DOI :10.13203/j.whugis2002.05.005
第 27 卷 第 5 期
2002 年 10 月
武汉大学学报·信息科学版
Geomatics and Information Science of Wuhan University
Vol. 27 No. 5
Oct. 2002

文章编号: 1000-050X(2002)05-0462-05
文献标识码: A

GIS 中线元的误差熵带研究

李大军[1]　龚健雅[1]　谢刚生[2]　杜道生[1]

(1　武汉大学测绘遥感信息工程国家重点实验室, 武汉市珞喻路 129 号, 430079)
(2　华东理工学院测量系, 江西抚州市环城西路 14 号, 344000)

摘　要: 基于现有的线元位置不确定性模型大多与置信水平的选取有关, 而置信水平的选取带有一定程度的主观性, 因而不能惟一确定。引入信息熵理论, 提出了线元的误差熵带模型, 并将它与"*E*-带"进行了比较, 计算了落入其内的概率。该模型根据联合熵惟一确定, 与置信水平的选取无关。

关键词: 线元; 位置不确定性; 联合熵; 误差熵圆; 误差熵带
中图法分类号: P208

空间数据的不确定性是 GIS 一个基础理论问题, 其中线元的位置不确定性是研究的一个重点。矢量 GIS 中, 线元常作为基本单元参与 GIS 的各种操作, 因此应将它作为一个整体来分析其位置不确定性。线元的位置不确定性研究是基于带的概念展开的, Pekal[1] 在 1956 年首先提出了"ε-带", Chrisman[2] 引用这个概念来描述线元位置的不确定性; Caspary[3] 对"ε-带"进行了扩展, 提出了"*E*-带"; 刘大杰教授等提出了改进的"ε_m"模型[4]; 史文中、刘文宝博士基于随机过程理论提出了广义误差带"*g*-带"[5]。上述模型随给定的置信水平而定, 而置信水平的选取带有一定程度的主观性。目前在置信水平的选取上尚未取得一致的认识, 这样会导致模型多样性, 给线元不确定性的可视化表示带来不便。信息论与概率论相结合可以为不确定性问题的研究提供新的途径。文献[6]研究了几种概率分布下的不确定区间, 文献[7]引入信息熵概念, 提出了熵意义下的"ε-带"。本文进一步引入联合熵的概念, 提出了度量线元位置不确定性的误差熵带模型, 所提模型不是由置信水平确定, 而是按联合熵惟一确定。

1　信息熵与联合熵

1.1　信息熵

熵是杂乱无章、不平衡、不确定等无序状态的度量。熵的概念已经推广应用到许多知识领域, 在热力学中用来度量热状态的不平衡程度, 在信息论中表示信源的平均不确定性度量。地理信息的采集和处理过程与信息的传输过程极为相似[7], 因此可以引入信息熵的理论来研究地理信息的不确定性问题。

对于一维连续变量 X, 设其概率密度为 $P(x)$, 则信息熵定义为:

$$H(X)=-\int_R P(x)\log P(x)\mathrm{d}x$$

其中熵的单位由对数的底决定, 为了计算的方便, 常取以 e 为底。

1.2　联合熵

N 维连续矢量 $\boldsymbol{X}=[X_1, X_2, \cdots, X_n]^{\mathrm{T}}$ 的联合熵为:

$$H(\boldsymbol{X})=-\int_R\cdots\int_R P(x_1, x_2, \cdots, x_n)\cdot \log P(x_1, x_2, \cdots, x_n)\mathrm{d}x_1\cdots\mathrm{d}x_n$$

假如 N 维连续随机矢量服从正态分布, 各变量的均值 $m_r(r=1, 2, \cdots, n)$的相关中心矩为:

$$R_{ij}=E[(X_i-m_i)(X_j-m_j)]$$

其中, R_{ij}构成一个 $n\times n$ 矩阵 $\boldsymbol{R}$, $i, j=1, 2, \cdots, n$; $R_{ii}=\sigma_i^2$, $R_{ij}=R_{ji}$, $|\boldsymbol{R}|$ 代表矩阵的行列式值, 且不为零, 它的逆矩阵 $\boldsymbol{r}=\boldsymbol{R}^{-1}$。令它的第 i 行第 j 列的元素是 r_{ij}, N 维连续矢量的概率密度为:

收稿日期: 2002-02-05。
项目来源: 国家自然科学基金资助项目(46071068)。

$$P(\boldsymbol{X})=\frac{1}{(2\pi)^{\frac{n}{2}}\mid \boldsymbol{R}\mid^{\frac{1}{2}}}\exp\left[-\frac{1}{2}\sum_i\sum_j r_{ij}(x_i-m_i)(x_j-m_j)\right]$$

$$H(\boldsymbol{X})=-\int_R\cdots\int_R\frac{1}{(2\pi)^{\frac{n}{2}}\mid \boldsymbol{R}\mid^{\frac{1}{2}}}\exp\left[-\frac{1}{2}\sum_i\sum_j r_{ij}(x_i-m_i)(x_j-m_j)\right]\cdot$$

$$\left[\log\frac{1}{(2\pi)^{\frac{n}{2}}\mid \boldsymbol{R}\mid^{\frac{1}{2}}}-\frac{\log e}{2}\sum_i\sum_j r_{ij}(x_i-m_i)(x_j-m_j)\right]\mathrm{d}x_1\mathrm{d}x_2\cdots\mathrm{d}x_n$$

对上式进行积分计算[8], 有:

$$H(\boldsymbol{X})=\log\left[(2\pi)^{\frac{n}{2}}\mid \boldsymbol{R}\mid^{\frac{1}{2}}\right]+\frac{\log e}{2}\sum_i\sum_j r_{ij}R_{ji}\qquad(1)$$

由于 $\boldsymbol{r}$ 和 $\boldsymbol{R}$ 互为逆矩阵, 且都是对称矩阵, 有:

$$\sum_j r_{ij}R_{ji}=1$$

$$\sum_i\left(\sum_i r_{ij}R_{ji}\right)=n$$

故有:

$$H(\boldsymbol{X})=\frac{1}{2}\log\mid \boldsymbol{R}\mid+\frac{n}{2}\log(2\pi e)\qquad(2)$$

如果各变量之间相互独立, 有:

$$\mid \boldsymbol{R}\mid=\prod_{r=1}^{n}\sigma_r^2$$

公式(2)变成:

$$H(\boldsymbol{X})=\frac{n}{2}\log(2\pi e)+\frac{1}{2}\sum_{r=1}^{n}\log\sigma_r^2$$

设 $\boldsymbol{X}$ 是二维连续随机矢量, 则 $\boldsymbol{X}=[x, y]^{\mathrm{T}}$ 的联合熵为:

$$H(\boldsymbol{X})=\frac{1}{2}\log\sigma_x^2\sigma_y^2(1-\rho^2)+\log(2\pi e)\qquad(3)$$

如果随机变量 x、y 相互独立, 则随机矢量 $\boldsymbol{X}$ 的联合熵为:

$$H(\boldsymbol{X})=\frac{1}{2}\log\sigma_x^2\sigma_y^2+\log(2\pi e)$$

2　误差熵带的确定

2.1　二维随机点的误差熵圆的确定

设二维连续随机矢量 $\boldsymbol{X}=[x, y]^{\mathrm{T}}$ 服从均匀分布, 它的概率密度为:

$$P(x,y)=\begin{cases}1/\pi r^2, 0\leqslant\sqrt{x^2+y^2}\leqslant r\\ 0,\quad \sqrt{x^2+y^2}>r\end{cases}$$

则它的联合熵为:

$$H(\boldsymbol{X})=-\iint_{R^2}\frac{1}{\pi r^2}\log\frac{1}{\pi r^2}\mathrm{d}x\mathrm{d}y=\log\pi r^2\qquad(4)$$

按照确定一维随机变量误差熵的基本思想[9], 根据上式可确定二维随机点的熵不确定区域半径:

$$r_E=\sqrt{\frac{e^{H(X)}}{\pi}}\qquad(5)$$

将正态分布的联合熵代入, 即将式(3)代入, 则熵不确定区域半径为:

$$r_{EN}=\sqrt{\frac{\sigma_x\sigma_y\sqrt{(1-\rho^2)}(2\pi e)}{\pi}}=\sqrt{2e\sqrt{(1-\rho^2)}\sigma_x\sigma_y}\qquad(6)$$

