DOI:10.13203/j.whugis2002.05.002
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)05-0445-06
文献标识码:A

基于一般抽样原理的GIS属性数据质量评定方法

史文中[1]　刘　春[2]　刘大杰[2]

(1　香港理工大学土地测量与地理资讯学系,香港九龙红磡)
(2　同济大学测量与国土信息工程系,上海市四平路1239号,200092)

摘　要:在一般抽样原理的基础上,给出了缺陷率统计模型,并推导出缺陷率的两个统计特征值均值和方差。理论结果和应用实例说明了缺陷率对GIS属性数据精度度量的可行性。在此基础上,利用缺陷率的均值和方差,详细给出了GIS数据生产时单幅数字地图以及多幅数字地图对应属性数据质量的评定应用。

关键词:质量评定;一般抽样;缺陷率;属性数据
中图法分类号:P208;P207

属性不确定性是在采集、描述和分析真实世界中客观实体的过程中,实体属性的量测、分析值围绕其属性真值,在时间和空间内的随机不确定性变化域。也可以认为,属性不确定性是更广义上的属性误差问题,或者属性误差问题的纵横延伸。连续属性数据的不确定性可以用量测误差度量,使用和位置不确定性相同的误差传播定律等方法;非连续属性数据的不确定性可以通过评价一组分类结果来实现。研究属性不确定性可概括为探讨不确定性的产生、传播和控制等几个方面的问题。

长期以来,许多学者在数据质量与不确定性方面作了深入的研究并取得了许多成果,但多数都是对几何图形数据精度模型和应用进行研究和分析,而对属性精度的研究仍然比较欠缺。在对属性数据的研究中,很多成果是在遥感影像分类中探讨分类属性精度的度量与描述的。如Congalton和Green(1999)对遥感数据精度估计的方法和应用作了较为综合的论述;Card(1982)、Chrisman(1982)和Hay(1988)探讨了遥感影像分类中误分类对面积估算的影响;Stephen(1996)讨论了分层随机抽样中,分类属性精度的Kappa系数估计及其方差;Ma和Redmond(1995)则讨论了遥感影像分类精度的误差矩阵,并采用Kappa系数、Tau系数和概率估计方法对分类属性精度进行度量。近年来,在属性数据精度的研究方面又有了一些新的方法。例如,模糊集合理论(Zadeh,1965)被应用于对属性分类不确定性的度量(Wang & Hall,1996;Burrough & Faran,1996;Brown,1998);粗集理论(Pawlak,1982)被探讨应用于空间数据分类精度的描述和度量(Schneider,1997;Worboys,1998a,1998b;Ahlqvis, keukelaar & Oukbir,2000);朱光(1997)采用数学关系来处理数字形式的属性数据误差,用逻辑关系处理字符形式的属性类别数据;张景雄、杜道生(1999)采用场模型对属性精度的性质及度量从模型的角度进行研究。

用于属性不确定性与精度分析的方法中,目标模型和域模型是较经典的数据处理模型,概率论、概率矢量、证据理论和空间统计学具有统计意义,而粗集、模糊集合、云理论、遗传算法、混沌理论、灰色理论和不确定数学则用于实现计算和模拟。

属性数据在GIS系统中多是没有数值关系的属性描述值,而研究大量这些类型的数据时,抽样方法是一个简单而实用的方法(Schilling,1982;Cspary,1999;Joos,2000)。在一般抽样原理的基础上,本文给出了缺陷率统计模型,同时在数学上严格给出了一般抽样条件下缺陷率的两个统计特征值均值和方差,以此说明采用缺陷率对GIS属性数据精度度量的可行性,在此基础上利用缺陷率的均值和方差详细给出了GIS数据生产时,单幅数字地图和多幅数字地图对应属性数据质量的评定方式以及应用。

收稿日期:2002-06-10。
项目来源:国家自然科学基金资助项目(40171078);香港特别行政区资助局资助项目(3-ZB40);香港理工大学资助项目(1.34.9709)。

1 属性数据精度的缺陷率模型

在GIS空间数据中，几何图形数据与属性数据库中的记录是通过关键字相互对应的。属性数据目前多以数据库存储，采用抽样方法抽取图形实体，可以依据关键字对应相应的属性数据记录。大多数的属性数据值不是量化值，因此其质量的好坏很难直接量化确定。比如，区分和标识空间数据的标识码、空间实体的描述值以及逻辑关系值等。所以需要采用一定的量化方法把属性数据的质量量化，然后寻找评价质量的指标以确定属性数据质量的好坏，这正是本文所要探讨的内容。

