DOI:10.13203/j.whugis2004.03.014
第29卷 第3期 武汉大学学报·信息科学版 Vol.29 No.3
2004年3月 Geomatics and Information Science of Wuhan University March 2004
文章编号:1671-8860(2004)03-0244-05 文献标识码:A

数字土地信息中属性数据的质量控制

刘 春[1] 史文中[2] 刘大杰[1]

(1 同济大学测量与国土信息工程系,上海市四平路1239号,200092)
(2 香港理工大学地球资讯科技研究中心,香港九龙红[illegible])

摘 要:提出了一种数字土地信息属性数据质量的检验、度量和分析的方法。首先给出了基于简单随机抽样和分层抽样的属性数据缺陷率度量数学模型,基于该统计模型,以某工业开发区的农村土地利用现状数据为例,探讨了土地利用属性数据的质量抽样方案、质量度量和质量分析的具体思路。

关键词:土地利用现状数据;属性数据;质量
中图法分类号:P208;P271

随着土地信息系统应用的扩大和分析功能的扩充,对数据的要求也越来越高,尤其是土地信息系统中的属性数据在一定程度上涉及法律产权关系,数据本身常常会引起法律上的纠纷。对于空间数据的不确定性以及质量控制问题,文献[1,2]对此进行过详细和系统的研究与分析;针对土地利用数据的质量问题,许多学者对基于遥感的土地利用现状的边界确定和分类的精度作了大量的研究[3];对由大比例尺数字图综合产生的土地利用现状图,文献[4]也对其质量检验和控制作了分析;对于地籍图,文献[5]则探讨了图上面积和实地登记面积的一致性处理的平差方法。

然而,已有的许多研究内容主要是针对土地利用现状的图形数据,事实上,土地利用信息中,大量的是对土地进行描述的属性数据,对这些属性数据进行质量检验和分析也是整个土地利用数据质量控制中不可缺少的一部分。文献[6]对数字地图生产中属性数据精度的度量和质量控制作了初步的描述。

1 属性数据质量缺陷率度量模型

1.1 缺陷率简单抽样模型

缺陷率的基本模型建立在简单随机抽样的基础上。设有某一属性数据集合为 X,它包含数据单元 N 个(即总体容量为 N),采用不回放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,故

$$y_j=\begin{cases}1, & j\text{数据有缺陷}\\ 0, & j\text{数据无缺陷}\end{cases} \tag{1}$$

式中,j 为抽样中的第 j 个抽样数据单元。则缺陷数 y 为:

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

因抽样的样本容量为 n,而抽样检验的缺陷数为 y,通常称其比值为抽样的数据缺陷率估值,记

$$\hat{u}=\frac{y}{n},\ E(\frac{y}{n})=u \tag{3}$$

若 X 的总体容量为 N,总缺陷数为 Y,则 $\hat{u}$ 是总体数据缺陷率 $u=Y/N$ 的估值。

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[7]。

考虑到缺陷数 y_i 的取值总是取1或0,可以得到缺陷率的数学期望和方差分别为[8]:

$$\begin{cases}E(\hat{u})=u\\ V(\hat{u})=V(\frac{y}{n})=\frac{N-n}{n(N-1)}u(1-u)\end{cases} \tag{4}$$

收稿日期:2003-12-10。
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)。

缺陷率的方差是属性数据抽样样本数据中缺陷率离散程度的度量, 它反映了采用缺陷率对属性数据质量进行度量的可靠性, 所以与缺陷均值作为两个主要统计指标对属性数据的质量进行度量。缺陷率均值反映数据质量的好坏, 而方差反映缺陷率指标的可靠程度。

1.2　缺陷率分层抽样模型

一般随机抽样是最基本的抽样方法, 但是对于大批量的检验总体, 有时总体各部分质量的实际分布差别很大, 对不同部分采用不同的检验程序就比较符合实际。GIS 数据质量的抽样检验中, 可引入分层概念。在对不同数据层分层抽样时, 考虑估计每一层的缺陷率, 层的缺陷率估值通过适当加权就能得到对整个总体的缺陷率的估计量。

