文章编号：1007-4619 (2004)09-0009-05

GIS 线元的平均熵不确定带

李大军[1,2]，龚健雅[1]，于海龙[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，湖北 武汉 430079；2. 东华理工学院 测量系，江西 抚州 344000)

摘 要： 在 GIS 线元的位置不确定性方面，国内外学者已提出了“ε-带”、“e-带”、“g-带”、“H-带”等模型，然而就应用而言，由于“ε-带”具有不变带宽，因而应用最为广泛。但是“ε-带”的宽度往往难以确定，从而限制了它的使用范围。在“H-带”的基础上，提出了根据线元的平均信息熵确定“ε-带”宽度的思想，建立了线元的平均熵不确定带，并以此作为线元位置不确定性的度量。

关键词： 线元；位置不确定性；ε-带；平均信息熵；平均熵不确定带
中图分类号： P208 **文献标识码**： A

1 引 言

空间数据的不确定性是 GIS 的一个基本理论课题。由于线元不仅是面域不确定性的基础，其本身也是 GIS 叠置分析、缓冲区分析等的基本元素，因此线元的位置不确定性是 GIS 不确定性研究的一个重点[1]。线元的位置不确定性研究是基于带的概念展开的，1956 年美国学者 Perkal 首次提出了“ε-距离”概念，1982 年 Chrisman 引入上述概念提出了度量线元位置不确定性的“ε-带”模型。在此基础上一些学者进行了扩展，Gaspary 等提出了“e-带”模型[2]；史文中、刘文宝基于随机过程理论提出了“g-带”模型[3]；刘大杰等提出了改进后的“ε_m”模型[1]；范爱民等利用误差熵确定了“ε-带”的带宽，提出了“H-带”模型[4]。分析现有的线元 ε-带模型，可分为等带宽和不等带宽两种类型。就应用而言，等带宽的“ε-带”更为实用，但是“ε-带”带度的确定一直是个难点，很多情况下往往是根据经验人为选定的，这难免会受到主观性影响。文献[4]提出了一种很有价值的思想，即根据线元端点的误差熵确定“ε-带”的宽度，这样确定的带宽是唯一的，不受置信水平的影响。然而“H-带”仅仅根据线元端点的边缘概率分布的信息熵来确定带宽，并没有考虑线元上误差分布不均匀的特点，因而得到的带宽过大且过于保守。本文对“H-带”进行了发展，提出以整个线元边缘概率分布的平均信息熵作为确定“ε-带”带宽的依据，从而建立了线元的平均熵不确定带模型。

2 “H-带”模型

2.1 信息熵

对于连续随机变量 X，设它的概率密度函数为 $p(x)$，它的信息熵 $H(X)$ 定义为

$$H(X) = -\int_{-\infty}^{\infty} p(x)\log\ p(x)\mathrm{d}x \tag{1}$$

其中熵的单位由对数的底决定。通常为了计算的方便，取 e 为底，单位为奈特(nat)。

设一维连续正态随机变量 X，它的信息熵为

$$H(X) = -\int_{-\infty}^{\infty} P(x)\ln\ P(x)\mathrm{d}x = \frac{1}{2}\ln(2\pi\mathrm{e}\,\sigma^2) \tag{2}$$

2.2 熵不确定区间

根据误差熵的定义[5]，有

$$\Delta_E = \mathrm{e}^{H(x)}/2 \tag{3}$$

其中 Δ_E 为误差熵。

将(2)式代入上式，则正态分布的误差熵为

$$\Delta_{EN} = \frac{\sqrt{2\pi\mathrm{e}}}{2}\sigma = 2.07\sigma \tag{4}$$

收稿日期：2002-04-08；**修订日期**：2002-07-10
基金项目：国家自然科学基金项目(46071068)资助。

作者简介：李大军(1965—)，男，湖南澧县人，东华理工学院测量系副教授，武汉大学博士生，主要从事空间数据的不确定性与 GIS 应用研究，发表文章 18 篇。

其中 Δ_{EN} 表示正态分布的误差熵，σ 为标准差。

误差熵与标准差的比值定义为熵系数 k_{EN}，有

$$k_{EN} = \Delta_{EN} / \sigma \tag{5}$$

故正态分布的熵系数为

$$k_{EN} = \frac{\Delta_{EN}}{\sigma} = \frac{\frac{\sqrt{2\pi e}}{2}\sigma}{\sigma} = 2.07$$

