文章编号: 1000-050X(2002)06-0627-05　　文献标识码: A

城市人口 GIS 中数据的不确定性研究

柳宗伟[1]　王新洲[1]　陈顺清[2]

(1　武汉大学测绘学院, 武汉市珞喻路 129 号, 430079)
(2　广州城市信息研究所有限公司, 广州市天河软件园建工路 3 号, 510665)

摘　要: 针对系统数据源、数据模型的不确定性以及分析过程中引入的不确定性等问题进行了系统的研究, 并归纳了各种不确定性的影响模型, 进而提出了一套人口 GIS 数据优化建模以及分析过程中人口密度中心建模的思想和实用方法, 用于克服人口 GIS 中数据的不确定性的影响。

关键词: 人口 GIS; 不确定性; 优化建模; 密度中心
中图法分类号: P208; P207

近年来, 有关空间数据不确定性模型、GIS 数据质量控制等很多问题都得到了深入的研究[1~4]。总体看来, 有关 GIS 数据质量的研究集中在位置数据的不确定性上, 对属性数据不确定性的研究较少[5~6]。对于以人口普查信息为主的城市人口 GIS, 其属性数据质量的重要性甚至远大于空间数据, 因此, 有关城市人口 GIS 中数据的不确定性研究必须统筹考虑空间数据和属性数据的不确定性问题。

城市人口 GIS 中, 作为系统基础数据的城市电子地图存在着复杂的位置与属性的不确定性[1,2], 而人口普查作为一种社会调查方法也不可避免地导致了某些普查内容的不确定性[7]。

1　数据源的不确定性

城市人口 GIS 的数据源主要包括城市矢量电子地图、普查区界线数据、人口普查数据, 以及其他相关的工业、社会经济等方面的数据[8]。其中影响城市矢量电子地图的数据质量的因素包括地图测量、绘制不确定性, 图纸变形, 数字化不确定性, 数据转换不确定性等许多方面[9]。本文主要讨论了人口 GIS 中普查区界线位置及其属性数据的不确定性。

1.1　普查区界线的不确定性

人口普查区一般按行政界线进行分级划分。以城市为例, 普查区分为区、街道、居委会等不同的等级[7]。普查小区是人口 GIS 中最基本的地理单元。

普查小区的获得主要有两种方法: 在已有地形资料上进行划分和人口普查员绘制普查小区草图。对于在已有地形资料上进行划分, 位置不确定性产生的原因主要有: ①地形资料的质量问题; ②普查员的判读出错; ③数字化人员的失误。对于人口普查员绘制普查小区草图, 位置不确定性产生的原因主要有: ①数字化底图的质量问题; ②普查区草图质量问题; ③数字化人员的失误。

要减少这方面的影响, 必须采用高质量的地形资料, 依据严格的规范进行数据录入。图 1 所示为采用最新正射影像图进行普查区界线录入的实例; 图 2 所示为针对复杂区域结合大比例尺地形图进行普查区界线的辅助判读、定位。

1.2　人口普查数据的不确定性

1.2.1　有关人员工作失误造成的属性不确定性

有关人员工作失误可能造成普查区划分重、漏和普查区编号填错等情况, 导致图形数据与属性数据无法一一对应。这些错误可以在普查区界线数字化后, 通过有关的 GIS 软件进行可视化浏览、查错。

1.2.2　人口普查中特殊情况造成的属性不确定性

人口普查过程中还存在一些特殊情况和特殊规定[7]。例如虚拟居委会, 是由特殊人员(如被劳教人员等)组成的没有明确地理位置的单位,

收稿日期: 2002-07-13。
项目来源: 国家高技术研究发展计划(863 计划)资助项目(2001AA135081)。

图 1 以最新正射影像进行界线录入

Fig. 1 Getting the Digital Census Borderlines on the Latest Orthophoto

图 2 以 1∶500 地形图进行界线的辅助定位

Fig. 2 Getting the Digital Census Borderlines on the Large Scale Relief Map

在相关的普查区草图上没有明确的地理范围，这也会导致属性数据的不确定性。在系统建设中，应根据国家有关规定对以上特殊情况进行适当调整，如将虚拟居委会的普查数据在本街道内进行均分。

