DOI:10.13203/j.whugis2000.01.009
第25卷第1期 武汉测绘科技大学学报 Vol.25 No.1
2000年2月 Journal of Wuhan Technical University of Surveying and Mapping Feb. 2000

用随机模拟方法建立矢量数据的误差模型*

张景雄[1] 杜道生[1] 孙家炳[2]

(1 武汉测绘科技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 武汉测绘科技大学信息工程学院,武汉市珞喻路129号,430079)

摘 要 探讨利用空间统计学进行位置数据的误差模型建立,以期求得对位置误差的空间相关性实行有效的量化和模拟。用随机模拟方法进行矢量数据误差的条件模拟,并用摄影测量数据对该方法进行了实验。

关键词 矢量数据;误差模型;实现;随机模拟;空间统计学
分类号 P207;P208

随着GIS在各行业的推广应用,空间数据库日渐庞大,误差问题及与之相联系的数据质量问题愈来愈显得紧迫[1]。概括地说,GIS误差的产生是因为:①计算机化的空间数据的管理和处理必须对极其复杂的现实世界进行必要的取舍和近似(即抽象化或离散化过程);②对抽象模型的量测(即获取数据的过程)也包含误差;③GIS空间数据的处理过程中也可能产生误差;④一个GIS工程所涉及的源数据的各种误差在分析处理过程中传播。

研究误差是为了探讨误差的产生、传播和控制问题。GIS中一类比较重要的数据是由地面测量或空中三角测量提供的点位数据,即二维或三维坐标。在经典测量中,点的精确定位一直是关注的重点。但GIS所涉及的数据类型异常复杂,对这些数据的误差进行有效的处理超出了传统测量平差的范畴。

1 现实世界的目标模型

经典的测量数据处理方法是假设现实世界可由目标模型来描述。目标模型是一种常用的、标准的空间数据模型,它认为空间分布(二维)可以用一组离散的点、线和面来表达。

空间数据的目标模型适合表达有明确定义的空间实体,如埋石点、公路和地块可以分别用点、线和面来精确地表示。目标的定位数据借助于点、线和面等基本几何实体来表达;属性数据用于进一步描述这些几何实体的其他定性或定量特性,如埋石点的等级、公路的路面材料和承压力、地块的地价或税收情况等。

目标模型的适用性必须从相对意义上来理解,因为纯几何意义上的点、线和面并不存在。又由于实用性考虑以及经济因素的制约,目标定位数据的采样只能是一种近似,如复杂的线(包括面的边界)用一系列离散的采样点来描述,这其中就有近似性和取舍问题。不过,明确定义的目标位置和属性原则上是可以精确测定的,所以在研究目标的不确定性问题时,位置和属性的真值(即参考数据)往往是充分和必要条件。

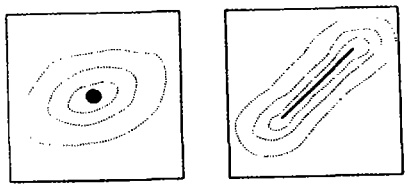
2 位置误差的场模型

本文讨论GIS中由点和弧段组成的矢量数据的位置误差的描述和模型建立。一般地,点的误差椭圆和线(即若干弧段)的ε-误差带使用的度量指标(如均方根差、标准离差、误差带宽等)是数值型的。因此,常规的基于目标的数据库均将上述误差指标看作是目标的扩展属性,使用地学关系数据结构来存储它们。当得知某一点的误差椭圆参数后,可以进一步根据误差椭圆概率密度函数,利用计算机模拟该误差分布的若干"实现"(realizations)。这些不同的"实现"作为样本输入到某一空间过程,用来探求点的位置误差在该空间过程中的传播以及该误差对某一派生数据产品的影响。上述过程均称为误差模型的建立[3]。

然而,线的位置误差情形就复杂得多。虽然ε-误差带不失为一个具有直观和简易等特点的线

收稿日期:1999-05-31.
* 国家自然科学基金及国家测绘局测绘科技发展基金、教育部优秀青年教师基金资助项目,编号49671063及97013。

的位置误差度量,但仅由带宽参数还不足以获得线目标的不同"实现"。现有的误差模型大都是解析式,即使用某种形式的概率密度函数,如包含一线段两端点的正态分布函数。为了建立某个空间运算(过程)的误差特性,实际中更有效的方法是模拟线段的不同"实现",以建立反映该过程的统计量,据此模拟数据。点的误差椭圆、线的 ε-误差带如图 1 所示,其中虚线表示位置不确定性的二维概率分布等值线。