对于一维连续正态随机变量，可以取熵不确定区间$(-2.07\sigma, 2.07\sigma)$作为它的位置不确定性度量。

2.3 *H*-带

“ε-带”模型简单实用，易于理解和 GIS 中实现，但带的宽度往往难以确定。文献[4]提出了线元的误差熵不确定带（*H*-带）模型，它定义为：以一维正态随机变量的误差熵（$\Delta_{EN}=2.07\sigma$）为带宽，所形成的缓冲区域。“*H*-带”具有唯一确定，不受置信水平影响的特点。但它未考虑到线元上各点误差分布的不均匀性，故所确定的带宽欠合理。

3 平均熵不确定带的确定

3.1 线元的平均信息熵

设线元由两端点 $Z_1=[x_1, y_1]^T$，$Z_2=[x_2, y_2]^T$ 组成，假设两端点误差不相关，且 Z_1 和 Z_2 分别服从二维正态分布

$$Z_1 = \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} \sim N_2\left[\begin{bmatrix} u_1 \\ v_1 \end{bmatrix}, \begin{bmatrix} \sigma_{x_1}^2 & \sigma_{x_1y_1} \\ \sigma_{y_1x_1} & \sigma_{y_1}^2 \end{bmatrix}\right],$$

$$Z_2 = \begin{bmatrix} x_2 \\ y_2 \end{bmatrix} \sim N_2\left[\begin{bmatrix} u_2 \\ v_2 \end{bmatrix}, \begin{bmatrix} \sigma_{x_2}^2 & \sigma_{x_2y_2} \\ \sigma_{y_2x_2} & \sigma_{y_2}^2 \end{bmatrix}\right]$$

线元上任一点 Z_t 可表示为

$$Z_t = (1-t)Z_1 + tZ_2 = \begin{bmatrix} (1-t)x_1 + tx_2 \\ (1-t)y_1 + ty_2 \end{bmatrix} \tag{6}$$

由于 Z_t 与 Z_1、Z_2 存在线性关系，则 Z_t 服从如下二维正态分布

$$Z_t \sim N_2\left[\begin{bmatrix} (1-t)\mu_1 + t\mu_2) \\ (1-t)v_1 + tw_2) \end{bmatrix}, \begin{bmatrix} \sigma_x^2(t) & \sigma_{xy}(t) \\ \sigma_{yx}(t) & \sigma_y^2(t) \end{bmatrix}\right]$$

设

$$D_{Z(t)} = \begin{bmatrix} \sigma_x^2(t) & \sigma_{xy}(t) \\ \sigma_{yx}(t) & \sigma_y^2(t) \end{bmatrix}$$

$$= \begin{bmatrix} (1-t)^2\sigma_{x_1}^2 + t^2\sigma_{x_2}^2 & (1-t)^2\sigma_{x_1y_1} + t^2\sigma_{x_2y_2} \\ (1-t)^2\sigma_{x_1y_1} + t^2\sigma_{x_2y_2} & (1-t)^2\sigma_{y_1}^2 + t^2\sigma_{y_2}^2 \end{bmatrix} \tag{7}$$

将坐标系 *xoy* 旋转至坐标系 $x'oy'$，如图 1。

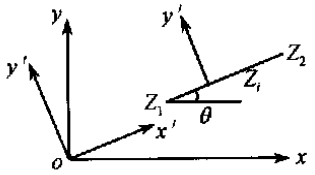

图 1 坐标系从 *xoy* 旋转至 $x'oy'$

Fig. 1 The Rotation of reference system from *xoy* to $x'oy'$

$Z_t'=[x_t', y_t']^T$ 为 Z_t 点在新坐标系 $x'oy$ 下的坐标，它们存在如下线性关系

$$\begin{bmatrix} x'_t \\ y'_t \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x_t \\ y_t \end{bmatrix} \tag{8}$$