1.2.3 人口数据建库产生的属性不确定性

人口数据建库也可能产生属性数据的不确定性。例如属性数据格式转换(如由文本文件向数据库转换)产生差错；另外，数据按专题(年龄、职业等)、按不同等级普查区(区、街道等)进行汇总时也可能出错。在系统建设过程中应从不同的角度对有关数据进行交叉检核，保证属性数据的正确性。

2 数据模型的不确定性

在人口 GIS 的应用中进行人口查询或分析时，所选范围与普查小区一般存在相离、相交、包含三种空间关系如图 3 所示。在确定被选择的普查区面积和属性时，可能因普查区内人口分布的情况不同产生数据模型的不确定性。

图 3 所选范围与普查小区存在不同的空间关系

Fig. 3 Three Kinds of Spatial Relationships Between the Selected Area and the Census Areas

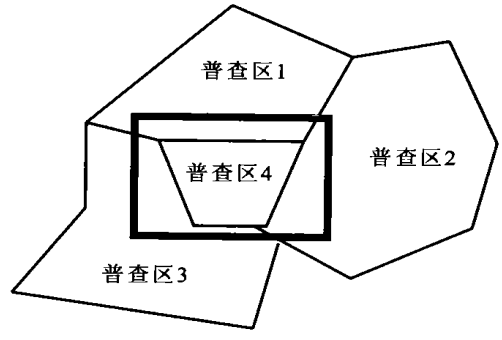

图 4 所选范围与普查小区相交的情况

Fig. 4 Selected Area Intersects the Census Areas

2.1 普查小区内人口分布均匀

普查小区内人口分布均匀的情况主要出现在城市老城区的密集的居民区、城市住宅小区等区域。以 P_i、P_j 分别表示与选择范围相交、包含的普查小区的人口数量，S_i 表示与所选范围相交的普查小区的面积(如图 4 中普查小区 1、2、3 的面积，不包括普查小区 4)，s_i 表示相交普查小区在所选范围内的面积(如图 4 中普查小区 1、2、3 在黑框内的面积)。以人口信息汇总为例，如果将所有与选择范围包含、相交的普查小区的属性进行汇总，结果势必会与实际情况有明显的出入。

例如，汇总人口数量为：

$$N_1 = \sum P_i + \sum P_j$$

实际的人口数量为：

$$N_2 = \sum (s_i / S_i) P_i + \sum P_j$$

汇总结果与实际人口数量的差别为：

$$\Delta N = \sum [(S_i - s_i)/S_i]\ P_i$$

产生的相对误差为：

$$E = \Delta N / N_2$$

因为 $S_i \geqslant s_i$，差别值 ΔN 总为正，即汇总结果总是

比实际人口多。可见, 在普查小区内人口分布均匀的情况下, 要避免这一不确定性问题, 必须精确计算相交区域面积 s_i, 并按所选的实际面积进行属性数据分配。

2.2　普查小区内人口分布局部均匀

针对学校、工厂、大面积绿地、水系等情况, 人口往往分布在普查小区的某一局部(或多个小区域), 其他区域为非住宅区。从整个普查小区来看, 人口分布不均匀, 但把区域分成明显的住人区、非住人区后, 则可以认为人口分布局部均匀, 如图 5 所示。设有人区面积分别为 S_{i_1}, 包含在选择区域内的有人区面积为 s_{i_1}, 其他设定与上面一致。在这种情况下, 汇总人口数量可用上述实际人口数量的计算公式计算。汇总人口数量(按人口均匀分布)为:

$$N_1 = \sum (s_i / S_i) P_i + \sum P_j$$

实际的人口数量为:

$$N_2 = \sum (s_{i_1} / S_{i_1}) P_i + \sum P_j$$

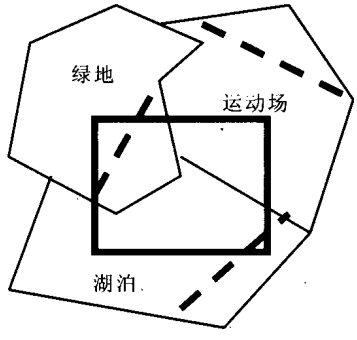

图 5　人口分布局部均匀

Fig. 5　Population Distrubutes Equally in Parts of the Census Area

汇总结果与实际人口数量的差别为:

