数据挖掘在人口GIS中的应用研究

李军利[1],查良松[2]
(1.滁州学院 国土信息工程系,安徽 滁州 239012;
2.安徽师范大学 国土资源与旅游学院GIS重点实验室,安徽 芜湖 241003)

摘 要:针对人口GIS中的海量人口数据,运用数据挖掘技术对人口信息的时空特征进行研究,挖掘人口数据背后的知识与规律。时间上,通过GM(1.1)、三次指数平滑与一元线性回归模型对人口信息变化趋势做定量研究,以近十年安徽全省年末户籍人口数量为例进行预测,发现安徽未来五年内人口总量平稳增长,人口压力较大。空间上,通过空间自相关理论对人口信息的聚集程度进行度量,以1997-2003年安徽区县人口增长率为例进行探测,依据Moran散点图,Moran's I为0.1444,发现人口分布在整体上呈聚集状,具有较强的空间自相关。人口信息变化与经济发展、自然地貌有关,将人口信息诸多社会学特征一并纳入地域空间上考虑,通过数据挖掘能可以分析其时空演变规律。

关键词:人口GIS;数据挖掘;GM(1,1)模型;空间自相关
中图分类号:TP311;P208 **文献标识码**:A **文章编号**:1673-629X(2009)12-0213-04

Application Study of Data Mining Technology in Population GIS

LI Jun-li[1], ZHA Liang-song[2]
(1. Department of Land Information Engineering, Chuzhou University, Chuzhou 239012, China;
2. Key Labortary of GIS, College of Territorial Resources & Tourism,
Anhui Normal University, Wuhu 241003, China)

Abstract: In view of the massive population data in population GIS, conducts an application study on spatial and temporal characteristic of the population information with data mining technology to mine the knowledge and the rule behind population data. In terms of time, the demonstration shows that the year-end household quantity population in Anhui province is about to be steady growth and the pressure will be huge in five years, with carrying on the population prediction by GM(1.1) model, three indices smooth as well as linear regress; In terms of space, then it explores the aggregation level of population information through the spatial autocorrelation theory, for example, taking a study of 1997-2003 years county population growth, analyzing Moran scatter plot of population growth ratio of county city in Anhui province, the research obtains Moran's I is 0.1444, indicates the population distribute in the whole area forms the congregated shape, has the strong spatial autocorrelation. By the above study, population information change is related with economic development and natural landform. Integrating many sociology characteristics of the population information in the union region space, can be possible to analyze its space and time evolution rule through the data mining.

Key words: population GIS; data mining; GM(1.1) model; spatial autocorrelation

0 引 言

人口地理信息系统(简称人口GIS)是MIS与GIS结合的一种新技术,具有强大的空间分析和可视化功能,它最早源自美国人口调查局于1970年人口普查后开发的以街道信息为主的双重独立的地图编码系统GBF/DIME[1],后来美国的TIGER,英国的HMSO、SAPAC系统是其发展[2~4]。日本、加拿大、丹表等国在20世纪90年代左右也相继建立了人口GIS[3~5]。国内人口GIS真正兴起,始于2000年第五次人口普查后,国务院人口普查办公室要求:"有条件、有能力的省市可率先建设人口GIS。"上海、海南、北京、天津、银川、青岛、厦门、南京和广州等地相继建立了人口GIS。随着各地海量人口信息数据库的建立与数据挖掘技术在各个行业中的广泛应用,为在人口GIS中挖掘出可信的、有效的人口信息提供了可能[6]。运用数据挖掘技术,得到隐藏在人口数据背后的知识与规律,可以弥补人口GIS分析功能的不足,为具体应用提供准确

收稿日期:2009-03-24;**修回日期**:2009-06-15
基金项目:国家自然科学基金项目(40771207);安徽省高校自然科研项目(2006kj018B);滁州学院校级科研项目(2008kj007B)
作者简介:李军利(1976-),男,安徽无为人,讲师,研究方向为空间数据挖掘、GIS应用与制图。