对按比例配置的分层随机抽样, 数据总体缺陷率估计的均值与方差的估计为[8]:

$$\hat{u_w} = \sum_{k=1}^{H} W_k \hat{u_k} \tag{5}$$

$$V(\hat{u_w}) = \sum_{k=1}^{H} W_k^2 V(\hat{u_k}) = \sum_{k=1}^{H} W_k^2 (1-f_k)\frac{s_k^2}{n_k} = \frac{1-f}{n}\sum_{k=1}^{H} W_k s_k^2 \tag{6}$$

式中, $\hat{u_k}$ 表示 k ($k=1, 2, \cdots, H$)层抽样样本的缺陷率估值; $\hat{u_w}$ 表示总体样本的缺陷率估值, H 为属性数据分层的总数; W_k ($k=1, 2, \cdots, H$)是每一层的权, $\sum_{k=1}^{H} W_k = 1$; s_k^2 为各层样本缺陷率方差的估值; f 称为有限总体校正系数。

式(5)和式(6)是分层抽样条件下缺陷率的均值和方差的估计值。同样缺陷率均值反映数据质量的好坏, 而方差反映缺陷率指标的可靠程度。

2　土地利用现状属性数据质量抽样方案

笔者以某工业开发园区土地利用现状数据库为例, 来具体说明属性数据质量的抽样检验与分析。该区农村土地利用现状数据是在 1∶1 000 数字地形图、航空影像图以及外业调绘的基础上完成的。数据涉及下属 9 个行政区域范围内的 9 种用地大类。在完成土地利用现状数据库采集以后, 需要对整个土地利用信息中的属性数据的质量进行度量, 以保证良好的数据质量, 服务于土地利用信息系统的分析和统计。

2.1　一般简单随机抽样检验

由于数据量非常大, 一般随机抽样取抽样的置信水平 $\alpha=0.05$, 根据最大相对偏差原则[7, 8]确定抽样容量, 这里最大相对偏差取为 $r=0.2$。该土地利用现状属性数据表的抽样方案见表 1。

表 1　一般随机抽样方案表

Tab. 1　Scheme of Simply Random Sampling

总体数据量 N	抽样置信水平 α	最大相对偏差 r	抽样量 n
19 559	0.05	0.2	1 270

2.2　根据土地用地类型的分层抽样

由于属性数据量非常大, 同时具有区域的特点, 各种用地类型在数据库中占的比例是不一样的, 例如由于例中的城市处于江南水乡, 所以水域非常丰富, 而其表达的方法也非常丰富, 鉴于这些特点, 可以根据用地类型对数据进行分层抽样。各数据层的权重依据各层数据量在总数据量中占的比重来确定。抽样方案见表 2。

表 2　根据用地类型的分层随机抽样方案表

Tab. 2　Scheme of Stratified Random Sampling According to the Land Types

层号 k	用地类型名	各层总体数据量 N_k	各层权重 W_k	各用地类型面积/m²	各层抽样量 n_k
1	耕地	8 935	0.457	163 299 912	469
2	园地	420	0.021	1 065 601	21
3	林地	358	0.018	516 839	19
4	牧副渔业用地	1 286	0.066	12 146 884	68
5	居住及工矿用地	2 140	0.109	10 606 300	112
6	交通用地	686	0.035	2 049 240	36
7	水域	5 161	0.264	49 968 508	270
8	其他用地	252	0.013	657 106	13
9	附加用地	321	0.016	16 939 897	16
$\sum$		19 559	1.000	257 250 290	1 024

2.3　根据行政区域对土地利用属性数据分层抽样

试验区是正在开发的工业园区, 根据城市规划, 各行政区域开发的侧重点不一样, 有些区域注重开发工业用地, 有些区域注重农业用地的利用, 有些区域注重居住用地的开发等, 所以一定区域