(a)点的误差椭圆 (b)线的 ε-误差带

图 1 位置不确定性

Fig. 1 Positional Uncertainty

图 1 所示的概率分布等值线其实采用了 GIS 中的场模型[3]。场模型认为空间数据可以用定义在连续空间上的若干单值函数来表示。场模型描述的变量有类别型(称定性变量)和数值型(称定量变量),前者的实例为土地覆盖,后者的实例为地面高程。场模型的应用使得不确定性研究归结为所涉及变量的不确定性问题。

设所研究的数值型变量的场模型表示为函数 $z(x,y)$,(x,y)是二维空间坐标。人们所关心的是函数 $z(x,y)$能在多大程度上反映真值(设为 $Z(x,y)$)。函数值 $z(x,y)$与真值 $Z(x,y)$一般不吻合,设误差为 $e(x,y)=Z(x,y)-z(x,y)$,利用方差(σ^2)或均方根差(RMSE)度量该误差。进一步,当假设呈高斯分布特征时,概率可表示为:

$$pr\{z(x,y)-\sigma<Z(x,y)<z(x,y)+\sigma\}\approx 0.68$$

借助于场模型 $e(x,y)$来探讨建立位置误差模型,意味着位置误差不再是全局单一的,而是在空间上连续变化的;更重要的是可以借助连续场来考察位置误差的空间相关性,因为空间统计学可以为分析连续场空间相关性提供理论依据和量化工具[4,5]。这是离散目标模型所不能的。

3 位置误差建模的随机模拟方法

由上述可知,建立连续场误差模型的关键是生成连续的误差场。位置误差本来是针对离散目标的,即那些可以明确定义的现实客体(地物)。借助于连续场模型来研究位置误差,首先需要假想一个分辨力由仪器和人为因素所限制的密集的格网点,其上的点的位置误差呈现连续场的特性,这个连续场可以由若干已知点位上的误差推知。

位置误差模型的建立有赖于生成一系列等概率分布的位置误差的"实现",即产生几何上形似起始矢量数据但却是变形的若干组矢量数据。这可以由随机模拟方法来完成,所产生的一组模型数据即为一个"实现"[6,7]。

空间统计学的随机变量(random variables, RVs)是根据某一概率分布,能取不同值的一种变量,如污染物集中和人口密度。空间统计学中,随机变量的基本模型是[5]:

$$z(x)=u(x)+d(x)+\delta \qquad (1)$$

式中,$z(x)$是在点 x 的变量 Z 的值;$u(x)$是描述 Z 的结构部分的确定性函数;$d(x)$是描述 Z 的空间相关的、局部的随机变动;δ 是随机量或噪音。

空间统计学方法一个重要假设是平稳性,平稳性意味着任意两点所对应的两个值之间差值平方的期望只与这两点之间的距离和方向有关,即只要任意两点距离相等,方向相同,它们的差值平方的期望值相等:

$$\gamma(h)=E\{(z(x)-z(x+h))^2/2\} \qquad (2)$$

式中,$z(x)$、$z(x+h)$分别是变量 Z 在位置 x、$x+h$的值;h 是矢量滞后,表示这两个位置之间的距离和方向;$\gamma(h)$是量化空间相关特性并引导空间插值的所谓(半)变异函数。变异函数能从抽样数据估算,得到所谓的实验变异函数。

考察一个属性 $z(u)$在某一个场 A 的分布,可以表示为$\{z(u),u\in A\}$。随机模拟所建立的等概率、高分辨力有关 $z(u)$的空间分布的某一"实现"表示为:$\{z^{(l)}(u),u\in A\}$。建立条件模拟需满足条件 $z^{(l)}(u_\alpha)=z(u_\alpha)$,即所生成的"实现"遵从数据点位的数值。假设有 n 个数据点作为条件,则 $z(u)$在 A 的 N 个格网点的 N 个 RV's $\{Z_i,i=1,\cdots,N\}$的联合分布的条件累计分布函数(conditional cumulative distribution function, CCDF)可以表示为:

$$F_{(N)}(z_1,\cdots,z_N\mid(n))=$$
$$\text{Probability}\{Z_i\leqslant z_i,i=1,\cdots,N\mid(n)\} \qquad (3)$$