这里采用的量化方法就是用抽样方法抽取一定数量的属性数据单元，并对这些数据单元进行质量检验，凡是不符合数据生产质量标准的属性数据单元就可以认为是一个缺陷，并通过计数缺陷的方法来获得属性数据质量的检查值。获得被检验属性数据的缺陷计数值后，就可以寻找一个统计指标用以对整个属性数据的质量作出评价。

所以对于一个待评定的属性数据总体，有一个数据总体容量 N，它是指该检验数据的总体数据量，即如果检验过程中，以一个数据单元为检验对象，则总体容量就是整个检验总体中数据单元的总和。同样，检验过程中有一个抽样容量 n，它是指提供被检验的总抽取数据量，即如果检验过程中，以一个数据单元为检验对象，则抽样容量就是整个抽取数据单元的总和。

数学描述上，设有某一属性数据集合为 $\boldsymbol{X}$，它包含数据单元 N 个(即总体容量为 N)，采用不放回的简单抽样方法对数据单元进行抽样检查，若抽样的样本容量为 n，则抽样检验所得到的缺陷数为 y。

对于每个被检验的属性数据单元，总有是或不是缺陷两种情况，所以

$$y_j=\begin{cases}1, & \text{若第 } j \text{ 个数据单元是有缺陷的}\\ 0, & \text{若第 } j \text{ 个数据单元是无缺陷的}\end{cases} \tag{1}$$

其中，j 为抽样中的第 j 个抽样数据单元。则缺陷数 y 为：

$$y=\sum_{j=1}^{n} y_j \tag{2}$$

y 与 n 的比值通常称为抽样的数据缺陷率估值，记为：

$$\hat{u}=\frac{y}{n} \tag{3}$$

若总缺陷数为 Y，则 $\hat{u}$ 也是总体数据缺陷率 $u=Y/N$ 的估值。

这里提出的缺陷率模型简单地说就是单位数量数据单元中包含的缺陷数，由于这一指标计算中的缺陷数是属性数据质量的检验值，所以缺陷率可被用来对属性数据的质量进行度量。

2 缺陷率统计模型的均值与方差表达

显然，缺陷率模型是一个统计模型，同时缺陷率的取值与抽样大小、抽样方案以及抽样检验的结果有密切关系，所以需要得到在抽样方法下的缺陷率值的参数估计值，从而可以根据缺陷率的参数估计值对属性数据质量进行评定。对于一个统计量，常用其均值和方差来描述其性质，所以在利用缺陷率对属性数据精度进行度量前，获得该统计量严格的均值和方差表达有利于掌握缺陷率的统计性质，并能进一步证明缺陷率对属性数据质量度量的可行性。本文从抽样的一般性定义入手，推导缺陷率模型的均值与方差表达。

对于一个数据总体容量为 N 的属性数据集合 $\boldsymbol{X}$，考虑一般情况下，抽样对于总体属性数据中的任意一个数据单元，可令

$$a_j=\begin{cases}1, & \text{若第 } j \text{ 个数据单元抽入样本}\\ 0, & \text{其他}\end{cases} \tag{4}$$

$$j=1, 2, \cdots, N$$

兼顾式(1)，所以抽中的属性数据中的缺陷数和为：

$$y=\sum_{j=1}^{n} y_j=\sum_{j=1}^{N} a_j y_j \tag{5}$$

采用不放回的简单随机抽样，对于抽样中的任意一个数据单元，其入样的概率是一样的，可以得到 a_j 的概率分布如下：

a_j	1	0
p	$\frac{n}{N}$	$1-\frac{n}{N}$

所以 a_j 的均值 $E(a_j)$ 为：

$$E(a_j)=P(a_j=1)=\frac{n}{N} \tag{6}$$

根据式(3)，$\hat{u}$ 的均值 $E(\hat{u})$ 为：

$$E(\hat{u})=E\left(\frac{y}{n}\right)=\frac{1}{n}\sum_{j=1}^{N} E(a_j)y_j=$$

$$\frac{1}{n}\cdot\frac{n}{N}\sum_{j=1}^{N} y_j=\frac{Y}{N}=u \tag{7}$$

根据统计学原理，由于 a_j 是0、1取值，所以可认为其服从二项分布，则它的均值、方差和协方

差分别为:

$$E(a_j)=\frac{n}{N} \tag{8}$$

$$V(a_j)=\frac{n}{N}(1-\frac{n}{N})$$

$$\mathrm{cov}(a_j,\ a_k)=-\frac{n}{N(N-1)}(1-\frac{n}{N}) \tag{9}$$

令 $f=\frac{n}{N}$, 则 $\hat{u}$ 的方差 $V(\hat{u})$为:

$$V(\hat{u})=V\left(\frac{1}{n}\sum_{j=1}^{N}y_j\right)=\frac{1-f}{nN}\cdot\left(\sum_{j=1}^{N}y_j^2-\frac{2}{N-1}\sum_{j=1}^{N}\sum_{k>j}^{N}y_jy_k\right) \tag{10}$$

由于有$(\sum_{j=1}^{N}y_j)^2=\sum_{j=1}^{N}{y_j}^2+2\sum_{j=1}^{N}\sum_{k>j}^{N}y_jy_k$, 则式(10)为:

$$V(\hat{u})=\frac{1-f}{n}\cdot\frac{1}{N-1}\left[\sum_{j=1}^{N}{y_j}^2-\frac{1}{N}(\sum_{j=1}^{N}y_j)^2\right]=\frac{1-f}{n}S^2 \tag{11}$$

其中,

$$S^2=\frac{1}{N-1}\sum_{j=1}^{N}(y_j-Y)^2=\frac{1}{N-1}\sum_{j=1}^{N}(y_j^2-NY)^2 \tag{12}$$

由于 y_j 的取值为 0 或 1, 所以有$\sum_{j=1}^{N}y_j^2=\sum_{j=1}^{N}y_j$, 则式(12)为:

$$S^2=\frac{1}{N-1}(Nu-Nu^2)=\frac{N}{N-1}u(1-u) \tag{13}$$

把式(13)代入式(11)则有:

$$V(\hat{u})=\frac{N-n}{n(N-1)}u(1-u) \tag{14}$$

由于抽样方差 S^2 需要通过样本方差 s^2 来估计, 而且缺陷率 u 是未知的, 它也需要通过样本估计, 所以有:

$$s^2=\frac{1}{n-1}\sum_{j=1}^{N}(y_j-y)=\frac{n\hat{u}(1-\hat{u})}{n-1} \tag{15}$$

所以根据式(11), 缺陷率估值 $\hat{u}$ 的方差的估值 $V(\hat{u})$为:

$$V(\hat{u})=\frac{1-f}{n}\cdot\frac{n}{n-1}\hat{u}(1-\hat{u}) \tag{16}$$

由于缺陷率 u 的值较小, 同时影响其估值 $\hat{u}$ 精度的主要是n, 而不是 f, 因此当 f 很小(例如 $f<0.05$ 甚至 $f<0.1$)时可以忽略不计, 这时可以取 $1-f\approx1$, 则式(16)可为:

$$V(\hat{u})\approx\frac{\hat{u}(1-\hat{u})}{n-1} \tag{17}$$

对于抽样容量为 n 的一般随机抽样, 所得的属性数据缺陷率 $\hat{u}$ 的均值和方差表示为:

$$\left.\begin{aligned}E(\hat{u})&=u\\V(\hat{u})&\approx\frac{\hat{u}(1-\hat{u})}{n-1}\end{aligned}\right\} \tag{18}$$

n/N 较小时, 式(18)得到的是缺陷率 $\hat{u}$ 的均值和方差的估计值, 这两个值是缺陷率模型的主要统计特征值。从结果可以看出, 缺陷率对属性数据质量的度量是无偏的。所以缺陷率均值可以认为是属性数据抽样样本中缺陷多少的度量, 而缺陷率方差则是属性数据抽样样本数据中缺陷离散程度的度量。应用这两个统计特征值的总和就可以实现对属性数据精度的度量, 具体将体现在本文对数字地图属性数据质量的评定中。