其中, r_{EN}代表正态分布的熵不确定区域半径。

如果随机线元 Z_1Z_2 两端点的误差相互独立, 具有相同的方差, 且各向同性, 即 $\sigma_{x_1}^2=\sigma_{y_1}^2=\sigma_{x_2}^2=\sigma_{y_2}^2=\sigma^2$, $\sigma_{x_1y_1}=\sigma_{y_1x_1}=\sigma_{x_2y_2}=\sigma_{y_2x_2}=0$, 则式(12)简化为:

$$r_{EN}=\sqrt{2e}\sigma\qquad(7)$$

将 r_{EN}与标准差之比定义为熵系数 k, 有:

$$k=\frac{r_{EN}}{\sigma}=\sqrt{2e}=2.332$$
$$r_{EN}=k\sigma\qquad(8)$$

这里把半径为 $\sqrt{2e}\sigma$ 的圆定义为误差熵圆。在误差熵圆内集中了二维随机变量的主要的不确定性信息, 它是二维随机变量可能出现的基本区域。

2.2　线元的误差熵带

2.2.1　误差带(*E*-带)模型

人们在"ε-带"的基础上, 通过进一步研究提出了"*E*-带"。"ε-带"与"*E*-带"的差别在于:"ε-带"认为无论在何处带的宽度都是相等的, 而"*E*-带"是中间窄两端宽的哑铃型。

在线段 Z_1Z_2 上的任意一点 Z_i 在 x 和 y 方向上的方差分别为 $\sigma_{x_i}^2$、$\sigma_{y_i}^2$, 线段 Z_1Z_2 上的任意一点的坐标按下式计算:

$$X_i=(1-r)X_1+rX_2$$
$$Y_i=(1-r)Y_1+rY_2\qquad(9)$$

式中, $r=S_i/S$, 且 $0<r<1$, S 为Z_1Z_2 的长度, S_i 为Z_1Z_2 的长度; (X_1, Y_1)和(X_2, Y_2)分别为 Z_1、Z_2 两点的坐标。按照"*E*-带"的假定条件: 两端点的误差相互独立, 具有相同的方差和协方差, 且各向同性, 即有 $\sigma_{Z_1}^2=\sigma_{Z_2}^2$且 $\sigma_X^2=\sigma_Y^2=\sigma^2$。

根据误差传播定理, 有:

$$\sigma_{X_i}=\sigma_{Y_i}=\sqrt{1-2r+2r^2}\,\sigma\qquad(10)$$

Z_i 点的均方差为:

$$\mathrm{RMS}(P_i)=\sqrt{2}\,\sigma_{x_i} \tag{11}$$

误差带的形状由 $\mathrm{RMS}(P_i)$ 来决定，如图 1 所示。

2.2.2 误差熵带模型的确定

假定线段 Z_1Z_2 的两个端点 Z_1 和 Z_2 相互独立，且服从以下二维正态分布：

$$Z_1=\begin{bmatrix}x_1\\y_1\end{bmatrix}\sim N\left[\begin{bmatrix}u_1\\v_1\end{bmatrix},\begin{bmatrix}\sigma_{x_1}^2 & \sigma_{x_1y_1}\\ \sigma_{y_1x_1} & \sigma_{y_1}^2\end{bmatrix}\right]$$

$$Z_2=\begin{bmatrix}x_2\\y_2\end{bmatrix}\sim N\left[\begin{bmatrix}u_2\\v_2\end{bmatrix},\begin{bmatrix}\sigma_{x_1}^2 & \sigma_{x_1y_1}\\ \sigma_{y_1x_1} & \sigma_{y_1}^2\end{bmatrix}\right]$$

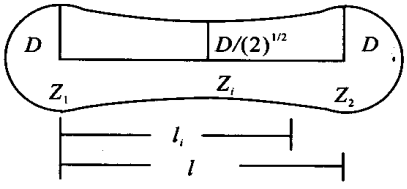

图 1 误差带（E-带）

Fig. 1 Error Band (E-band)