连续地使用条件概率关系可以看出,从式(3)取一个 N 变量的样本通过以下 N 个连续步骤获得(每一步只涉及单变量的条件累计分布函数,但条件的限制渐增):

1)给定原始数据(n),从 Z_1 的单变量的条件累计分布函数抽取一个数值 $z_1^{(l)}$,该数值添加到

(n),因此,条件更新为$(n+1)=(n)\cup\{Z_1=z_1^{(l)}\}$;

2)给定更新数据$(n+1)$,从 Z_2 的单变量的条件累计分布函数抽取一个数值 $z_2^{(l)}$,条件进一步更新为$(n+2)=(n+1)\cup\{Z_2=z_2^{(l)}\}$;

3)依此类推,从而得到所有 N 个 RV 的模拟数值 $z_i^{(l)}(i=1,2,\cdots,N)$。

假如 X 和 Y 方向的误差场是相互独立的,并按上述方法分别得到了模拟的 X 和 Y 方向的稠密的误差场,则可以内插求得离散点位的平面位置误差。空间内插处理的理论依据是,在地理现实世界里,纯集合点、线是不存在的,内插过程中涉及的计算误差可以控制在容许范围内。一般地,原始的点位误差数据需进行标准正态化,以便计算机的随机模拟,模拟后的结果需要逆变换到原始的误差尺度。

4　实　验

使用航空摄影测量方法来获取实验数据,所用的数据源还包括地面测量数据、大比例尺平面图。首先,采用常规点位误差估计方法,将从各数据层次上所选取的若干个明显地物点的坐标测量值与参考数值(即真值)作比较,其误差以均方根差的形式表达,结果列于表 1 中[8]。

上述点的位置误差估计值只是各数据层在整个研究地区的独立点的平均误差。若要对线状地物的位置误差进行估计,可以使用随机模拟方法。为此,利用 1:5 000 和 1:24 000 比例尺航空影像,在解析测图仪 AP190 上获取了两套矢量数据,以前者为参考数据对后者进行评估。所提取的线状地物包括铁路、围墙(篱笆)、建筑物轮廓、人行小径和小湖的水涯线等。图 2 中的实线和虚线分别表示从 1:5 000 和 1:24 000 比例尺航空影像上所获得的矢量数据。

表 1　点位误差估计

Tab.1　Estimation of Point Error

试验数据层次	位置误差/m
地面测量点	0.08~0.18
1:5 000 比例尺航片加密点	0.18
地图数字化	
1:1 250 比例尺平面图	0.9~0.13
1:2 500 比例尺平面图	0.19~0.23
解析法测图(1:24 000 比例尺航片)	0.57~0.60

所获取的矢量数据输出到 Arc/Info 系统,以方便后续处理。测试数据与参考数据的同名点的"匹配"依据是距离最近和属性一致,然后进行位置偏差的计算,结果存入匹配点的属性 PAT 文件。位置偏差数据转换到一个空间统计学软件系统 GSLIB[6],计算实验变异函数,并用球面函数拟合出理论变异函数,如图 3 所示,其中点和线分别代表实验和理论的变异函数。