由于 Z_t 服从二维正态分布，则 Z'_t 也服从二维正态分布，它的方差-协方差矩阵为

$$D_{Z'(t)} = \begin{bmatrix} \sigma_{x'}^2(t) & \sigma_{x'y'}(t) \\ \sigma_{y'x'}(t) & \sigma_{y'}^2(t) \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} D_{Z(t)} \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \tag{9}$$

将(7)式代入上式，经整理有

$$\sigma_{y'}^2(t) = (1-t)^2[\sin^2\theta\sigma_{x_1}^2 - \sin 2\theta\sigma_{x_1y_1} + \cos^2\theta\sigma_{y_1}^2] + t^2[\sin^2\theta\sigma_{x_2}^2 - \sin 2\theta\sigma_{x_2y_2} + \cos^2\theta\sigma_{y_2}^2] \tag{10}$$

根据文献[6]，Z_t 点沿 y' 方向上的边缘概率分布的密度函数为

$$f(y') = \int_{-\infty}^{+\infty} f(x', y')\mathrm{d}x' = \frac{1}{(2\pi\sigma_{y'}^2)^{1/2}} \exp[-(y' - v_{y'})^2/2\sigma_{y'}^2] \tag{11}$$

其中

$$y' = -\sin\theta[(1-t)x_1 + tx_2] + \cos\theta[(1-t)y_1 + ty_2]$$

$$v_{y'} = -\sin\theta[(1-t)u_1 + ru_2] + \cos\theta[(1-t)v_1 + tw_2]$$

线元上任一点 Z_t 边缘概率分布的信息熵为

$$H(y'_t) = -\int_{-\infty}^{+\infty} f(y')\ln f(y')\mathrm{d}y' = \frac{1}{2}\ln(2\pi e\,\sigma_{y'_t}^2) \tag{12}$$

由于线元上任意一点的方差 $\sigma_{y'_t}^2$ 是参数 t 的函数，故 $H(y_t')$ 在闭区间$[0, 1]$内是连续变化的。根据闭区间上连续函数的性质，则 $H(y_t')$ 必有最大值和最小值，设它们分别记为 $H_{\max}$ 和 $H_{\min}$。利用后面表 1 中的数据，我们绘出了线元在 4 种情况下 $H(y_t')$ 随 t 变化的曲线图，同时标出了 $H(y_t')$ 取最大值和最小值时的位置，如图 2。

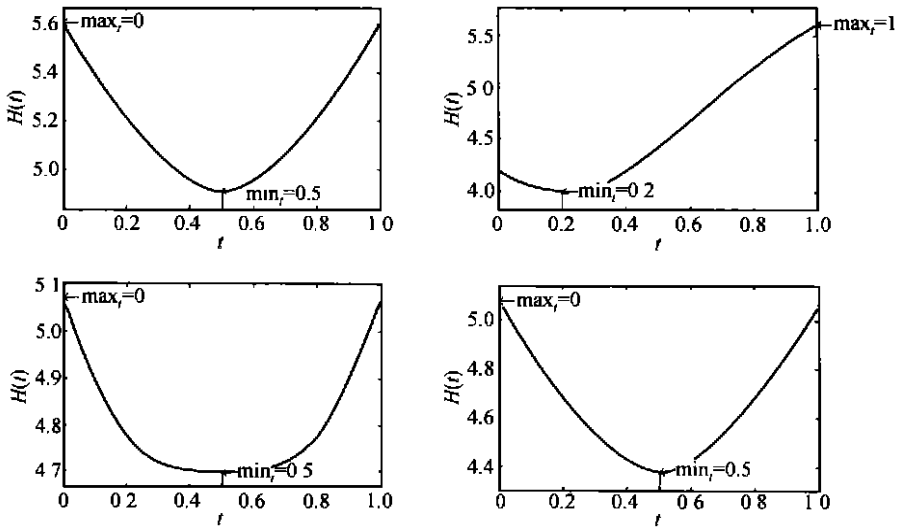

图 2　$H(y_t')$在 4 种情况下的曲线图

Fig. 2　The graph of $H(y_t')$ in four cases

考虑到线元上各点的熵值是连续变化的,为了得到线元位置不确定性的整体度量,我们求整个线元上的平均信息熵,有

$$\overline{H}=\int_0^1 H(y_t')\mathrm{d}t=\ln(\sqrt{2\pi\mathrm{e}})+\frac{1}{2}\int_0^1 \ln\sigma_{y't}^2\mathrm{d}t \tag{13}$$