所拥有的用地类型差别比较大。鉴于这些差别，可以考虑根据行政区域对用地进行划分，各数据层的权重也是依据各层数据量在总数据量中占的比重来确定的。抽样方案见表 3。

表 3 根据行政区域的分层随机抽样方案表

Tab. 3 Scheme of Stratified Random Sampling According to Districts

层号 k	行政区域	各层总体数据量 N_k	各层权重 W_k	各用地类型面积/m^2	各层抽样量 n_k
1	*A*	4 914	0.081	20 918 515	87
2	*B*	839	0.027	6 995 464	30
3	*C*	3	0.022	5 649 582	3
4	*D*	3 463	0.507	130 376 572	523
5	*E*	10	0.012	2 965 330	10
6	*F*	102	0.074	19 101 490	79
7	*G*	5 406	0.118	30 280 875	124
8	*H*	4 537	0.132	33 965 914	138
9	*I*	285	0.027	6 996 544	31
$\sum$		19 559	1.000	257 250 290	1 025

3 土地利用现状属性数据质量缺陷率度量和分析

属性数据缺陷的类型可以分为轻缺陷、重缺陷和严重缺陷 3 类。不同数据项其缺陷类型是不同的，因为其缺陷对整体数据质量的影响是不同的。缺陷类型的划分一方面根据国家相关数据质量标准，另一方面需要考虑系统对数据的特别要求。在本文中，综合评判了不同数据项的缺陷类型划分(表 4)。缺陷的划分有利于对属性数据质量进行更全面的评价。表 5 给出了各种抽样方法所获得的缺陷率估计结果，其中包括轻缺陷、重缺陷和严重缺陷以及总缺陷率的估计。

表 4 土地利用属性数据表字段缺陷等级

Tab. 4 Degree of disfigurement of fields in the Attribute Data Table

字段名	mslink	土地标识码	面积	一级地类代码	二级地类代码	一级地类名	二级地类名	乡镇名	村名
缺陷等级	严重缺陷	严重缺陷	重缺陷	重缺陷	重缺陷	轻缺陷	轻缺陷	轻缺陷	轻缺陷

表 5 土地利用属性数据缺陷率抽样估计结果

Tab. 5 Result of Estimating of Disfigurement Rate on Attribute Data of Land Use

抽样类型		轻缺陷		重缺陷		严重缺陷		总缺陷率	
		缺陷率	缺陷率标准差	缺陷率	缺陷率标准差	缺陷率	缺陷率标准差	缺陷率	缺陷率标准差
一般随机抽样		0.047	0.003	0.047	0.004	0.022	0.003	3.311%	0.003 2
用地类型划分的分层抽样	耕地	0.037	0.004	0.032	0.005	0.008	0.003		
	园地	0.029	0.019	0.032	0.023	0.013	0.018		
	林地	0.024	0.018	0.045	0.028	0.080	0.046		
	牧副渔业用地	0.034	0.011	0.037	0.013	0.009	0.008		
	居住及工矿用地	0.045	0.010	0.034	0.010	0.110	0.022	2.794%	0.002 4
	交通用地	0.034	0.015	0.045	0.020	0.019	0.016		
	水域	0.056	0.007	0.034	0.006	0.010	0.004		
	其他用地	0.024	0.021	0.045	0.034	0.009	0.019		
	附加用地	0.020	0.018	0.032	0.026	0.013	0.020		
行政区域划分的分层抽样	*A*	0.045	0.011	0.045	0.013	0.087	0.022		
	B	0.056	0.022	0.056	0.025	0.095	0.040		
	C	0.035	0.054	0.061	0.082	0.012	0.045		
	D	0.055	0.005	0.035	0.005	0.011	0.003		
	E	0.045	0.034	0.044	0.038	0.087	0.066	3.020%	0.002 3
	F	0.033	0.010	0.047	0.014	0.013	0.009		
	G	0.034	0.008	0.044	0.011	0.021	0.009		
	H	0.050	0.010	0.034	0.009	0.011	0.006		
	I	0.032	0.016	0.025	0.016	0.006	0.010		

由表 5 给出的结果可以看出，3 种抽样方法所获得的土地信息属性数据的缺陷率基本一致，然而由图 1 可以看出分层抽样方法比一般随机抽样方法具有更好的抽样精度。由于数据层间的质量差异不大，所以一般随机抽样方法和分层抽样的抽样精度差别不是很大。但是分层抽样方法抽