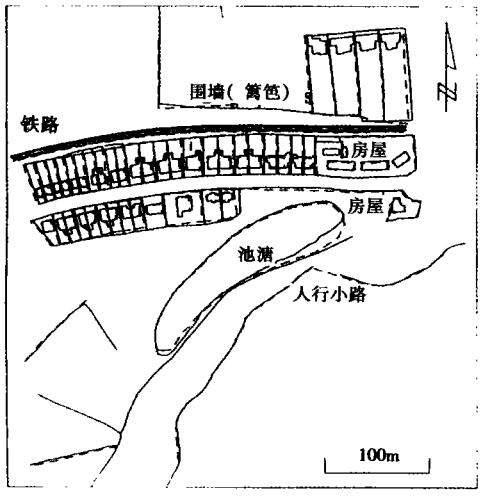


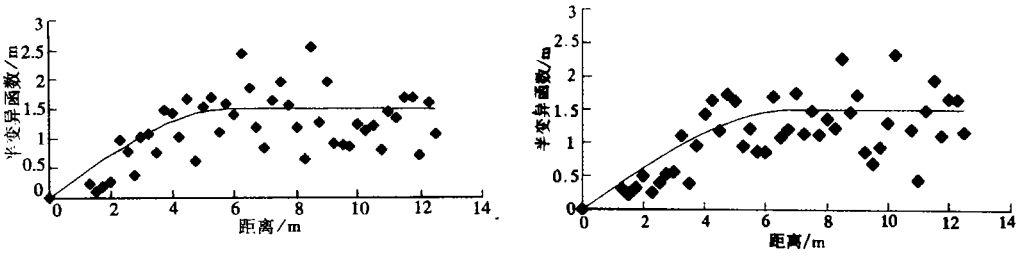
图 2　航空摄影测量方法获取的矢量数据

Fig.2　Vector Data Derived from Aerial Photogrammetry

在拟合出理论变异函数的基础上,利用 GSLIB 的高斯序贯模拟程序,经过适当的参数(如伪随机数"种子")设置和调整,选择规则格网尺寸为 8m,输出了 X 和 Y 方向各 10 个条件模拟图像。这些模拟数据经过格式变换,输入到 Arc/Info,内插得到匹配点的扩展属性 PAT 文件。这样,从 PAT 文件可以组合生成 10 个模拟矢量数据。图 4 的实线和虚线分别表示原始的和模拟的矢量数据。

利用随机模拟方法,得到了符合特定空间差异和相关特征的矢量数据的若干"实现"。实验表明,空间统计学方法可以量化实际数据所具有的空间相关性,并按照严密的条件,生成反映特定特性的矢量数据,以模拟相同概率分布的矢量数据。这些矢量数据可以认为是相同主客观条件下所能获取的,如同典型量测过程中呈现的随机误差一样。

通过计算机模拟含有误差的矢量数据,使得误差模拟过程可以灵活实现,而不局限于某个特定的数据和其分布类型。另外,随机模拟方法是假设所分析的数据不受粗差或系统误差的影响,仅限于对随机误差的计算机模拟。在此前提下,一旦按照本文所描述的随机模拟步骤生成了足够样本数的"实现"数据,并且作为数据输入到一特定的空间进行分析和处理,即可实施 MonteCarlo



(a) X 方向 (b) Y 方向

图3 位置误差的变异函数

Fig.3 Semivariograms for Positional Errors

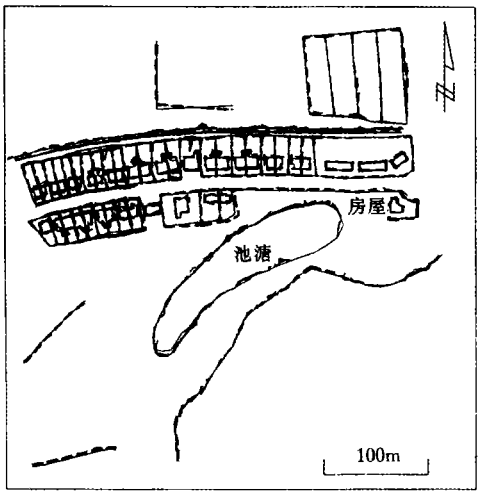


图4 原始(实线)和模拟生成(虚线)的矢量数据

Fig.4 The Original (Solid Lines) and Simulated (Dashed Lines) Vector Data

模拟。所以,本文所描述的随机模拟方法为探索不确定的地理现实所用,而非验证的工具。

5 结 论

矢量数据是GIS的重要数据类型。空间数据的误差问题讨论也大多侧重于对地图数字化的误差处理、模型建立等方面,并取得了可喜的进展[9~14]。随着对空间数据误差问题的关注和研究,场模型的方法也越来越受到重视[15~17]。