的决策依据,尽量减少不确定因素[5]。

1 时间域人口数据挖掘

人口预测是人口 GIS 的重要功能。人口信息时间域分析方法很多,如傅里叶周期分析、自回归模型、分形理论、灰色理论等。而目前在人口 GIS 中用的较多的仍然是马尔萨斯人口模型、离散指数增长模型、阻滞增长模型(Verhust;Logistic)和人口发展方程。基中几何增长法在十年以内的短期预测较好,Logistic 模型适合二十年左右的中长期预测,人口动力学方程主要对长于五十年的预测也能产生较好效果。而 Logistic 模型主要考虑种群类内竞争与最大人口容量,动力学方程要在稳定的社会环境下进行,数据采集较复杂,灰色 GM(1,1)对较短时间预测较准。模型都是有局限性的,使用多种方法进行建模,相对合理。以下借助灰色理论、三次指数平滑与一元回归模型预测,取其三者预测均值对人口信息变化趋势做定量研究。

1.1 模型原理

GM(1,1)是用某一指标(现象或要素)的过去行为来预测未来,预测结果是该指标在未来各个时刻的具体化数值。通过时间序列历史数据揭示现象随时间变化的规律,将这种规律延伸到未来,从而对该现象的未来做出预测,也可以反推过去[7~9]。基于 GM(1,1)的人口灰色预测的实质,是找出某一序列数据间的动态关系,将同一序列各年度人口数据"一视同仁"地当作研究对象,找出各数据间的灰指数关系[10]。而人口数量在不同时期的数值大小,也常受不确定因素的影响,使用指数平滑可以消除偶然因素影响,使其随时间发展变化的趋势和方向明显化。

因篇幅限,这里只给出 GM(1,1)模型按一般计算步骤[9,10]:首先生成数列的累加矩阵 $\boldsymbol{B}$,再建立微分方程为 $\frac{dx^{(1)}}{dt} + ax^{(1)} = b$。$a$ 为发展系数,b 为灰作用量,是微分方程的参数,可以通过如下最小二乘法拟合得到。建立所对应的时间响应函数方程,$x^{(1)}(t+1) = \left[x^{(0)}(1) - \frac{b}{a}\right]e^{-at} + \frac{b}{a}$,此即为 GM(1,1) 预测模型,通过此模型计算出各期的 $x^{(1)}(t+1)$,再反生成时间序列的预测值 $\hat{x}^{(0)}(t) = \hat{x}^{(1)}(t) - \hat{x}^{(1)}(t-1)$。

1.2 实证分析

安徽是人口大省,对 1949~2008 年末全省户籍人口数量作分析,除 1958~1962 年外,人口一直平稳上升,改革开放以来人口成平稳发展态势。考虑到年限太长,波动太大,发展趋势无规律可循,文中选取 1997~2005 年总人口进行预测。使用 GM(1,1)模型,得到下式:

$x(t+1) = 700727.109522e0.008890t + 694521.109522$ (t 从 1999 年起算,$t = 0,1,2,\cdots,n$,灰色预测精度 $p = 1.0000$ 很好,$c = 0.1091$ 很好)

因上式中 p,c 结论很好,否则要进行修正。经过散点模拟,发现随时间推移预测结果偏大,这是因为 GM(1,1) 模型是以灰色模块为基础的,在灰色模块中未来预测值的上界和下界所夹的平面为灰平面,成一喇叭型展开,即未来时刻越远,预测的灰区间越大,精度也就会受到影响。人口系统复杂,随着时间的推移,未来的一些扰动因素将不断地进入系统,发生影响,必然导致越往未来发展,灰度越大,预测值的实际意义越小[11]。因此再选用三次指数平滑与一元回归模型预测,然后取三者的平均值。三次指数平滑方程如下:

$x(t+1) = 6744.8082 + 82.0993t + 3.0106t^2$

($\alpha = 0.40$, 均方误差 $= 25.2951$, $A_0 = 6205.7705$, $B_0 = 1.9585$, $C_0 = 2.0001$)

一元回归模型:

$Y = 57.088x - 107919$(相关系数 $R = 0.9874$,置信水平可靠)

取三种预测方法的平均值,得到未来五年人口合理预测值,单位为万人(见表 1)。

表 1 安徽省未来人口预测

模型	2009 年	2010 年	2011 年	2012 年	2013 年
GM(1.1)	6778.6	6839.2	6900.2	6961.8	7024
三次指数平滑	6829.9	6921	7018.2	7121.4	7230.6
一元线性回归	6770.8	6827.8	6884.9	6942	6999.9
平均值	6793.1	6862.7	6934.4	7008.4	7084.8

三种模型很好地反映了未来五年的人口增长水平,预测结果也非常接近,结果具有可信性,预测表明未来安徽人口增长平稳。因此控制人口增长任务艰巨,人口压力较大。同此原理,可在人口 GIS 中进行诸如人口性别数量、人均纯收入等预测。