3　单幅数字地图属性数据质量的评定

对于要进行质量检验的一个数据集合(如单幅数字地图的属性数据), 其中的所有属性数据可作为抽样检验的总体, 所以检验方案可以描述为: 从数据总体 N 中抽取容量为 n 的样本, 则获得缺陷数的检验值 y, 并计算出缺陷率 $\hat{u}$。以数字道路的属性数据集为例, 表 1 是路段部分属性数据列表, 共有 22 个属性数据项, 如果区域内有 1 000 条路段描述, 则数据总体容量 N 为 1 000×22=22 000。

表 1　路段部分属性数据列表

Tab. 1　Attribute Description for Road Segment

编号	属性内容	编号	属性内容
1	路段编号	12	最大允许车辆长度
2	路段平均速度	13	最大允许承载重量
3	交通指向	14	最大允许车辆宽度
4	路段长度	15	车道数
5	路段宽度	16	最大车道数
6	路段等级	17	最少车道数
7	路段形式	18	最大交通容量
8	路段功能等级	19	是否收费公路
9	路面建筑状况	20	允许车辆数量
10	路面建筑状况	21	道路所有权
11	通行限制	22	开放时间

根据统计学原理, 属性数据总体数据量 N、抽样量 n 以及 $N-n$ 较大时, 缺陷数 y 的分布可以近似为正态分布, 所以单幅数字地图属性数据质量的评定方式如下。

1) 以样本缺陷率 $\hat{u}$ 估值作为单幅数字地图属性数据质量的度量值, $\hat{u}$ 值越小则属性数据的质量越好, 反之则属性数据质量越差。

2) 样本缺陷率估计值 $\hat{u}$ 的标准差为 $\sqrt{V(\hat{u})}$

$\approx\sqrt{\frac{\hat{u}(1-\hat{u})}{n-1}}$，该值能反映出样本缺陷率估计值的抽样精度。

3）由于有 $P_r=(u-\hat{u}<d)=1-\alpha$，可以理解为缺陷率真值与其估计值之间的绝对偏差在一定可信度下小于一个常数。当样本容量足够大时，可以认为缺陷率估值 $\hat{u}$ 服从正态分布 $N(\hat{u}, V(\hat{u}))$，于是有 $\frac{u-\hat{u}}{\sqrt{V(\hat{u})}}$ 服从 $N(0,1)$ 分布，所以可以得到缺陷率估值 $\hat{u}$ 的上限值 u_c：

$$u_c=\hat{u}+\mu_{(1-\alpha)}\sqrt{V(\hat{u})} \tag{19}$$

其中，$\mu_{(1-\alpha)}$ 为在置信度为 $1-\alpha$ 时的正态分布分位点。

由于 $V(\hat{u})\approx\frac{\hat{u}(1-\hat{u})}{n-1}$，可得：

$$u_c=\hat{u}+\mu_{(1-\alpha)}\sqrt{\frac{\hat{u}(1-\hat{u})}{n-1}} \tag{20}$$

当属性数据缺陷率 $\hat{u}$ 小于该上限值时，可以认为用 $\hat{u}$ 评定属性数据质量具有概率为 $1-\alpha$ 的可信度。

4 多幅数字地图属性数据质量的评定

当对 K 个数据集合（如 K 幅数字地图属性数据）进行质量评定时，可以认为是对多个检验总体进行抽样检验。所以检验方案可以描述为：从 $k(k=1, 2, \cdots, K)$ 个数据总体（总体容量为 N_k）中分别抽取容量为 n_k 的样本，获得并计算出各数字地图属性数据的缺陷数 y_k 和缺陷率估值 $\hat{u}_k$。

这 K 个数据集合的缺陷率估值应为：

$$\hat{\bar{u}}=\frac{1}{k}\sum_{k=1}^{K}\hat{u}_k \tag{21}$$

所以多幅数字地图属性数据质量的评定方式如下。

1）多幅数字地图样本属性数据缺陷率 $\hat{\bar{u}}$ 是多幅数字地图属性数据质量的度量值，$\hat{\bar{u}}$ 值越小则这批数字地图的属性数据的质量越好，反之则属性数据质量越差。

2）多个数据集合样本缺陷率估计值 $\hat{\bar{u}}$ 的标准差为 $\sqrt{V(\hat{\bar{u}})}\approx\sqrt{\frac{\hat{\bar{u}}(1-\hat{\bar{u}})}{\bar{n}-1}}$，该值能反映出多个数据集合样本估计缺陷率的抽样精度。

