基于回溯地理编码方法的人口研究

邹 海,周 倩
(安徽大学 计算机科学与技术学院,安徽 合肥 230039)

摘 要:文中的目的是描述在人口研究上使用多相地理编码回溯建立地理信息系统的方法。介绍了回溯建立一个地理信息系统的挑战并提出了提高地理编码成功率的实用技巧。使用4800个人的详细情况作为基准数据,在多相过程中完成这些人的地理编码。经过初步删除不能地理编码的地址(38例),使用Arc GIS 96%地址被编码。在这些样本的记录中使用数据抽象的回溯方法,利用地址的核查,除19个地址外,成功地取得了所有地理编码,近99%(4743例)被地理编码。

关键词:地理信息系统;地理编码;多相;回溯
中图分类号:TP39 **文献标识码**:A **文章编号**:1673-629X(2011)11-0201-03

Population Research Based on Retrospective Geocoding

ZOU Hai, ZHOU Qian
(School of Computer Science and Technology, Anhui University, Hefei 230039, China)

Abstract: The main purpose of is to describe the method of using the multiphase geocoding to retrospectively create a GIS. It also introduces the challenge of retrospective creating a GIS and advances some practical tips that improve the success. 4800 people's detailed information was used as baseline data, finish this people's geocoding in multiphase process. Using Arc GIS 96% addresses were coded after delete the addresses which can not be coded. An interactive method using data abstraction from sample records, use of verification of existence of address, yielded successful coding of all but 19 addresses. Overall, nearly 99% of the addresses were coded.

Key words: GIS; geocoding; multiphase geocoding; retrospective

0 引 言

环境正日益被视为与个人和邻里健康[1~3]有决定性的联系。地理编码[4~7]和地理信息系统是调查区域环境和健康[2,8]之间的空间关系必不可少的工具。现在有助于健康的重点生态和环境因素,可以作为地理信息系统技术进步的结果被收集、管理和评估。

许多人口研究已经接受了使用绘图作为一种方法来调查空间聚类和疾病及健康结果的决定性作用,其他人可能会在回溯地理分析上有兴趣,他们的资源或研究设计的功能排除了基线上这些数据的收集。很少有研究所描述的方法运用地理编码和地理信息系统[9]。缺乏方法论细节限制了地理信息系统数据可靠性和有效性的使用和评价,阻碍了多方面的环境对健康的检查。为了拓宽对空间数据有效性[10]的讨论,研究人员也开始报告关于地理编码匹配结果或准确性更详细的介绍。

文中描述了多相地理编码方法和用来评估基线样本特点的GIS结果以及决定大型人口研究中的城市和农村地理位置,此大型人口研究在基线处没有收集地理数据。提供了可用于其他研究中的方向以增加空间样本大小和提高地理数据的可靠性。一个关于地理编码现有数据集的详细的协议将变得越来越有价值,就像GIS方法和技术的进展被用于流行病学和其他研究一样越来越有价值。

1 地理编码概述

首先要创建一个可靠的地理数据库,它可用于在未来的研究中检查空间和周边环境的关系。

目标是:

(1)成功的地理编码95%或更大的地址到正确的样本块组;

(2)为每一个地理编码分配纬度和经度坐标;

(3)不增加资源和成本并最小化位置误差。

使用了多相过程,其中包括:确保全部样本地址的准确性,地理编码地址的坐标,验证程序,分类和地理参照样本到样本块组上,以及连接社会人口特征到每个样本块组上。

收稿日期:2011-03-24;**修回日期**:2011-06-27
基金项目:国家科技重大专项资助项目(2008ZX05039-004)
作者简介:邹 海(1969-),男,博士,副教授,硕士研究生导师,主要研究领域有数据挖掘与信息检索、中间件理论与技术、工作流理论与技术、图像处理技术。

1.1 地址验证

包含邮编或农村路线的不能得到改善的地址被认为有不正确的定位和地理编码错误的可能性。按照研究记录地理编码误差,通过使用邮编的地理编码,如果没有可利用的街道地址也被视为不能接受的地址精确度,并且无法管理。