实验表明,本文所提出的空间统计学方法能生成反映特定空间差异和相关特征的矢量模拟数据。实验所用的技术和平台为地理信息产业的通用系统。因为本文假设 X 和 Y 方向的位置误差相互独立,但实际情况不一定如此,所以未来主要研究目标与场模型的相互补充和利用。另外,需要将GIS和有关的空间分析软件、空间统计学软件系统更好地集成起来,以便使机助误差处理更好地为科技人员和生产部门使用。

致谢:本文所涉及的数据由爱丁堡大学地理学系提供。李德仁院士为笔者的研究提供了最广义的基础设施。

参 考 文 献

1 Guptill S, Morrison J. Elements of Spatial Data Quality. Oxford: Elsevier Scientific, 1996

2 NCDCDS (National Committee for Digital Cartographic Data Standards). The Proposed Standards for Digital Cartographic Data. The American Cartographer, 1988, 15(1):9~140

3 Goodchild M F. The State of GIS for Environmental Problem Solving. In: Goodchild M F, Parks B O, Steyaert L T eds. Environmental Modelling with GIS. New York: Oxford University Press, 1993. 8~15

4 Burrough P A, Frank A U. Geographic Objects with Indeterminate Boundaries. Basingstoke: Taylor and Francis, 1996

5 Cressie N. Geostatistics: a Tool for Environmental modellers. In: Goodchild M F, Parks B O, Steyaert L T eds. Environmental Modelling with GIS. New York: Oxford University Press, 1993. 414~421

6 Deutsch C, Journel A G. GSLIB: Geostatistical Software Library and User's Guide. New York: Oxford University Press, 1992

7 Journel A G. Modelling Uncertainty and Spatial Dependence: Stochastic Imaging. International Journal of Geographical Information Systems, 1996, 10(5):517~522

8 Zhang J. A Surface-based Approach to Handling Uncertainties in an Urban-orientated Spatial Database: [Ph. D Thesis]. Edinburgh: The University of Edinburgh, 1996

9 Chrisman N R, Yandell B S. Effects of Point Error on Area Calculations: a Statistical Model. Surveying and Mapping, 1982, 48(4):241~246

10 Prisley S P, Gregoire T G, Smith J L. The Mean and

Variance of Area Estimates Computed in an Arc-node Geograhical Information System. Photogrammetric Engineering and Remote Sensing, 1989, 55(11): 1 601～1 612

11 Dunn R, Harrison A R, White J C. Positional Accuracy and Measurement Error in Digital Databases of Land Use: An Empirical Study. International Journal of Goegraphical Information Systems, 1990, 4(4): 385～398

12 Goodchild M F, Hunter G J. A Simple Positional Accuracy Measure for Linear Features. International Journal of Geographical Information Systems, 1997, 11(3): 299～306

13 Garcia J A, Fdez-Valdivia J, Perez de la Blanca N. An Autoregressive Curvature Model for Describing Cartographic Boundaries. Computers & Geosciences, 1995, 21(3): 397～408

14 Huang Y C, Liu W B. Building the Estimation Model of Digitizing Error. Photogrammetric Engineering and Remote Sensing, 1997, 63(10): 1 203～1 209,

15 Kiiveri H T. Assessing, Representing and Transmitting Positional Uncertainty in Maps. International Journal of Geographical Information, 1997, 11(1): 33～52

16 Jain A K, Yu Z. Object Matching Using Deformable Tempaltes. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1996, 18(3): 267～277

17 张景雄, 杜道生. 一个基于场的 ε-误差带模型. 武汉测绘科技大学学报, 1997, 22(3): 212～215

张景雄,男,35岁,副教授,博士后。现主要从事地理信息系统及摄影测量与遥感研究。代表成果:异值地理空间数据不确定性模型的建立;GIS 与 LIS 空间数据处理的集成。

E-mail: jxz@hp01. wtusm. edu. cn

Modeling Errors in Vector Data Using Stochastic Simulation

ZHANG Jingxiong[1]　*DU Daosheng*[1]　*SUN Jiabing*[2]

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

(2 School of Information Engineering, WTUSM, 129 Luoyu Road, Wuhan, China, 430079)

