文章编号: 1671-8860(2006)01-0082-04 文献标志码: A

GIS中空间数据不确定性的混合熵模型研究

史玉峰[1,2,3] 史文中[3] 靳奉祥[4]

(1 山东理工大学建筑工程学院, 淄博市张店区张周路12号, 255049)
(2 武汉大学地球空间环境与大地测量教育部重点实验室, 武汉市珞喻路129号, 430079)
(3 香港理工大学土地测量与地理资讯学系地球资讯科技研究中心, 香港九龙红磡)
(4 山东科技大学校长办公室, 青岛市经济技术开发区前湾港路579号, 266510)

摘 要: 基于信息理论和模糊集合理论, 针对GIS中部分空间数据既具有随机性又具有模糊性的特点, 建立了空间数据不确定性的混合熵模型。以GIS中线元不确定性为例, 讨论了线元不确定性的统计熵、模糊熵和混合熵估计方法, 并针对特例给出了线元不确定性的熵带分布。

关键词: 不确定性; 空间数据; 混合熵; 统计熵; 模糊熵; 线元
中图法分类号: P208

空间数据的不确定性泛指空间数据所具有的误差、不精确性、模糊性和含混性[1], 一般可分为位置不确定性、属性不确定性、时域不确定性、逻辑不一致性和数据不完整性, 而空间数据特征的不确定性是当今GIS学术界研究的主要问题之一。目前, 空间数据不确定性的研究主要集中在位置不确定性的产生、模型和传播方面。基于数理统计理论, 许多研究人员研究了GIS中点位、线元和面元的不确定性模型[1-10]。严格地说, 基于信息熵的不确定性模型仍然属于统计不确定性, 因为误差熵是基于概率密度函数推导出来的, 这里的误差熵模型是一种基于统计理论的不确定性模型。

由于GIS中的一些空间数据常常既具有随机性, 又具有模糊性, 同时, 空间数据的采集和处理过程与信息传输模型极为相似, 因此, 基于信息论和模糊集合理论, 本文分别建立了空间数据位置不确定性的统计熵模型和属性不确定性的模糊熵模型, 并将随机性与模糊性综合起来考虑, 建立了空间数据不确定性的混合熵模型。对于非明确定义的地理现象, 由于空间数据的模糊性和随机性是以连续体的形式存在的, 因此, 混合熵更能体现其不确定性。

1 混合熵模型

信息熵是信息论中的重要概念, 它表示信源的平均不确定性[11]。对于取值离散的样本空间(信源), $[X \cdot P]: \{X: a_1, a_2, \cdots, a_n; P(X): p_1, p_2, \cdots, p_n\}$, p_i 为事件 a_i 出现的概率, 且 $p_i \geqslant 0$, $\sum_{i=1}^{n} p_i = 1$, 则

$$H_s(X) = E[-\lg p_i] = -\sum_{i=1}^{n} p_i \lg p_i \quad (1)$$

为信源的信息熵。

模糊熵是模糊集合理论中度量模糊子集模糊不确定性的测度之一。许多学者对模糊熵的建立方法进行了研究, 提出了多种模糊熵模型[12-14]。在不考虑概率分布函数的情况下, Deluca和Termini提出的模糊熵模型[12]为:

$$H_f(A) = -k\sum_{i=1}^{n}\{\mu_A(x_i)\lg\mu_A(x_i) + (1-\mu_A(x_i))\lg(1-\mu_A(x_i))\} \quad (2)$$

式中, k 是大于0的常数, 常取 $k=1$。

在现实问题中, 一个系统中可能既含有随机不确定性, 又含有模糊不确定性。当这两种不确定性同时存在时, Deluca和Termini提出了一种

收稿日期: 2005-10-16。
项目来源: 香港特区资助项目(1.34.37.A222); 山东省基础地理信息与数字化技术重点实验室开放研究基金资助项目(SD2003-10)。

测度来度量系统的统一不确定性, 称之为总熵(total entropy)[11], 并定义为:

$$H_{total} = H_r(A) + H_f(A) \quad (3)$$

式中, $H_r(A)$ 为系统的统计熵; $H_f(A)$ 为模糊熵。

式(3)中的统计熵和模糊熵分别是在概率空间和模糊空间中计算的, 它们之间没有建立有机的联系。但实际问题中的随机性和模糊性常常是联系在一起的, 如 GIS 中土壤的分界线等[3, 5], 其位置具有随机性, 属性具有模糊性, 因此, 为了统一考虑由随机性和模糊性所引起的总的不确定性, 应建立一个由随机空间(R)和模糊空间(F)所确定的共同积空间 $R \times F$, 这样, 表征系统随机不确定性和模糊不确定性的总分布函数就是如下一个映射:

$$f: R \times F \rightarrow [0, 1] \quad (4)$$

本文将随机性和模糊性所产生的总的不确定性用混合熵 $H_h(R, F)$ 来表示, 它应满足如下基本条件: 当模糊性消失时, $H_h(R, F)$ 退化为统计熵模型; 当随机性不存在时, $H_h(R, F)$ 退化为模糊熵模型。