在地理编码之前,地址被分类和检查明显的数据输入和拼写错误。信息丢失,不完整和不准确的地址将在数据库中被识别,并且数据表被检索到抽象的地址信息。如果需要的话,可以用网络来获取匹配的街道地址信息。一个先验的决定是为 GIS 基线排除住在区域地图之外的家庭子研究样本的地理编码,因为这些人数很少(13 例)并且分布在全国。邮局和不能得到改善的农村路线的地址可能由于位置错误而被排除,和不属于研究区的地址一样被排除。

1.2 街道地址地理编码

ArcGIS 9.2 可以查询数据,创建地图和多用户使用,也可以用于许多行业包括环境、政府和公共卫生机构,所以用它来进行地理编码。使用 1983 年北美地理坐标系统的美国等距圆锥投影拓扑集成地理编码和参照地图文件。

地理信息系统街道图的地址定位器,地理编码引擎执行地址的标准化和基于参数的匹配,和 ESRI 2005 街道地图被用来给每个地址分配一个地理位置。Arc GIS 提供两种地理编码方法:基于地址匹配参数的自动匹配和交互式匹配,其中地址可以进行审查,并根据需要纠正地址。文中的地理编码是用批处理匹配,然后再互相重新匹配。样本被地理编码并分配 X,Y 坐标,使用 70% 的拼写敏感性,10% 最小候选分,80% 的最小匹配值为参数。匹配值是基于地址属性与地理编码的地址定位器匹配的好坏来设置的,并受拼写、门牌号、方向属性,和街道、街道名、城市和邮编的不同的影响。最好的匹配值为 100%。在地理编码期间分配坐标。简言之,在地理编码时,ESRI 街道图使用地址定位器和地址属性,索引来找出非空间地址在地图上的位置。从地址定位器中的空间数据推算出与已知位置坐标相一致的点,包括街道交叉口处的 XY 坐标和街道终点的 XY 坐标和地图上的空间。因此,每一个点的地理编码地址是一对 XY 坐标。

如果认为未匹配的地址是可以接受的(例如,如果街道名称拼写错误),就分别对其进行审查和匹配。匹配值低于 80% 将会在互相重新匹配过程中被评估。使用街道图定位器将未匹配的文件输出并重新匹配。试图进一步定位包括手动编码的每一个未匹配地址:

(1)为基线检查购买区域纸质街道图;

(2) 核实现实存在的地址与文件上记录的地址是否匹配的网络地图。

1.3 验 证

第二个地理编码过程是不记录值丢失和不完整值,给每个地址分配地理位置。这就允许了两个作为结果的 GIS 相比较,促成了第一个地理编码过程的验证和精确性。通过可用的 XY 坐标和每一个过程的匹配值来实施 25% 的质量控制检查,通过 XY 坐标计算出距离的精确性。

1.4 块组中参加者的分类和地理参照

基于城市区域的特点每个块被分为市区或非市区。任何块不被列为城市就被认为是农村。因为分析单位是块组,如果块组内所有块同样被分类,块就被分为城市或者农村。由于块组分为城市块组和农村块组,基于城市或农村块数量上的优势定义了混合块组;农村混合块组包含较多的农村块。最后每一个样本以及它们的基准数据被地理参照到块组上。

1.5 连接邻里社会人口特点

选择的社会人口特点块组通过特别的块组代码被连接到 GIS 中,生成一系列的地图来阐释研究区域的社会人口分布特点。

2 地址匹配

缺失邮箱地址的样本(12 例),邮箱地址无法通过提取或使用电话号码反向查找记录得到改善(26 例),或在第一阶段被排除在区域之外地区的地址(14 例),剩下的 4748 例在第二阶段地理编码。大多数删除的邮箱地址被定位到市区的邮政局(18 例)。批处理和交互式处理匹配后,4092 个地址(86%)被匹配了,匹配值为 80% 或者更高,说明了样本地址与地址定位器的匹配很好,453 个地址被匹配的匹配值低于 80%。203 例不能匹配的项中 32 个额外的地址可以使用街道地图文件而得到匹配。剩下的 171 例手工匹配。这其中 158 例被手工标注在区域内,8 例被标注在区域外,还有 5 例不能被标注。最后两组(13 例)从 GIS 中移除。最后 GIS 中包括 4735 例(见图 1)。