Abstract Vector data are important components in geographical information systems (GISs), which are represented via discrete points and lines that are often topologically structured. It is a known fact that various errors exist in vector data and their geo-processing, where positional errors are of major concern for well-defined objects, though focus should be shifted to attribute errors otherwise. Research on GIS errors is oriented to describing, modeling and visualizing them for spatial decision support. Modeling errors is a key issue. One method is to use analytic tools such as variance propagation to ascribe mathematical formulae for specific data sets or geo-processing. Error modeling is more effectively and flexibly carried out by simulating alternative, equal-probable realizations of a data set so that it is possible to analyze the propagation of errors from source data sets such as polygon coverages to an overlaid coverage. Based on error modeling, both data producers and data users are able to assess the fitness of a particular data product for a certain purpose. With geostatistics, spatially distributed phenomena are conceived as random variables, which are assumed to take values drawn from population conforming to a specific distribution. The notion of random variables conveys spatial variabilities, which are central to capturing spatially varying errors in spatial data. Geostatistics is of particular usability for research on GIS error issues on at least two aspects: one is spatial interpolation known as Kriging, which produces variance surfaces as by-products along with interpolated surfaces, the other is stochastic simulation, which generates alternative, equal-probable surfaces. The latter approach has been widely used for error modeling in environmental and geographical problem-solving. However, its application for vector data has been rare. This paper presents a novel use of geostatistical simulation for modeling positional errors in vector data.

Successive application of the conditional probability relation shows that drawing an N-variate sample from the equation above can be done in N successive steps, each using a univariate CCDF.

An Edinburgh suburb was chosen at the test site, with the 1:24 000 scale aerial photographs being used to generate tests data and 1:5 000 scale aerial photographs to provide reference data, for which coordinates at a certain point or verticcel located at x are denoted $x(x)$ and $X(x)$ respectively. The positional error at this location $\varepsilon(x)$ can then be expressed as: $\varepsilon(x) = X(x) - x(x)$. The underlying rationale is that, using photogrammetric techniques in urban areas for increased efficiency, aerial photographs at large and medium scales are normally used for topographic and thematic mapping.

While it is common practice to use $\varepsilon(x)'s$ at checking points to derive error measures such as RMSE in position and elevation, it is not adequate for those error descriptors to be used as vector error models to predict the accuracy in derivative data products such as line lengths and polygon areas by means of variance propagation, unless homogeneity and spatial independence of positional errors among points are assumed. The method used in this experiment was to apply conditional simulation to simulate equal-probable $\varepsilon(x)'s$ and, in turn, alternative versions of the test data ($x(x)'s$) in order to model errors in the source data and assess the consequences of using them in a certain map operation. Stochastic simulation was performed using a Gaussian sequential simulation program SGSIM provided in GSLIB. The parameter file was supplied with suitable data including grid cell size (8 by 8 metres), ranges, sills and nugget effects describing semivariogram models. Ten realizations were created from SGSIM for X, Y independently, and were put as new data items in the Arc/Info PAT file mentioned above. Ten versions of vector data were generated from the expanded PAT file.

The results confirm that spatial variability in positional errors can be usefully explored via geostatistics, and desired number of simulated vector data sets can be generated from conditional simulation approach supported in public domain geostatistical software systems such as GSLIB, which are, unfortunately, often packaged independent of GIS platforms. It would thus be desirable to integrate error modeling functions such as that described above into mainstream GIS software systems so that information on spatial data errors is accessible for general GIS end users. Further research should be directed towards foundmental issues related to uncertain vector data modeling, such as simulation of vector data with geometric and topological constraints, which could be reached unduly by simulated artifacts.

Key words vector data; error models; realization; stochastic simulation; geostatistics

---

ZHANG Jingxiong, male, 35, associate professor and a postdoctoral research scientist. He is with research interests in GIS, remote sensing and photogrammetry, exploring novel methods for monitoring China's dryland salinity. His main work is on modeling uncertain geo-spatial data of heterogeneous nature, improved spatial data handling in integrated GIS/LIS. His paper "Fully-fuzzy supervised classification of sub-urban land bover from remotely sensed imagery: statistical and artificial neural network approaches" was in press with Int. J. Remote Sensing.

E-mail: jxz@hp01.wtusm.edu.cn