假定 $\sigma_{x_1}^2=\sigma_{y_1}^2=\sigma^2$ 以及 $\sigma_{x_1y_1}=\sigma_{y_1x_1}=0$，则服从如下正态分布[9]：

$$Z_i=\begin{bmatrix}x_i\\y_i\end{bmatrix}\sim N\left[\begin{bmatrix}(1-r)u_1+ru_2\\(1-r)v_1+rv_2\end{bmatrix},\left[(1-r)^2+r^2\right]\begin{bmatrix}\sigma_{x_1}^2 & 0\\ 0 & \sigma_{x_1}^2\end{bmatrix}\right]$$

根据公式(7)，线段上任意一点的误差熵圆半径为：

$$r_{EN}=\sqrt{2\mathrm{e}}\,\sigma_{x_i} \tag{12}$$

将式(10)代入式(12)，有：

$$r_{EN}=\sqrt{2\mathrm{e}\left((1-r)^2+r^2\right)}\,\sigma \tag{13}$$

把线段上的无数个误差熵圆重叠构成的区域定义为误差熵带，每一个误差熵圆的半径均为 $\sqrt{2\mathrm{e}\left((1-r)^2+r^2\right)}\,\sigma$。

同“E-带”一样，误差熵带也是中间窄两端宽。当 $r=1/2$ 时，误差熵带最窄。根据式(13)可计算最小带宽：

$$(r_{EN})_{\min}=\sqrt{\mathrm{e}}\,\sigma$$

根据式(11)和式(12)，可计算相同条件下误差熵带与“E-带”的带宽之比为：

$$\frac{r_{(EN)i}}{\mathrm{RMS}(P_i)}=\frac{\sqrt{2\mathrm{e}}\,\sigma_{x_i}}{\sqrt{2}\,\sigma_{x_i}}=\sqrt{\mathrm{e}} \tag{14}$$

误差熵带的面积为：

$$A=2\pi\mathrm{e}\sigma^2+2l\sigma\sqrt{2\mathrm{e}}\int_0^1\sqrt{1-2r+2r^2}\,\mathrm{d}r=$$

$$\pi\mathrm{e}\sigma_p^2+2l\sigma_p\sqrt{\mathrm{e}}\int_0^1\sqrt{1-2r+2r^2}\,\mathrm{d}r=$$

$$\sqrt{\mathrm{e}}\left(\sqrt{\mathrm{e}}\pi\,\sigma_p^2+2l\sigma_p\int_0^1\sqrt{1-2r+2r^2}\,\mathrm{d}r\right)=$$

$$\sqrt{\mathrm{e}}\left(\pi\,\sigma_p^2+2l\sigma_p\int_0^1\sqrt{1-2r+2r^2}\,\mathrm{d}r\right)=\sqrt{\mathrm{e}}\,A_1 \tag{15}$$

式中，σ_p 为位置中误差，$\sigma_p=\sqrt{2}\,\sigma_{x_1}=\sqrt{2}\,\sigma$；$A_1$ 为相同条件下“E-带”的面积。

误差熵带的形状如图 2 所示。

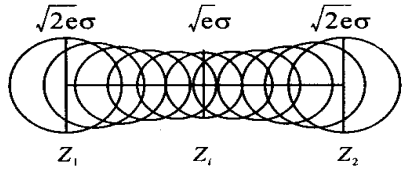

图 2 误差熵带

Fig. 2 Error Entropy Band

3 随机线元落入熵误差带内的概率分析

随机线元的误差熵带分为 A、B、C 三部分，如图 3 所示。根据文献[11]的分析方法，误差熵带划分为相互独立的沿线元方向的($A+C$)区域和垂直于线元方向的 B 区域。