对那些不能被匹配的地址使用批处理,交互式,地图文件集匹配以及手工绘制,64% 被定位在城市块组中,13% 定位在农村,12% 定位在混合农村区域还有小于 11% 被定位在混合城市块组中。城市地址使用邮政服务地址很容易被找到。混合城市和混合农村地区地址找起来稍微困难,特别是如果有细分出来的新街道。只包括 13% 绘制地址的农村地址,需要使用购买的地图找到。

没有记录抽样的独立地理编码过程产生一个非常低的成功率:4080 例(85%)被成功的地理编码并且匹配值为80% 或者更高,234例被地理编码的匹配值低

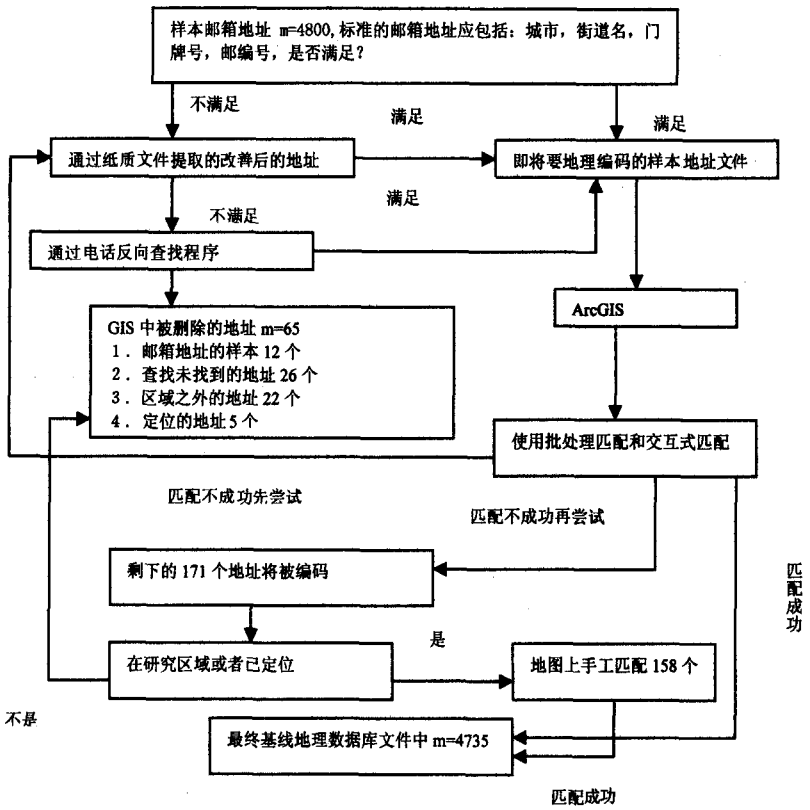

图 1　人口研究中地理编码过程

于 80%。其余的记录无法匹配并且不在 GIS 的独立地理编码过程中。25% 质量控制检查结果是 100% 的匹配值,距离等于 0。样本人口分布见图 2。

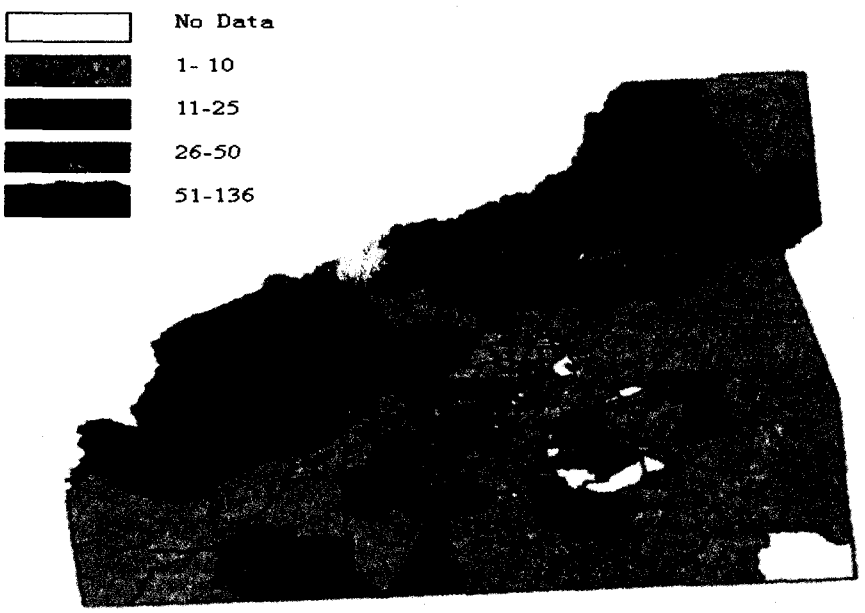

图 2　样本人口分布图

研究区域中 309 个块组中的大多数被分为城市(205 块,66.3%),剩下的几乎全被分为农村(36 块,11.6%),混合农村(38 块,12.3%)和混合城市(30 块,9.7%)。3567 名住在城市块中,489 名住在农村块中,其次混合农村块 417 人,城市混合块 262 人。

3　实验分析

越来越多关于环境和城镇对健康的作用的关注成为有效可靠地理数据集所必需的。这个方法的详细说明和创建 GIS 的结果为完成大量流行病学研究回溯编码提供了资源。这种方法的细节对于评估用在周围环境和空间分析地理数据质量是很重要的。回溯创建 GIS 的阶段,过程和结果的描述为未来的研究提供了基础。这个 GIS 数据为地方缓解及其对健康表现不同的作用的理解提供了机会。

通过规定删除不可编码和不可接受的地理编码的地址,地址删除之后在 Arc GIS 中 96% 通过自动匹配和缺省设置编码。随着多级利用和方法描述,几乎 99% 的样本被编码。