3.2　平均熵不确定带的带宽确定

根据式(3),线元的平均误差熵 $\overline{\Delta}$ 为

$$\overline{\Delta}=\frac{1}{2}\mathrm{e}^{\overline{H}} \tag{14}$$

我们把线元的平均误差熵 $\overline{\Delta}$ 作为"ε-带"的带宽,所形成的缓冲区称为平均熵不确定带。

下面讨论平均熵不确定带同"H-带"的关系。

假定线元处于水平位置,两端点具有相同的方差、协方差且各向同性,即 $\theta=0$, $\sigma_{x_i}^2=\sigma_{y_i}^2=\sigma^2$, $i=1,2$。根据式(10)、(12)和式(3),有

$$H_{\max}=\ln(\sqrt{2\mathrm{e}\pi}\sigma),\ \Delta_{\max}=\frac{1}{2}\sqrt{2\mathrm{e}\pi}\sigma=2.07\sigma$$

$$H_{\min}=\ln(\sqrt{\mathrm{e}\pi}\sigma),$$

$$\Delta_{\min}=\frac{1}{2}\sqrt{\mathrm{e}\pi}\sigma=1.46\sigma=\Delta_{\max}/\sqrt{2}$$

由此可见,"H-带"就是这种情况下的最大熵带,带宽为最大误差熵。"H-带"的带宽是依线元端点边缘分布的熵确定的,而本文所提的平均熵不确定带的带宽是根据整个线元上的平均熵确定的,显然本文的做法要合理一些。

4　算例分析

参照文献[7]的起算数据,表 1 中给出了线元 Z_1Z_2 在 4 种情况下的坐标、方差和协方差数据。图 3(a)绘出了它们的"g-带"可视化图形,其中第一种情况是两端点方差相同、协方差为 0 时的情形;第二种情况是在两端点方差不相等、协方差为 0 时的情形;第三种情况是两端点方差相同、协方差的绝对值相等的情形、第四种情况是两端点方差、协方差均相同时情形。从图 2 中可以看出"g-带"的形状随两端点的误差状态而定。

根据本文所提的平均熵不确定带模型,图 3(b)绘出了在上述 4 种情况下对应的可视化图形。从图中可以发现不同形状的"g-带"图形转变为相应的"ε-带"图形。通过计算,线元在 4 种情况下的平均熵不确定带的带宽分别为 6.636cm, 4.937cm, 7.250cm, 7.250cm。平均熵不确定带的带宽由整个线元边缘概率分布的平均信息熵决定,而线元的平均信息熵与线元两端点的方差、协方差以及线段的倾角有关,这可以从公式(10)中看出。当线段处于水平位置时,平均熵不确定带的带宽与协方差无关,故线元 3 和线元 4 的平均熵不确定带的带宽均为 7.25cm。

表 2 中给出了线元 Z_1Z_2 在 4 种情况下的最大、最小和平均信息熵及其对应带宽的计算结果。从表 2 中可以看出平均熵不确定带介于最大熵不确定带

与最小熵不确定带之间。以线元 1 为例，最大熵带的带宽为 $\varepsilon=8.224=2.07\sigma$，其中 $\sigma^2=15.84$；最小熵带的带宽为 $\varepsilon=5.815=1.46\sigma$；而平均熵带的带宽只为 $\varepsilon=6.636=1.67\sigma$。同时计算结果也表明：第一种情况下的最大熵不确定带就是文献[4]中的“*H*-带”，带宽为 2.07σ。