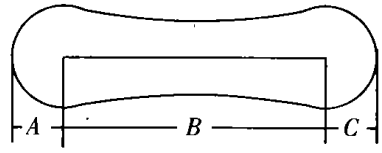

图 3 误差熵带的 A、B、C 区域

Fig. 3 A, B and C Region of Error Entropy Band

由概率的可加性，随机线元落入误差熵带的概率为：

$$P((A+C)\cup B)=P(A+C)+P(B)-P((A+C)B) \tag{16}$$

由于随机线元 Z_1Z_2 在沿线元方向上的随机偏差与垂直于线元方向的随机摆动是互不干涉的，即区域($A+C$)和区域 B 的偏差是相互独立的，根据概率论，有：

$$P((A+C)B)=P(A+C)P(B)$$

将 $P((A+C)\cup B)$ 记为 P_k，$P(A+C)$ 记为 P_s，$P(B)$ 记为 P_u，则式(16)变为：

$$P_k=P_s+P_u-P_sP_u \tag{17}$$

根据前面的假设，随机线元 Z_1Z_2 两端点的误差相互独立、方差相同，且各向同性，则两端点的概率密度为：

$$f(x,y)=\frac{1}{2\pi\sigma^2}\exp\left[-\frac{1}{2\sigma^2}(x^2+y^2)\right]$$

根据文献[10],线元上任意点在线段垂直方向上的概率密度为:

$$f(y')=\frac{1}{\sqrt{2\pi}\,\sigma_{y'}}\exp\left[-\frac{1}{2\sigma_{y'}^2}(y-v_{y'})^2\right]$$

式中, $y'=-\sin\theta((1-r)X_1+rX_2)+\cos\theta((1-r)Y_1+rY_2)$; $v_{y'}=\sin\theta((1-r)u_1+ru_2)+\cos\theta((1-r)v_1+rv_2)$; $\sigma_{y'}^2=((1-r)^2+r^2)\sigma^2$。

A 和 C 的面积为误差熵圆的一半, $(A+C)$ 正好是误差熵圆的区域, 落入 $(A+C)$ 的概率为:

$$P_s=P(A+C)=\iint_{x^2+y^2\leqslant r_{EN}^2} f(x,y)\,\mathrm{d}x\,\mathrm{d}y=$$

$$\int_0^{2\pi}\mathrm{d}\theta\int_0^{k\sigma}\frac{1}{2\pi\sigma^2}\exp\left(-\frac{1}{2\sigma^2}r^2\right)r\,\mathrm{d}r=$$

$$\int_0^k\exp\left(-\frac{t^2}{2}\right)t\,\mathrm{d}t=1-\mathrm{e}^{\left(-\frac{k^2}{2}\right)}\qquad(18)$$

B 区域为随机线元在垂直于线元方向上进行平移时产生的, 随机线元落入区域 B 的概率等于所有一维截口随机变量 y' 在区间 $(v_i-k\sigma,\ v_i+k\sigma)$ 上的截面面积 $S(k)$ 对线长的积分。

随机线元截口随机变量的截面面积为:

$$S(k)=\int_{v_i-k\sigma_{y'}}^{v_i+k\sigma_{y'}}f(y')\,\mathrm{d}y'=$$

$$2\int_0^k\frac{1}{\sqrt{2\pi}}\mathrm{e}^{\left(-\frac{t^2}{2}\right)}\mathrm{d}t$$

把随机线元作为一个整体, 则 P_u 为:

$$P_u=\int_0^l S(k)\,\mathrm{d}l=S(k)=2\int_0^k\frac{1}{\sqrt{2\pi}}\mathrm{e}^{\left(-\frac{t^2}{2}\right)}\mathrm{d}t$$

对于线元的误差熵带, 将 $k=\sqrt{2\mathrm{e}}=2.332$ 代入式(17)、式(18)、式(19), 可计算 $P_k=0.998\,7$, $P_s=1-\mathrm{e}^{-\mathrm{e}}=0.934\,1$, $P_u=2\int_0^{\sqrt{2\mathrm{e}}}\frac{1}{\sqrt{2\pi}}\mathrm{e}^{\left(-\frac{t^2}{2}\right)}\mathrm{d}t=0.980\,3$。

表 1 比较了随机线元落入不同误差带的概率。

4　结　论

1) 通过引入联合熵理论, 提出了线元的误差熵带模型。同"E-带"一样, 误差熵带的形状也是呈中间窄两端宽的哑铃型。误差熵带与"E-带"的带宽之比为$\sqrt{\mathrm{e}}$倍, 面积之比近似为$\sqrt{\mathrm{e}}$倍; 随机线元落入误差熵带的概率为 99.87%。