3）同样，在置信度为 $1-\alpha$ 时，这批地图属性数据质量的上限值为：

$$u_c=\hat{\bar{u}}+\mu_{(1-\alpha)}\sqrt{\frac{\hat{\bar{u}}(1-\hat{\bar{u}})}{\bar{n}-1}} \tag{22}$$

当某一数字地图属性数据缺陷率 $\hat{u}$ 小于该限值时，可以认为该幅数字图的质量合格且在控制之中，并具有概率为 $1-\alpha$ 的可信度。

一批道路数字地图的属性数据需要对其进行检验，由于区域内道路的分布比较均匀，所以每一图幅的属性数据记录约为 4 000 条，属性数据检验的内容主要涉及表 1 中对路段的属性描述。抽样量的选择主要兼顾了抽样误差以及抽样费用。抽样量太大虽然可以控制抽样误差，但无疑带来较大的抽样费用；反之，虽然降低了抽样费用，但很难保证抽样的可靠性。一般情况下，主要将抽样方差控制在一定范围内，以此作为依据确定抽样量的大小。这里通过计算确定，从每一图幅中抽取约 350 条记录进行检验，置信度水平 $\alpha=0.05$。得到各图幅的缺陷率估值如表 2 所示。

表 2 多幅道路数字地图属性数据缺陷率

Tab.2 Rate of Disfigurement Estimations for Batch of Digital Road Maps

数字图幅号 k	各数字地图属性数据总体容量 N_k	样本容量 n_k	缺陷率 u_k
1	4 019	351	0.012
2	4 032	352	0.017
3	4 023	352	0.023
4	3 987	348	0.022
5	4 203	367	0.014
6	3 976	347	0.030
7	4 095	358	0.009
8	4 100	358	0.015
9	3 900	341	0.023
10	3 987	348	0.021
11	4 023	352	0.013
12	3 926	343	0.019
13	4 093	358	0.020
14	4 033	352	0.008
15	4 000	350	0.028
16	4 024	352	0.022
17	3 987	348	0.021
18	3 872	338	0.019
19	3 986	348	0.019
20	4 023	352	0.020
21	3 893	340	0.015
22	4 026	352	0.023
23	4 082	357	0.022
24	4 100	358	0.019
25	3 890	340	0.014

则多幅图缺陷率的期望值为：

$$\hat{\bar{u}}=\sum_{k=1}^{25}u_k=0.018\ 72$$

$\hat{\bar{u}}=0.018\ 72$ 是这批数字地图属性数据质量的度

量值, 不考虑抽样比 f 时, 这批数字地图属性数据质量缺陷率的上限值为:

$$u_c = 0.018\ 72 + \mu_{(1-0.05)} \cdot \sqrt{\frac{0.018\ 72(1-0.018\ 72)}{350-1}} = 0.030\ 645$$

当考虑抽样比 f 时, 这批数字地图属性数据质量缺陷率的上限值为:

$$u_{c(f)} = \hat{\bar{u}} + \mu_{(1-\alpha)} \sqrt{\frac{\hat{\bar{u}}(1-\hat{\bar{u}})(1-f)}{\bar{n}-1}} = 0.018\ 72 + \mu_{(1-0.05)} \cdot \sqrt{\frac{0.018\ 72(1-0.018\ 72)(1-0.087\ 375)}{350-1}} = 0.030\ 121$$

可见在不考虑抽样比 f 时, 对属性数据质量缺陷率上限值的估计略偏大, 但与缺陷率估值差两个数量级, 因而在缺陷率方差的估计中, 忽略抽样比的影响对公式的简化在实际应用中是合理的。

图 1 是各数字图缺陷率估计值的趋势分布, 图中的上界即为可信度为 95% 的属性数据缺陷率的上限值(不考虑抽样比 f), 中线即代表了这批数字图属性数据的质量值。把各幅数字图的缺陷率在图中展布, 一方面可以直观地看出属性数据的质量分布, 同时从该图中能够容易得到多幅数字图整体的属性数据质量。