考虑上述基本条件, 根据 Shannon 熵和 De-luca-Termini 模糊熵模型, 可定义离散混合熵 $H_h(R, F)$ 为:

$$H_h(R, F) = -\sum_{i=1}^{n}\{p_i\mu_i\lg(p_i\mu_i) + p_i(1-\mu_i)\lg(p_i(1-\mu_i))\} \quad (5)$$

式中, 各参数意义同前。

对式(5)进行分解, 可得:

$$H_h(R, F) = -\sum_{i=1}^{n}p_i\lg p_i - \sum_{i=1}^{n}\{\mu\lg\mu_i + (1-\mu_i)\lg(1-\mu_i)\} + \sum_{i=1}^{n}(1-p_i)\{\mu_i\lg\mu_i + (1-\mu_i)\lg(1-\mu_i)\} = H_s + H_f - H_{sf} \quad (6)$$

显然, 式(6)右端的 H_s、H_f 分别是由式(1)和式(2)所定义的系统的统计熵和模糊熵。定义 H_{sf} 为随机性与模糊性的交叉熵, 它可以看作随机性和模糊性的交叉影响, 即

$$H_{sf} = -\sum_{i=1}^{n}(1-p_i)\{\mu_i\lg\mu_i + (1-\mu_i)\lg(1-\mu_i)\} \quad (7)$$

由式(6)可以看出, 当系统为一明晰集合, 即 $\mu(x) = \{0, 1\}$, 式(6)就退化为统计熵模型; 当系统为一模糊集合, 即 $p(x) = 1$ 时, 式(6)就退化为模糊熵模型。

对于连续随机分布和连续模糊分布的变量, 定义混合熵为:

$$H_h(R, F) = -\int_{-\infty}^{+\infty}\{p(x)\mu(x)\lg\{p(x)\mu(x)\} + p(x)(1-\mu(x))\lg\{p(x)(1-\mu(x))\}\}dx \quad (8)$$

式中, $\mu(x)$ 和 $p(x)$ 分别为连续分布的模糊隶属度函数和概率密度函数。对式(8)作变换得:

$$H_h(R, F) = -\int_{-\infty}^{+\infty}p(x)\lg p(x)\,dx - \int_{-\infty}^{+\infty}\{\mu(x)\cdot\lg\mu(x) + (1-\mu(x))\lg(1-\mu(x))\}dx + \int_{-\infty}^{+\infty}(1-p(x))\{\mu(x)\lg\mu(x) + (1-\mu(x))\cdot\lg(1-\mu(x))\}dx = H_s + H_f - H_{sf} \quad (9)$$

式中,

$$H_s = -\int_{-\infty}^{+\infty}p(x)\lg p(x)\,dx \quad (10)$$

$$H_f = -\int_{-\infty}^{+\infty}\{\mu(x)\lg\mu(x) + (1-\mu(x))\lg(1-\mu(x))\}dx \quad (11)$$

$$H_{sf} = -\int_{-\infty}^{+\infty}(1-p(x))\{\mu(x)\lg\mu(x) + (1-\mu(x))\lg(1-\mu(x))\}dx \quad (12)$$

由式(6)和式(9)可以看出, 在由随机空间和模糊空间组成的联合空间中, 系统总的不确定性(混合熵)等于统计熵与模糊熵的和再减去它们的交叉熵。

2　线元不确定性的混合熵模型

目标模型认为, 空间数据的分布可以用一组离散的点、线和面来表示, 它适合表征具有明确定义的空间实体。假设有一多边形数据分类模型如图 1 所示, A、B 为不同属性的地类, l 为地类的边界线。分界线的点位不确定性具有随机性, 且服从正态分布, 即其概率密度分布函数为 $p(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{x^2}{2\sigma^2}\right\}$, 则由式(10)可以求得其统计熵为:

$$H_s(x) = -\int_R p(x)\ln p(x)\,dx = \ln\sigma\sqrt{2e\pi} \quad (13)$$

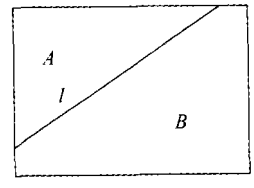


图 1　多边形数据分类模型

Fig. 1　Polygon Data Classing Model

由式(13)可以看出, 点位的统计熵的大小与点位分布的方差有关。根据文献[15], 随机点点

位的不确定性熵区间为:

$$R_{\text{entropy}} = \frac{1}{2}e^{H_s(x)} = \sqrt{\frac{e\pi}{2}}\sigma \approx 2.066\,4\sigma \quad (14)$$