样量少于一般随机抽样方法, 所以分层抽样方法能以比一般随机抽样方法少的抽样费用而得到更好的抽样精度, 因此分层抽样方法是切实可行的。从图 1 中可以发现, 采用基于用地类型划分的分层抽样和基于行政区域划分的分层抽样方法具有基本相当的抽样精度, 但可以从不同方面反映土地信息中属性数据的质量。

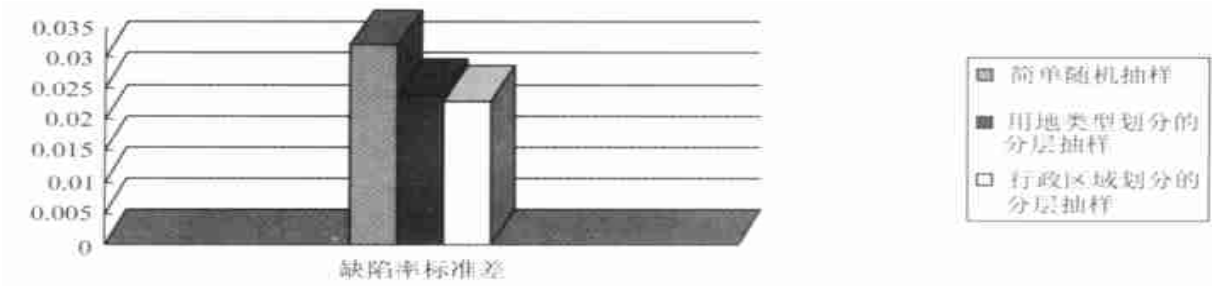

图 1　各种抽样方法抽样精度

Fig. 1　Sampling Accuracy of Various Sampling Methods

采用分层抽样方法另一个优点是可以获得各数据层的质量抽样结果, 从而可以对各数据层采用缺陷率度量其质量。图 2 是各用地类型数据分层抽样的质量结果, 图 2 中给出了各种用地类型数据的轻缺陷率、重缺陷率、严重缺陷率以及总和缺陷率, 该图能够比较清楚地反映不同缺陷类型数据的质量。从图 2 可以发现, 居住及工矿用地缺陷率的值比较高, 即相对质量比较差, 分析其原因是由于案例数据来自于一个工业开发区, 居住和工矿用地的变更非常频繁, 地块的划分也比较细, 所以这些用地属性数据质量相对比较差是正常的, 采用缺陷率模型对这一情况也能反映出来。

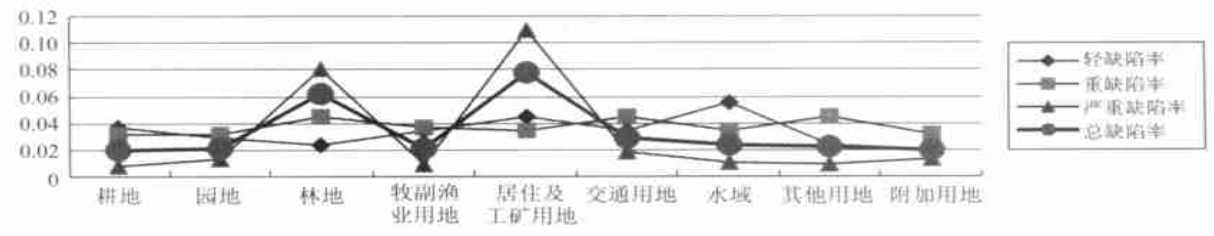

图 2　用地类型数据分层抽样质量估计

Fig. 2　Rate of Disfigurement for Different Land Types

同样, 图 3 是各行政区域数据分层抽样的详细质量结果, 图 3 能够确定不同行政区域采集数据的质量, 数据质量也从 3 种缺陷类型的数据上综合体现。从图 3 上发现, 行政区域 A、B 和 E 也具有相对较差的属性数据质量, 其原因在于这些区域是该工业园区首期开发区域, 是引资建设的重点区域, 所以这些区域的土地利用率非常高, 土地利用的属性经常发生变更, 从而导致这些区域的属性数据的质量相对比较差。