2) 误差熵带能惟一确定, 与置信水平的选取无关。实际上, 不管随机矢量服从何种分布, 只要它的联合熵存在, 都可以确定它的不确定范围。在误差熵带内集中了随机线元的主要不确定性信息, 它是随机线元不确定性可能出现的基本区域, 它是一个平均意义下的指标, 与传统的误差带指标有本质的区别。

表 1　随机线元落入不同误差带的概率

Tab. 1　Probability of Falling into Different Error Bands

误差带名称	k 值	落入误差带的概率/%
标准差带	1.0	80.77
或然误差	0.551 5	50
地图 ε 精度	1.24	90
误差熵带	2.332	99.87
极限误差	3.0	99.99

3) 线元的不确定性存在于有限范围内, 不能一概运用概率论的无限区域概念来讨论, 应从有限范围入手。信息论与概率论相结合可以为二维、三维乃至多维线元和面元不确定性问题的解决提供有效途径。

参　考　文　献

1　Perkal J. On Epsilon Length. Bullitin de I' Academic Polonasise Des Sciences, 1956(4): 399~403

2　Chrisman N R. A Theory of Cartographic Error and Its Measurement in Digital Databases. Auto-Carto5, 1982. 159~158

3　Caspasy W, Scheuring R. Error Band as Measurers of Geometrical Accuraey. EGIS' 92, Utrecht, 1992

4　刘大杰, 华　慧. GIS 线要素不确定性模型的进一步探讨. 测绘学报, 1998, 27(1): 45~49

5　史文中, 刘文宝. GIS 中线元位置不确定性的随机过程模型. 测绘学报, 1998, 27(1): 37~43

6　孙海燕. 熵与不确定区间. 武汉测绘科技大学学报, 1994, 19(1): 63~70

7　范爱民, 郭达志. 误差熵不确定模型. 测绘学报, 2001, 30(1): 48~53

8　周　炯. 信息理论基础. 北京: 人民邮电出版社, 1983

9　诺维茨基, 佐格拉夫. 测量结果误差估计. 北京: 中国计量出版社, 1990. 58~68

10　史文中. 空间数据误差处理的理论与方法. 北京: 科学出版社, 1998

11　戴洪磊. 矢量 GIS 位置不确定度量与传播的理论: [博士论文]. 武汉: 武汉测绘科技大学, 2000

作者简介: 李大军, 副教授, 博士生。现从事 GIS 不确定性理论与 GIS 应用研究。

E-mail: lidajun66@sina.com

Error Entropy Band for Linear Segments in GIS

LI Dajun[1] *GONG Jianya*[1] *XIE Gangsheng*[2] *DU Daosheng*[1]

(1 National Laboratory for Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, 129 Luoyu Road, Wuhan, China, 430079)

(2 Dep. of Surveying, East China Institute of Technology, 14 West Huancheng Road, Fuzhou, China, 344000)

Abstract: Uncertainty of spatial data in GIS can be in the aspect of position, attribution, temporary, logical relation and completeness. Among them, positional uncertainty of linear segments is an important aspect. There are uncertainty models of linear segments, such as "E-band", "g-band" and so on, in some confidence bands, and they are relevant to different confident levels, but choice of confident levels exists a certain extent subjectivity and therefore cannot be completely determined.

In this paper, on the basis of union entropy and maximum entropy theorem in information theory, an error entropy band is put forward and the comparison between error entropy band and "E-band" is given, and then the probability of falling into it is calculated. Finally, some conclusions are drawn as follows:

1) Error entropy band for linear segment is an average uncertainty measure of linear segments, and can be solely determined by union entropy and is independent of the choice of confident level. Therefore, it is a comparatively impersonal index.

2) The ratio of band width between error entropy band and E-band is $\sqrt{e}$ times, and the ratio of area is about times, and the probability of falling into them is 99.87%.

3) Information theory can provide a new approach for solving uncertainty problems in GIS.

Key words: line segments; positional uncertainty; union entropy; entropy circle; error entropy band

About the author: LI Dajun, associate professor, Ph. D candidate. His main interest is focused on uncertainty theory of spatial data and application research of GIS.

E-mail: lidajun66@sina.com