$$\Delta N = \sum (s_i / S_i - s_{i_1} / S_{i_1}) P_i$$

产生的相对误差为:

$$E = \Delta N / N_2$$

在这种情况中, 汇总结果与实际人口数量的差别 ΔN 可能为 0(汇总结果与实际人口相同), 可能为负(汇总结果小于实际人口), 也可能为正(汇总结果大于实际人口)。具体到某一个普查小区, 当面积比值 s_i / S_i 大于 s_{i_1} / S_{i_1} 时, 该小区的汇总人口偏大, 其他情况可以类推。从上式还可以看到, 人口分布均匀的情况可以看成人口分布局部不均匀的特例, 当 $S_i = S_{i_1}$ 时, 两种情况一致。

针对这一情况, 笔者提出了细分普查小区的方法来加以改进。细分即在普查小区范围内将分散的居民区勾绘出来, 并按居民区的规模将普查信息进行适当的分配。其中地理编码可以专门制定细分区界层的编码方式, 也可以给无人区某一特殊编码, 例如"XXX99", 进行标识。这样处理可以优化系统以普查小区为最小地理单元的局限, 能够确定 s_{i_1} 和 S_{i_1} 的精确数值, 也能保证属性数据的合理分配。

2.3　普查小区内人口分布不均匀

普查小区内人口分布不均匀的情况在城市中也普遍存在, 例如, 住宅楼与办公楼、厂房在同一普查小区的情况, 就会产生人口分布的不均匀。如图 6 所示, 从表面看普查小区范围内房屋分布均匀, 但实际上由于房屋性质的不同导致其人口分布并不均匀, 这种情况下采用上面提到的按所选面积进行属性分配的方式同样也会产生不确定性。以 S_{i_2} 表示普查小区内住宅区面积、s_{i_2} 表示该普查小区包含在所选范围以内的住宅区面积, 则汇总人口数量(按人口均匀分布)为:

$$N_1 = \sum (s_i / S_i) P_i + \sum P_j$$

实际的人口数量为:

$$N_2 = \sum (s_{i_2} / S_{i_2}) P_i + \sum P_j$$

汇总结果与实际人口数量的差别为:

$$\Delta N = \sum (s_i / S_i - s_{i_2} / S_{i_2}) P_i$$

产生的相对误差为:

$$E = \Delta N / N_2$$

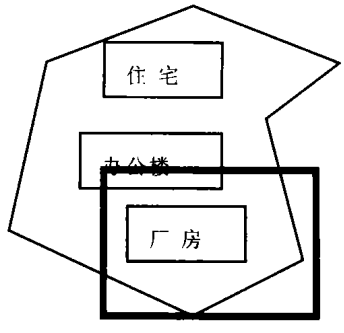

图 6　人口分布不均匀

Fig. 6　Population Distributes Unequally in a Census Area

汇总结果与实际人口数量的差别 ΔN 可能为 0(汇总结果与实际人口相同), 可能为负(汇总结果小于实际人口), 也可能为正(汇总结果大于实际人口)。具体到某一个普查小区, 当面积比值 s_i / S_i 大于 s_{i_2} / S_{i_2} 时, 该小区的汇总人口偏大, 其他情况可以类推。从上式还可以看出, 上文所述

的人口分布均匀和局部均匀等情况都可以看成人口分布不均匀的特例，当 $S_{i_2}=S_i$ 时，即为人口分布均匀；而当 $S_{i_2}=S_{i_1}$ 时，则为人口分布局部均匀的情形。对于人口分布不均匀的情况，可以结合城市电子地图，借助相关的地物属性，通过叠置分析来获得普查小区具体的人口分布情况。按这种思路可以建立更灵活的数据模型，精确确定 s_{i_2} 和 S_{i_2} 的数值，并能够处理各种复杂的人口分布情况下的数据不确定性问题。