类似图 1 的缺陷率展布图还可以用来直观表达多幅数字地图属性数据质量状况, 展布图中的上界是该批图属性数据质量在一定可信度下的上界, 缺陷率超过该上界的图幅其属性数据质量可认为发生变化。图 1 中的各幅数字地图属性数据质量都不大于上限值, 所以可以认为这批数字图的属性数据质量是稳定而且是在控制之中的。同时从该图上也能反映这批图属性数据质量的分布没有明显的系统性, 即其属性数据质量在一定范围内随机跳动, 没有质量的趋势变化, 所以该趋势图可用于对数据质量的监测和控制。

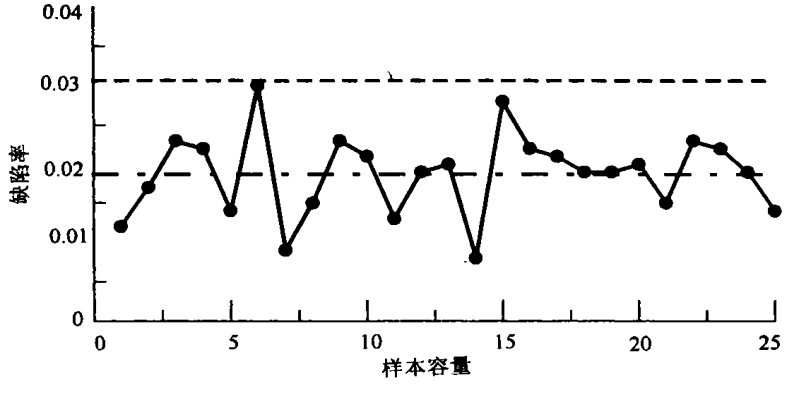

图 1　多幅道路数字地图属性数据缺陷率趋势图

Fig. 1　Trend Map of the Rate of Disfigurement for Attribute Data of Digital Road Maps

5　结　论

传统属性数据质量的研究多集中于分类属性数据的精度分析中, 而对于矢量地图属性数据的质量研究相对比较薄弱。本文着重探讨矢量地图属性数据质量的评定方法。首先在属性数据质量抽样检验的基础上, 提出了一般抽样方法下的缺陷率统计模型, 基于统计模型的特点, 推导了缺陷率模型中缺陷率均值和方差这两个统计特征值。结果发现, 由于缺陷率估计的无偏性以及缺陷检验值本身反映了属性数据的质量, 所以采用缺陷率度量属性数据的精度是合适的。在此基础上, 利用缺陷率均值和方差这两个统计特征值, 进一步给出了实际情况下单幅数字地图和多幅数字地图的属性数据质量评定方法。由于缺陷率统计模型基于抽样的基本方法, 所以在实际应用中具有较强的易操作性, 同时又有严密的理论依据, 对属性数据质量估计和评定具有实用价值, 这也是 GIS 属性数据质量研究中的一个新方法, 在生产实际中可考虑推广使用。当然缺陷率模型和质量评定方法与抽样方法和抽样数量的选取有很密切的关系, 因此, 探讨适合不同类型 GIS 属性数据的抽样方法以及优化后的抽样容量是需要进一步研究的内容。

参　考　文　献

1　Bonin O. New Advances in Error Simulation in Vector Geographical Databases Accuracy 200. The 4th Interna-

tional Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, Amsterdam, 2000

2 Burrough P A. GIS and Geostatistics: Essential Partners for Spatial Analysis. The International Symposium on Spatial Data Quality' 1999, Hong Kong, 1999

3 Caspary W, Joos G. Statistical Quality Control of Geodata. The International Symposium on Spatial Data Quality' 1999, Hong Kong, 1999

4 Caspary W, Joos G. Quality Criteria and Control for GIS Databases. The IAG SC4 Symposium, Eisentadt, 1998

5 Goodchild M F. Measurement-based GIS. The International Symposium on Spatial Data Quality' 1999, Hong Kong, 1999

6 Liu C, Liu D J. Study on Sampling Inspection Schemes to Digital Products in GIS. Geo-Spatial Information Science, 2001, 4(1): 62~67

7 Russell G C, Green K. Assessing the Accuracy of Remotely Sensed Data Principles and Practices. New York: CRC Press, Lewis Publishers, 1999

