by TFL is always positive. When  $\rho(I_i) > 1$ , UEPL generates the special-affected areas [such as the shading areas in Figure 5 and 7(b)], which are always surrounded by the concave boundaries. According to *String theory* (Zwiebach, 2004), these areas represent a new type of space structures which may generate “black holes” (Obanian and Ruffini, 1994).

## 6. CONCLUSION

We have proposed in this paper a new spatial statistics theory – Universal Error Propagation Law to handle errors for spatial random sets (or objects). The specific contributions are:

- Based on Generalized Gaussian Distribution, UEPL has been presented for modelling spatial errors;
- By using Monte-Carlo simulation, the efficiency and quality of error processing based on UEPL has been evaluated;
- According to UEPL, practical examples are given for generating the error models for different spatial objects. Comparing with the exiting error models, such as  $\epsilon$ -band model and error-band model, our approach put the end to the long discussions about spatial data uncertainty in GIS. This means that without considering spatial dependences rigorously, no method can get the accurate estimation result for modelling spatial errors.
- The relationships between UEPL and NUGL/TFL show that UEPL is a new discovered natural law for spatial information field.

Our future works will be concentrated on two major directions: (1) in the theoretical aspect, we will extend UEPL to high-dimensional spatial and spatial-temporal data sets; (2) in the application aspect, we will use UEPL to establish various simplified and hierarchical error models for processing real-world spatial errors.

## REFERENCES

Choi, S., A. Cichocki and S. Amari, 1998. Flexible independent component analysis, *Neural Networks for Signal Processing VIII*, pp. 83-92.

- Chrisman, N. R., 1982. A theory cartographic error and its measurement in digital database, *Proceedings of Auto-Carto 5*, Bethesda, MD, USA, pp. 159-168.
- Cressie, N., 1993. *Statistics for Spatial Data*, Wiley, New York.
- Frieden, B. R., 2004. *Science from Fisher Information*, Cambridge University Press.
- Goodchild, M.F., 1989. Modelling error in objects and fields, *Accuracy of Spatial Databases*, Taylor and Francis, pp. 107-114.
- Hampel, F., E. Ronchetti, P. Rousseeuw and W. Stahel, 1986. *Robust Statistics: The approach based on influence functions*, John Wiley and Sons.
- Huber, P., 1981. *Robust Statistics*, John Wiley and Sons.
- Kuruoglu, E., P. Rayner and W. Fitzgerald, 1998. Least Lp-norm impulsive noise cancellation with polynomial filters, *Signal Processing* 69, pp. 1-14.
- Leung, Y. and J. Yan, 1998. A locational error model for spatial features, *International Journal of Geographical Information Science*, 12(6), pp. 607-620.
- Matheron, G., 1963. Principles of geo-statistics, *Economic Geology*, 58, pp. 1246-1266.
- Mikhail, E. M. and F. Ackermann, 1976. *Observations and Least Squares*, IEP, New York.
- Müller F., 1993. Distribution shape of two-dimensional DCT coefficients of natural images, *Electron. Lett.* 29 (22), pp. 1935-1936.
- Obanian, H. C. and R. Ruffini, 1994. *Gravitation and Space-time*, W.W. Norton & Company, New York.
- Papoulis, A., 1991. *Probability Random Variables and Stochastic Processes*, McGraw-Hill, New York.
- Schilder, M., 1970. Some structure theorems for the symmetric stable laws, *Ann. Math. Statist.*, 41(2), pp. 412-421.
- Sharifi K. and A. Leon-Garcia, 1995. Estimation of shape parameter for generalized Gaussian distribution in sub-band decompositions of video, *IEEE Trans. Circuits Syst. Video Technol.* 5 (1), pp. 52-56.
- Shi, W., 1994. *Modelling Positional and Thematic Uncertainties in Integration of Remote Sensing and GIS*, ITC Publication, The Netherlands.
- Tobler, W.R., 1970. A computer movie simulating urban growth in the Detroit region, *Economic Geography*, 46, pp. 234-240.
- Zwiebach, B., 2004. *A First Course in String Theory*, Cambridge University Press.