3 分析过程中引入的不确定性

在人口 GIS 中，普查区的几何中心通常被作为属性数据的点位。以图 7 所示的情况为例，在分析某活动中心为居民提供服务的效果时，可采用居委会的几何中心作为居民位置点，把相关的人口信息作为该点的属性进行分析。而这种处理方法可能在分析过程中引入中心点位的不确定性以及相关点位属性的不确定性。

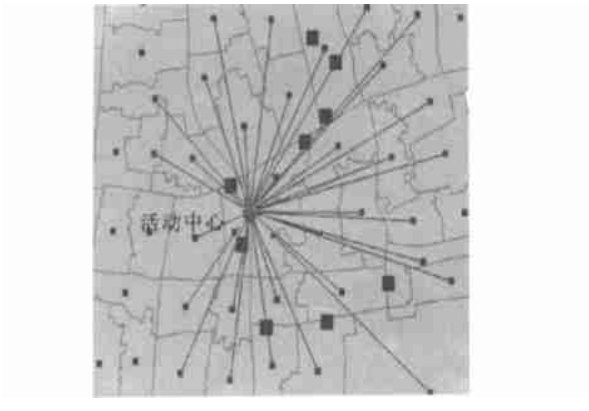

图 7 计算居民到活动中心的距离

Fig. 7 Measuring the Distance Between the Location of People and the Entertainment Center

3.1 中心点位的不确定性

如图 8 所示的某居委会，由于人口分布整体不均匀导致其几何中心和人口密度中心点位存在较大的差异，在分析中不进行区别会产生不确定性。

在有关效果评价的分析中经常采用权重模型 P_i/D_i^2，其中 P_i 表示居民点人口；D_i 则表示居民点到活动中心距离[9]。设活动中心到居民区几何中心的距离为 D_{i_1}，活动中心到居民区人口密度中心的距离为 D_{i_2}，则采用几何中心的得到的权重模型值为：

$$R_1 = P_i/D_{i_1}^2$$

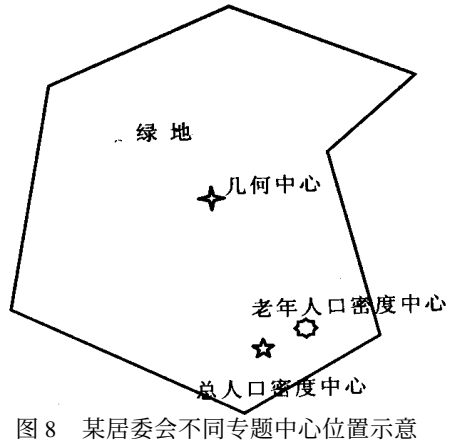

图 8 某居委会不同专题中心位置示意

Fig. 8 Different Centers Locate in Different Locations Within the Same Census Area

采用密度中心的权重模型值为：

$$R_2 = P_i/D_{i_2}^2$$

两者差别为：

$$\Delta R = P_i(D_{i_2}^2 - D_{i_1}^2)/(D_{i_1}^2 \times D_{i_2}^2)$$

中心点位不确定性的影响可以分为以下几种情况：

1）当 $D_{i_1}=D_{i_2}$ 时，$\Delta R=0$，即居民区几何中心与人口密度中心一致时（人口分布均匀），不存在中心点位的不确定性；

2）当 $D_{i_1}>D_{i_2}$ 时，$\Delta R<0$，活动中心到几何中心比到人口密度中心的距离大，采用几何中心计算的权重模型值偏小；

3）当 $D_{i_1}<D_{i_2}$ 时，$\Delta R>0$，活动中心到几何中心比到人口密度中心的距离小，采用几何中心计算的权重模型值偏大。

3.2 相关属性的不确定性

在实际分析中，不同的分析主题必须明确对应的相关属性。例如活动中心可以按年龄分为少儿、青年、老年活动中心；也可以按类型分为妇女活动中心、铁路职工活动中心等。进行有关老年活动中心的分析时，点位属性应该选择老年人口，笼统地采用总人口会给分析结果带来偏差甚至错误。