这样, 区间$(-2.066\,4\sigma, 2.066\,4\sigma)$就是统计熵意义下的点位误差分布区间, 在该区间中, 集中了随机变量的主要不确定性信息, 它是随机不确定性的客观测度。

假设在垂直于边界线 l 的方向上, A、B 两类属性都为线性模糊分布, 其隶属函数分别为:

$$\mu_A(x) = \begin{cases} 1, & x < a_1 \\ (x + a_1)/2a_1, & a_1 \leqslant x \leqslant 0 \\ (b_1 - x)/2b_1, & 0 \leqslant x \leqslant b_1 \\ 0, & x > b_1 \end{cases}$$

$$\mu_B(x) = \begin{cases} 0, & x < a_2 \\ (a_2 - x)/2a_2, & a_2 \leqslant x \leqslant 0 \\ (b_2 + x)/2b_2, & 0 \leqslant x \leqslant b_2 \\ 1, & x > b_2 \end{cases} \quad (15)$$

隶属函数分布如图 2 所示, 即在分界线 l 上, 其模糊性最大, 隶属函数值为0.5; 随着离边界线 l 距离的增大, 其模糊性逐渐减小, 直至最大隶属度为 1 或最小隶属度为 0。则由式(11)可以计算其模糊熵为:

$$H_f(A) = (b_1 - a_1)/2 \quad (16)$$

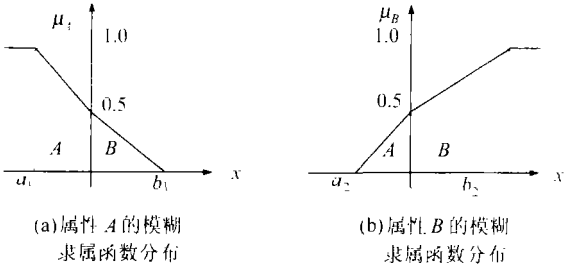
$$H_f(B) = (b_2 - a_2)/2 \quad (17)$$

由式(16)、式(17)可以看出, 线性模糊分布函数的模糊熵仅与分布区间有关, 其大小等于分布区间长度的一半。根据模糊熵的可加性, 由模糊不确定性所产生的总模糊熵为:

$$H_f = \{(b_1 - a_1) + (b_2 - a_2)\}/2$$

则其交叉熵为:

$$H_{sf} = -\int_{a_1}^{b_1}(1 - p(x))\{\mu_A(x)\lg\mu_A(x) + (1 - \mu_A(x))\lg(1 - \mu_A(x))\}dx - \int_{a_2}^{b_2}(1 - p(x))\cdot\{\mu_B(x)\lg\mu_B(x) + (1 - \mu_B(x))\lg(1 - \mu_B(x))\}dx \quad (18)$$



(a)属性 A 的模糊隶属函数分布　(b)属性 B 的模糊隶属函数分布

图 2　属性 A、B 的模糊隶属函数分布

Fig. 2　Distributions of Fuzzy Membership Function of Category A and B

式(18)无法用符号代数式表示, 只能在已知积分区间的情况下用数值积分求出其近似数值。

现假设位置不确定性的概率密度函数为标准正态分布, 即 $\sigma = \pm 1$; 属性 A、B 的模糊隶属函数相同, 且为对称分布, 令 $a_1 = a_2 = -1$, $b_1 = b_2 = 1$, 则统计熵 $H_s = \ln\sqrt{2e\pi} \approx 1.418\,9$, 模糊熵 $H_f = 2$。采用 Matlab 数值积分工具计算式(18), 得 $H_{sf} \approx 0.720\,5$, 则 $H_h(R, F) = 1.418\,9 + 2 - 0.720\,5 = 2.698\,4$。由式(14)可以求得线元的统计熵、模糊熵和混合熵的不确定性熵带半径分别为 2.006 3、3.694 5 和5.616 7。图 3 给出了以统计熵、模糊熵和混合熵为半径的线元不确定性的分布情况。