表 1 已知数据

Table 1 Known data

线元号	Z_1					Z_2				
	x_1/m	y_1/m	$\sigma^2_{x_1}$/cm	$\sigma^2_{y_1}$/cm	$\sigma_{x_1y_1}$/cm	x_2/m	y_2/m	$\sigma^2_{x_2}$/cm	$\sigma^2_{y_2}$/cm	$\sigma_{x_2y_2}$/cm
1	300.00	100.00	15.84	15.84	0.00	300.00	120.00	15.84	15.84	0.00
2	500.00	100.00	3.96	3.96	0.00	500.00	120.00	15.84	15.84	0.00
3	500.00	100.00	18.91	8.71	8.83	500.00	120.00	18.91	8.71	−8.83
4	500.00	100.00	18.91	8.71	8.83	500.00	120.00	18.91	8.71	8.83

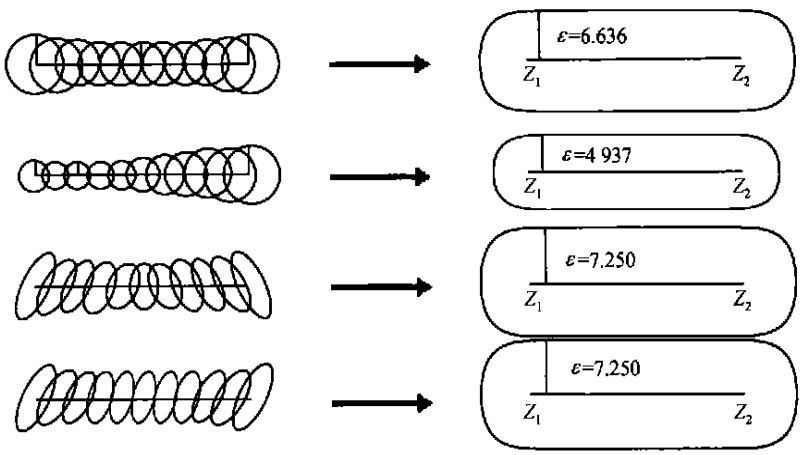

图 3 (a) *g*-带的可视化图；(b) 平均熵不确定带的可视化图

Fig. 3 (a)Visualization of *g*-band in four states; (b)Visualization of average entropy band in four states

表 2 最大、最小和平均熵带在 4 种情形下的带宽

Table 2 The band-width of the band of maximum, minimum and average entropy in four cases

线号	最小值			平均值			最大值		
	t	H/ *nat*	ε/cm	t	H/ *nat*	ε/cm	t	H/ *nat*	ε/cm
1	0.5	2.454	5.815	0.775	2.586	6.636	1	2.800	8.224
2	0.2	1.995	3.678	0.558	2.290	4.937	1	2.800	8.224
3	0.5	2.542	6.354	0.775	2.674	7.250	1	2.889	8.986
4	0.5	2.542	6.354	0.775	2.674	7.250	1	2.889	8.986

5 结论与建议

通过本文的研究，得出了以下的结论：

(1) 本文引入了信息熵的理论，提出以整个线元边缘概率分布的平均熵作为确定带宽的依据，建立了平均熵不确定模型。该模型是对“*H*-带”的进一步发展，体现在带宽的确定上不是仅仅根据端点的信息熵，而是根据整个线元上平均信息熵而定。通过本文的理论推导和实例计算，证明了“*H*-带”就是线元在第一种情况下的最大熵不确定带。

(2) 平均熵不确定模型本质上是一种平均熵意

义下的“ε-带”。它既满足传统“ε-带”等带宽的要求,又考虑到了线元上误差分布不均匀的特点,是对线元位置不确定性的一种整体度量。虽然在描述的精细程度方面不及“g-带”,但它着重于把握线元位置不确定性的总体态势,不拘泥于局部细节,符合系统论所倡导的思想;同时由于本文所提的模型具有“ε-带”的性质,因而比较适合于应用。