地理参照样本的地理编码数据到块组上提供改善地图的机会,改善地图以评估样本区的分布情况。虽然这不是在每个块组中获得平衡样本的目标,却是获得一个调查心血管疾病的基于人口的有代表性的样本的目标,GIS 提供了一个成功策略的简介。把样本块组分为城市、农村、混合城市/农村,并允许将其社会人口特征与相应区域人口信息进行比较。样本被地理参照到研究区域近 93% 的块组中。

通过 GIS,整个生命过程中环境对健康的影响的复杂分析将得到提高,跟踪样本的流动性以及获得其在具体生活点的实际地址可以加强纵向的研究设计。收集这些数据是困难的。每年对样本进行的跟踪和监视将允许对时刻改变的地址地理编码。

4　结束语

文中描述了在大型人口研究中创建 GIS 的许多挑战并提出了提高匹配成功率的切实可行的解决方案。使用这多种策略几乎 99% 样本的地址被地理编码[11]了。无论在研究规划阶段还是在回溯使用现有数据阶段,其他研究人员都可以利用这些解决方案在大型人

(下转第 207 页)

控制炸弹的投放。

电路图如图 5 所示。

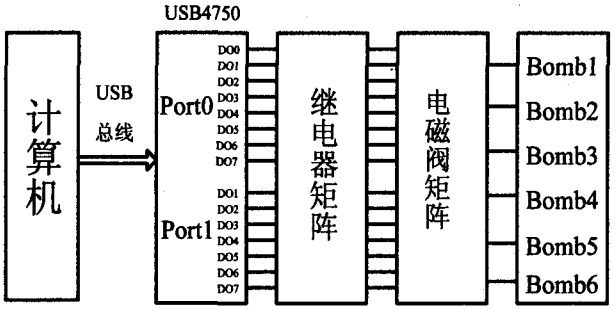

图 5　电路图

4　结束语

通过利用 GL Studio 对该项目的开发,发现利用 GL Studio 建模不但形象逼真、效率高、速度快,而且生成代码可读性好,确实适用于多仪表多旋钮系统的仿真开发。将其应用于某型飞机外挂物管理仿真系统的开发设计中,取得了良好的应用效果。

参考文献:

[1] 张继夫,陈　蕾,邓　华,等. 基于面向对象技术的飞行仿真研究[J]. 计算机技术与发展,2010,20(7):211-215.

[2] 郭奇胜,董志明,单家元. 系统仿真学报[M]. 北京:国防工业出版社,2006:168-175.

[3] 姚合生. 基于 OpenGL 的快速原形制造系统仿真[J]. 计算机技术与发展,2008,18(11):176-179.