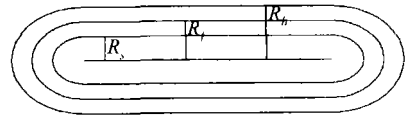


图 3　线元的随机熵带、模糊熵带和混合熵带

Fig. 3　Statistic Entropy Band, Fuzzy Entropy Band and Hybrid Entropy Band of Line Segment

3　结　语

本文仅考虑了 GIS 中空间数据同时含有随机性和模糊性情况下的不确定性混合熵模型的建立, 但空间数据可能还同时含有其他的不确定性, 如灰性和时域不确定性等, 如何估算这些情况下空间数据的不确定性, 还有待继续研究。

参　考　文　献

[1] Goodchild M F. Geographical Information Science [J]. International Journal of Geographical Information System, 1992(6): 31-46

[2] Cheung C K, Shi Wenzhong, Zhou X. A Probability-based Uncertainty Model for Point-in-Polygon Analysis in GIS[J]. Geoinformatica, 2004, 8(1): 71-98

[3] Foody G M. Approaches for the Production and Evaluation of Fuzzy Land Cover Classifications from Remotely-sensed Data[J]. Int. J. Remote Sensing, 1996, 17: 1 317-1 340

[4] Goodchild M F, Hunter G F. A Simple Positional Accuracy Measure for Linear Features[J]. International Journal of Geographical Information System, 1997, 11: 299-306

[5] Zhang J X, Goodchild M F. Uncertainty in Geo-

graphical Information[M]. London: Taylor&Francis, 2002

[6] 史文中, 刘文宝. GIS 线元位置不确定性的随机过程模型[J]. 测绘学报, 1998, 27(1): 37-44

[7] 李大军, 龚健雅, 谢刚生, 等. GIS 线元误差熵带研究[J]. 武汉大学学报• 信息科学版, 2002, 27(5): 462-465

[8] 张景雄, 杜道生. 位置不确定性与属性不确定性的场模型[J]. 测绘学报, 1998, 28(3): 244-249

[9] 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

[10] 戴洪磊, 夏宗国, 黄杏元. GIS 中衡量位置数据不确定性的可视化度量指标族探讨[J]. 中国图像图形学报, 2002, 4(3): 239-249

[11] Shannon C E. A Mathematical Theory of Communication[J]. Bell System Technical Journal, 1948, 21: 379-423, 623-656

[12] Deluca A, Termini S. A Definition of Nonprobabilistic Entropy in Setting of Fuzzy Sets Theory[J]. Information Control, 1972, 20: 301-312

[13] Yager R R. Measure of Entropy and Fuzziness Related to Aggregation[J]. Information Science, 1995, 82: 147-166

[14] Zadeh L A. Probability Measures of Fuzzy Events[J]. Journal of Mathematics Analysis and Application, 1968, 23: 421-427

[15] 诺维茨基, 佐格拉夫. 测量结果误差估计[M]. 北京: 中国计量出版社, 1990

第一作者简介: 史玉峰, 副教授, 博士。主要从事信息模式识别理论与应用、空间数据不确定性研究。

E-mail: yufeng788@163.com

Hybrid Entropy Model of Spatial Data Uncertainty in GIS

SHI Yufeng[1,2,3]　*SHI Wenzhong*[3]　*JIN Fengxiang*[4]

(1 School of Architecture Engineering, Shandong University of Technology, 12 Zhangzhou Road, Zhangdian District, Zibo 255049, China)

(2 Key Laboratory of Geospace Environment and Geodesy, Ministry of Education, Wuhan University, 129 Luoyu Road, Wuhan 430079, China)

(3 Advanced Research Center for Spatial Information Technology, Department of Land Surveying and Geo-informatics, The Hong Kong Polytechnic University, Hunghom, Kowloon, Hong Kong)

(4 Presidential Secretariate, Shandong University of Science & Technology, 579 Qianwangang Road, Qingdao 266510, China)

Abstract: Based on the information theory and fuzzy set theory, in the light of the characteristics of randomicity and fuzziness of part spatial data in GIS, this paper proposes a hybrid entropy model of spatial data uncertainty, and hybrid entropy can be used to measure the total uncertainty of spatial data caused by stochastic uncertainty and fuzziness uncertainty. Taking the uncertainty of line segment as an example, this paper discusses the evaluating methods of statistic entropy, fuzziness entropy, and hybrid entropy of line segment uncertainty. And the entropy-band distribution of line segment uncertainty is shown.

Key words: uncertainty; spatial data; hybrid entropy; statistic entropy; fuzzy entropy; line segment

About the first author: SHI Yufeng, associate professor, Ph. D, majors in the theory and application of information pattern recognition and spatial data uncertainty.

E-mail: yufeng788@163.com