这里要特别强调的是，属性不确定性的影响甚至比位置不确定性的影响更大。设居委会总人口为 P_i，老年人口为 P_{i_1}，则该居委会的人口差 ΔP_i 为：

$$\Delta P_i = P_i - P_{i_1}$$

一般情况下，上式中 ΔP_i 很大，即总人口 P_i 要远大于老年人口 P_{i_1}，属性选择不当会导致分析结果完全错误。

点位与属性的不确定性密切相关。在分析老

年人口与老年活动中心的相关情况时, 点位中心应选择老年人口(如 60 以上)分布中心, 属性当然采用相应的老年人口数。有关专题的人口密度中心坐标采用专题人口数加权计算获得。以居委会老年人口分布中心坐标计算为例:

$$X_{i_1} = \sum (x_j p_j) / \sum p_j$$

$$Y_{i_1} = \sum (y_j p_j) / \sum p_j$$

式中, (X_{i_1}, Y_{i_1})为老年人口分布中心坐标, (x_j, y_j)和 p_j 分别为居委会内相关普查小区人口分布的中心坐标和老年人口数。进行其他专题分析时, 应采用类似的中心坐标计算公式和相对应的专题属性, 例如分析有关少儿的有关情况时, 上式中(x_j, y_j)应选用少儿人口密度中心坐标, p_j 则采用少儿人口数。其他情况可以根据以上模型进行类推。

参 考 文 献

1 Leung Y, Yan J P. A Location Error Model for Spatial Features. INT. J. Geographical Information Science, 1998, 12(6): 607 ~ 620

2 Crosetto M I, Tarantola S. Uncertainty and Sensitivity Analysis: Tools for GIS-based Model Implementation. INT. J. Geographical Information Science, 2001, 15(5): 415 ~ 437

3 刘大杰, 华　慧. GIS 线要素不确定性模型的进一步探讨. 测绘学报, 1998, 27(1): 46 ~ 49

4 史文中. 空间误差处理的理论和方法. 北京: 科学出版社, 1998

5 刘文宝, 邓　敏, 夏宗国. 矢量 GIS 中属性数据的不确定性分析. 测绘学报, 2000, 29(1): 76 ~ 81

6 Burrough P A, Mcdonnell R A. Principles of Geographical Information Systems. London: Oxford University Press, 1998. 231 ~ 240

7 广东省第五次人口普查领导小组办公室. 第五次全国人口普查工作细则, 2000

8 王新洲, 柳宗伟, 陈顺清. 城市人口地理信息系统建设模式探讨. 武汉大学学报·信息科学版, 2001, 26(3): 226 ~ 231

9 陈顺清. 城市增长与土地增值. 北京: 科学出版社, 2000

10 闫　正. 城市地理信息系统标准化指南. 北京: 科学出版社, 1998

作者简介: 柳宗伟, 博士生。现主要从事人口 GIS 和选址空间决策支持系统研究。

E-mail: zongwei1163@. com

Uncertainties of Data in Urban Census GIS

LIU Zongwei[1]　*WANG Xinzhou*[1]　*CHEN Shunqing*[2]

(1　School of Geodesy and Geomatics, Wuhan University, 129 Luoyu Road, Wuhan, China, 430079)

(2　Digital Cities Institute, High Technology Industry Development Zone, Tianhe, Guangzhou, China, 510665)

Abstract: In recent years, many researches are focused on the uncertainties of GIS data. It can be found that most of these researches are focused on the uncertainties of spatial data while a few are on the attribute data. In this paper, the authors try to integrate the uncertainties of spatial data and attribute data to analyze some problems on the data uncertainties in Census GIS. This paper suggests considering the uncertainties of spatial data and attribute data at the same time in GIS analysis and application. On the basis of the characters of Census GIS, the authors concludes three kinds of data uncertainties in Census GIS. Then, some related contents, reasons, and the influence of these uncertainties are discussed in detail. On the basis of the data uncertainties discussed in this paper, the authors propose some methods to resolve the related problems. Accordingly, some concise formulae are used to illustrate the uncertainties and the related models.

Key words: census GIS; uncertainty; model optimization; density center

About the author: LIU Zongwei, Ph. D candidate. His research interests are urban census GIS and SDSS for location selection.

E-mail: zongwei1@163. com