[4] 赵越超,李忠科,王　勇. 基于 OpenGL 的三维牙颌模型可视化研究[J]. 计算机技术与发展,2008,18(1):119-125.

[5] MultiGen-Creator Getting Started[M]. USA:MultiGen-Paradigm Inc,2003.

[6] Distributed Simulation Technology Inc. GL Studio User's Guide[M]. USA:Distributed Simulation Technology Inc,2004.

[7] GL Studio Version 2.1 API Documentation[M]. USA:Distributed Simulation Technology Inc,2003.

[8] 李　东,吕维涛,雷　震,等. 基于 GL Studio 的多仪表综合显示面板仿真[J]. 电脑知识与技术,2010,6(3):674-676.

[9] 李　秀,宋丽梅,周兴明,等. GL Studio 在直升机仪表面板仿真中的应用[J]. 计算机技术与应用,2009,29(2):42-44.

[10] 赵剑秋,朱　明. 用 VC 实现控制面板应用程序[J]. 计算机技术与发展,2006,16(6):110-112.

[11] 雷　超,陈　伟. 一种火控雷达终端显示器的仿真实现[J]. 计算机技术与发展,2008,18(4):195-198.

[12] 高　颖,邵亚楠,郑　涛,等. GL Studio 在飞行座舱模拟器中的仿真研究[J]. 弹箭与制导学报,2008,28(1):257-260.

[13] 于　辉,赵经成,付战平,等. GL Studio 虚拟仪表技术应用与系统开发[M]. 北京:国防工业出版社,2010:105-106.

[14] 狄旭明,吴建平. USB 多路数据采集器[J]. 中国测试,2010,36(2):81-96.

[15] 郎　峥,李晓峰. 基于 USB 的高精度多通道数据采集卡设计[J]. 电子科技,2010,23(2):86-89.

(上接第 203 页)

口研究中创建地理数据[12]而不用花费庞大的成本。该 GIS 允许分析和绘制慢性疾病模式,并可以调查有不同的慢性疾病和健康表现的邻里居住以及邻里特征的影响。此外,它允许分析建筑环境对健康和健康表现的影响,评估使用医疗资源的机会。许多此类研究目前正在进行。

参考文献:

[1] Goldberg D W, Wilson J P, Knoblock C A, et al. An effective and efficient approach for manually improving geocoded data [J]. Int J Health Geogr, 2008(7):60-60.

[2] Winkleby M, Sundquist K, Cubbin C. Inequities in CHD incidence and case fatality by neighborhood deprivation[J]. Am J Prev Med, 2007, 32(2):97-106.

[3] Augustin T, Glass T A, James B D, et al. Neighborhood psychosocial hazards and cardiovascular disease: the Baltimore Memory Study[J]. Am J Public Health, 2008, 98(9):1664-1670.

[4] 江　洲,李　琦. 地理编码(Geocoding)的应用研究[J]. 地理与地理信息科学,2003,19(3):22-25.

[5] 兰小机,彭　涛,王　飞. 赣州市地理编码系统及其关键技术[J]. 测绘科学,2009(2):231-232.

[6] 王凌云,李　琦,江　洲. 空间信息融合与地理编码数据库的开发[J]. 计算机工程,2007,30(5):1-2.

[7] 王凌云,李　琦,江　洲. 国内地理编码数据库系统开发与研究[J]. 计算机工程与应用,2004,40(21):167-168.

[8] Kazda M J, Beel E R, Villegas D, et al. Methodological complexities and the use of GIS in conducting a community needs assessment of a large U. S. municipality[J]. J Community Health, 2009, 34:210-215.

[9] 朱建伟,王泽民. 地理编码原理及其本地化解决方案[J]. 北京测绘,2004(2):24-27.

[10] 何　涛,张世禄. 基于 ArcGis 的县级林业资源管理信息系统研究[J]. 计算机技术与发展,2009,19(2):183-186.

[11] 陈细谦,迟忠先,金　妮. 城市地理编码系统应用与研究[J]. 计算机工程,2004,30(23):50-52.

[12] Bonham-Carter G F, Agterberg F P, Wright D F. Integration of geological datasets for gold exploration in Nova Scotia[J]. Photogram Eng Remote Sens, 1988, 54:1585-1592.