8 Fotheringham A S, Brunsdon C, Martin Charlton. Quantitative Geography. London: SAGE Publications Ltd, 2000

9 Kish L. 抽样调查. 北京: 中国统计出版社, 1997

10 梁小筠, 祝大平. 抽样调查的方法和原理. 上海: 华东师范大学出版社, 1998

11 史文中. 空间误差处理的理论和方法. 北京: 科学出版社, 1998

12 刘大杰, 史文中. 空间误差处理的理论和方法. 上海: 上海科学技术文献出版社, 1999

13 刘 春. GIS 属性数据的精度度量及质量控制的抽样原理与方法:[博士论文]. 上海: 同济大学, 2000

作者简介: 史文中, 博士, 副教授。主要从事地理信息系统和遥感研究, 出版学术专著 3 部, 编著 2 部, 发表学术论文 100 余篇。
E-mail: lswzshi@polyu.edu.hk

Quality Assessment for Attribute Data in GIS Based on Simple Random Sampling

SHI Wenzhong[1] *LIU Chun*[2] *LIU Dajie*[2]

(1 Dept. of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Kowloon, Hong Kong)

(2 Dept. of Surveying and Geo-Informatics, Tongji University, 1239 Siping Road, Shanghai, China, 200092)

Abstract: The research of spatial data accuracy and uncertainty began in the 1960s, and the NCDCDS (National Committee for Digital Cartographic Data Standards, 1988) identified several components of data quality as positional accuracy, attribute accuracy, logical consistency, completeness and lineage. At present, the issue of the representation of uncertainty in spatial data has become more and more of a concern. The description of error is take as a "function of information" and a "fundamental dimension of data" because of endemic nature of error in GIS widely recognized. Correspondingly, the development of what have been termed error-sensitive GIS has been deeply researched.

The using of sampling method in the analysis of spatial data accuracy and its measurement has two reasons: ① it conforms to the international standard that spatial data is taken as a product to be quality inspected; ② the data quality can be deduced and described with low expenses and high efficiency.

So, on the basis of the principles of the simple random sampling, the statistical model of rate of disfigurement (RD) is put forward and described in detail. Based on the definition of the simple random sampling for the attribute data in GIS, the mean and variance of the RD are deduced as the characteristic valve of the statistical model in order to explain the feasibility of the accuracy measurement of the attribute data in GIS using the RD. From the result of the deduced equation about the mean and variance of RD, the mean of RD is the measurement of the amount of the defects while the variance of RD is the measurement of the scatteration of the defects. So the mean

(下转第 461 页)

The discussion shows that the x and y coordinates of the same points are not correlative after coordinate translation, but correlation exists in the x and y coordinates of the different points. The number of the known points and the translating manner by which the conversion parameters were drawn determines the correlation of every point. The square error by using nine known points is smaller than that by using four known points after simulating conversion, when the knows points are the same. The precision of ground coordinate after conversion in affine conversion is lower than that in simulating conversion. With the calculating methods adopted in this paper, the value of the correlation changed from −0.235 3 to 0.542 3, which shows that the correlation of the coordinate conversion can not be ignored, and the error analyse should be made on the basis of correlation.

Key words: correlation; coordinates conversion; digitalized data; adjustment

About the author: YU Xiaohong, Ph. D candidate. Her majors include quality control of spatial data and attribute data in GIS.
E-mail: yxh196869@sina.com

（上接第 450 页）

of RD can be used to measure the attribute data accuracy while the variance of RD can measure the accuracy of the sampling and assure the sampling confidence. After that, the quality assessment method for attribute data of the singe or batch of vector maps during the procedure of the collecting is discussed based on the mean and variance of the RD. The RD spread graph is also drawn to see whether the quality of the attribute data is under control. The RD model can synthetic judge the quality of attribute data, which is different from other measurement coefficients that only discuss the accuracy of classification, so it can find its significance by realizing the measurement of the accuracy during the research of accuracy and uncertainty for attribute data in GIS.

Key words: quality assessment; simple random sampling; rate of disfigurement; attribute data

About the author: SHI Wenzhong, Ph. D., associate professor. His research interests include error modeling for GIS and remote sensing data, spatiotemporal and dynamic relationships of geographic objects, three-dimensional GIS models, integration of virtual reality, internet GIS and so on. He has published three research monographs, two edited books and near 100 papers.
E-mail: lswzshi@polyu.edu.hk