另外笔者提出如下建议:(a)在研究空间数据的不确定性问题时,有必要强调系统论的观点,从整体上把握空间数据的不确定性,以克服目前研究中的“只见树木,不见森林”的状况。(b)空间数据的不确定性问题是一个难度很大的基本理论课题,空间数据的不确定性的内涵极为丰富。尽管目前已经取得了可喜的进展,但众多的研究大都是在各自的方面独立展开的,至今也未能建立起空间不确定性的统一理论框架。人们还不十分清楚不确定性的本质、各种不确定性之间的相互关系、不确定性的传播机理等,因而还未找到从整体上处理各种不确定性的有效方法。因此笔者认为引入熵理论可能会对建立空间数据不确定性的理论框架提供有益的帮助。

参 考 文 献 (References)

[1] Liu Dajie, *et at*. Accuracy Analysis and Quality Control of Spatial Data in GIS [M]. Shanghai: Publishing House of Shanghai Technology Literature, 1999. [刘大杰等. GIS 空间数据的精度分析与质量控制[M]. 上海:上海科学技术文献出版社,1999.]

[2] Caspary W, Scheuring R. Error-band as Measurers of Geographic Accuray [A]. In Proceeding of EGIS'92 [C], 1992.

[3] Shi W Z, Liu W B. A Stochastic Process-based Model for the Positional Error of Line Segments in GIS [J]. *INT. J. Geographical Information Science*, 2000, **14**(1): 51—66.

[4] Fan A M, Guo D Z. The Uncertainty Band Model of Error Entropy [J]. *Acta Geodatica et Cartographica Sinica*, 2001, **30**(1): 48—53. [范爱民,郭达志. 误差熵不确定带模型[J]. 测绘学报, 2001, **30**(1): 48—53.]

[5] Kang Guangyong, Hu Naibin. Error Estimation of Measurement Result [M]. Beijing: Publishing House of China Computation and Measurement, 1990. [波·弗·诺维茨基,伊·阿·佐格拉夫著,康广庸胡乃滨译. 北京:测量结果误差估计. 中国计量出版社,1990.]

[6] Shi Wenzhong. Theory and Methods for Handling Errors in Spatial Data [M]. Beijing: Science Press, 1989. [史文中. 空间数据误差处理的理论与方法[M]. 北京:科学出版社,1989.]

[7] Dai Honglei. Theory and Method on Measurement and Propagation of Positional Uncertainty in Vector GIS [D]. Wuhan: Wuhan Technical University of Surveying and Mapping, 2000. [戴洪磊. 矢量 GIS 位置不确定度量与传播的理论[D.]武汉:武汉测绘科技大学, 2000.]

[8] Sun H Y. Entropy and Interval of Uncertainty [J]. *Journal of Wuhan Technical University of Surveying and Mapping*, 1994, **19**(1): 63—70. [孙海燕. 熵与不确定度区间[J]. 武汉测绘科学大学学报, 1994, **19**(1): 63—70.]

The Band of Average Entropy for Line Segments in GIS

LI Da-jun[1,2], GONG Jian-ya[1], YU Hai-long[2], DU Dao-sheng[1]

(1. *National Labaratory for Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan* 430079, *China*; 2. *Department of Surveying, East China Institute of Technology, Fuzhou* 344000, *China*)

Abstract: Spatial data is one of the fundamental parts of GIS. Uncertainty of spatial data can directly affect the quality of digital products and reliability of GIS-based decision-making and is regarded as one of the fundamental theoretical research issues. In recent years, positional uncertainty of line segments is a research focus. Several models have been presented by other scholars, for example ε-band, e-band, g-band, and H-band. Among these existing models, the ε-band has been widely used due to its advantage of fixed band-width. But it is difficult to determine the band-width, which limits its further application. In this paper, we expand the model of H-band in literature [4] and present an uncertainty band of average entropy based on average information entropy of the whole line segment. The new model absorbs the strongpoint of ε-band and at the same time takes the asymmetry of error distribution in line segment into consideration. It is a comparatively suitable measurement index of the positional uncertainty of line segments.

Key words: line segment; positional uncertainty; ε-band; average information entropy; band of average entropy