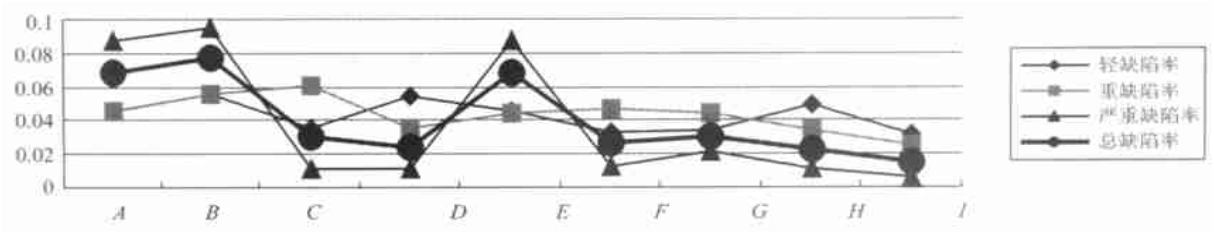

图 3　行政区域数据分层抽样质量估计

Fig. 3　Rate of Disfigurement for Different Districts

参　考　文　献

1　刘大杰, 史文中. GIS 空间数据的精度分析与质量控制. 上海: 上海科学技术文献出版社, 1999

2　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

3　Congalton R G, Green K. Assessing the Accuracy of Remotely Sensed Data: Principles and Practices. London:

Lewis Publishers, 1999
4 刘 春, 李 乔, 刘妙龙, 等. 基于 1∶1 000 数字地形的土地利用数字现状成图. 中国土地科学, 1999, 13(1): 42~44
5 刘大杰, 童小华. 附加尺度参数的地籍宗地面积处理. 同济大学学报, 2002, 30(4): 490~494
6 Liu C, Shi W Z, Liu D J. Quality Assessment of Attribute Data for Digital Maps Based on the Statistical Rate of Disfigurement Method. The 5th International Symposium on Spatial Accuracy. Melbourne, Austria, 2002
7 刘 春, 刘大杰, 史文中. 基于缺陷率的 GIS 属性数据的质量限差探讨. 同济大学学报, 2002, 30(1): 1 355~1 360.
8 刘 春, 史文中, 刘大杰. GIS 属性数据精度的缺陷率度量统计模型. 测绘学报, 2003, 32(1): 36~41

第一作者简介: 刘春, 博士, 博士后。主要从事 GIS 空间数据质量的基础理论研究和应用系统的开发。近年来参与完成 863 空间信息移动服务综合技术项目、国家自然科学基金项目(40171078)以及香港特别行政区资助项目等多个纵向项目。发表学术论文 40 余篇。
E-mail: liuchun@mail.tongji.edu.cn

Quality Control for Attribute Data in Digital Land Information

LIU Chun[1] *SHI Wenzhong*[1] *LIU Dajie*[2]
(1 Department of Surveying and Geo-Informatics, Tongji University, 1239 Siping Road, Shanghai 200092, China)
(2 Advanced Research Center for Spatial Information Technology, The Hong Kong Polytechnic University, Kowloon, Hong Kong)

Abstract: The rate of disfigurement is put forward to measure the accuracy of attribute data based on the sampling inspection. To a real GIS application as land information system, it is basic to collect the land use data and to establish its spatial database when developing the application of land use information system. Generally, the attribute data is regarded as an emphasis of the whole land use data set for its main content of the spatial analysis. So the enhancement of quality control for attribute data can be propitious to set up integrity land information system and provide the reliable service by land use analysis. The inspection, determination and analysis of the attribute data quality are discussed. The rate of disfigurement model is described in detail based on the simple random sample and the stratified sample.

Key words: land use data; attribute data; quality

About the first author: LIU Chun, Ph. D. postdoctoral in the professional of sea science. His research interests include error modeling for GIS and remote sensing data, application of GIS and so on. He has published over 40 papers.
E-mail: liuchun@mail.tongji.edu.cn

(责任编辑: 晓晨)

关于刘经南担任本刊主编的公告

经主办单位武汉大学研究决定, 中国工程院院士、武汉大学校长刘经南教授自 2004 年起担任本刊主编。同时, 编委会亦将作相应调整。

特此公告。