2 空间域人口数据挖掘

人口 GIS 的另一个重要功能就是人口统计信息空间化。传统的统计学模型是假定观测点(区)结果互相独立,而实际上在大多数观测结果都不具有独立性。空间统计学有一个基本假设,即地理学第一定理,相邻地理单元存在着某种联系,距离近事物之间的联系性要比距离远的事物之间的联系性更强[12]。当前正在应用并仍处于研究中的空间统计学方法,主要有空间关联分析研究、模式分析、尺度分析、地理分区、地理统计学、分类学、空间采样以及空间经济学等,而空间自相关理论作为空间域中的值聚集程度的一种度量方

法,其研究目前非常活跃[13]。

2.1 空间自相关理论

空间自相关理论反映的是一个区域单元上的某种地理现象或某一属性值,与邻近区域单元上同一现象或属性值的相关程度,是一种检测与量化从多个标点中,取样值变异的空间信赖性的空间统计方法[1,14]。它定义一个特殊的矩阵 γ,该矩阵由一个描述所有测量点(区)位置相关可能性的空间权重矩阵,与一个非空间相关性的矩阵(如经济关系、社会关系或者其它关系)相乘得到,如果这些矩阵信息是相似的,那么 γ 是高度正相关的,表示空间现象聚集性的存在[1]。

空间自相关按功能可分为全域自相关(Moran)与局域自相关(LISA)[5,14],前者表示某现象的整体分布状况,但不能确切表示哪些区域聚集,依据 Anselin 提出的 LISA 方法,局域自相关能够推算聚集范围,分别用 Moran's I 与 Local Moran's I 指数表示。Moran's I 值介于 -1 到 1 之间,大于 0 的为正相关,小于 0 的为负相关,趋于 0 时表示此空间呈随机分布,n 个局域自相关的 Moran's I 值之和为全域自相关 Moran's I 值。

Moran's I(I) 与 Local Moran's I(I_l) 按下式计算[14]:

$$I = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ji}} \times \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$I_l = \sum_{j}^{n} W_{ij}(x_i - \bar{x})(x_j - \bar{x})$$

其中,W_{ij} 是研究范围内每一个空间单元 i 与 j($j = \{1, 2, 3, \cdots, n\}$) 区空间单元的空间相邻权重矩阵。以 1 当作 i 与 j 相邻时,而以 0 表示 i 与 j 不相邻,x_i、x_j 表示统计单元信息属性值,$\bar{x}$ 表示统计属性均值。

2.2 实证分析

利用安徽省 1997~2000 年人口统计资料,以县区为单位,共 78 个研究单元(各市辖区表示合并为一个区,)用 ARCGIS 等软件提取图层,计算 1997~2003 年研究单元人口增长率。采用邻近标准,使用 ArcInfo 软件的 AAT 属性表中左右多边形共孤段原理,建立空间权重矩阵。经处理得到 Moran's I 的值(见图 1)为 0.1444(为图 1 中长回归斜率)>0,在国际通用标准 p <0.05 显著水平下检验通过。表明安徽区县人口在整体上分布呈聚集状,具有较强的空间自相关,不是呈随机状。

依据散点图绘制"高高区(HH)、低高区(LH)、高低区(HL)与低低区(LL)",分别表示第一、二、三与四象限。其中高高、低低表示区域人口增长率与周围地区高度的正相关,有较强的集聚与相似性。高低表示本区域人口增长率高,周围人口增长率低,低高与高低意思相反。表 2 列出了各县区的"高低"现象分布。计算 LISA CLUSTER 表明,皖南、皖西南部分地区低增长率聚集区显著,淮北高增长率聚集区显著。

表 2　Moran 散点关联类型

本身高周围高 (HH)	自身低周围高 (LH)	自身高周围低 (HL)	自身低周围低 (LL)
砀山、萧县、淮北、宿州、亳州、灵璧、濉溪、涡阳、泗县、太和、界首、固镇、蒙城、利辛、五河、怀远、阜南、阜阳、利辛、阜阳、临泉、凤阳、凤台、定远、寿县、霍邱	来安、全椒、肥东、肥西、舒城、铜陵、当涂	天长、滁州、合肥、巢湖、马鞍山、芜湖市、马鞍山、铜陵市、安庆	除其它象限列出的县区

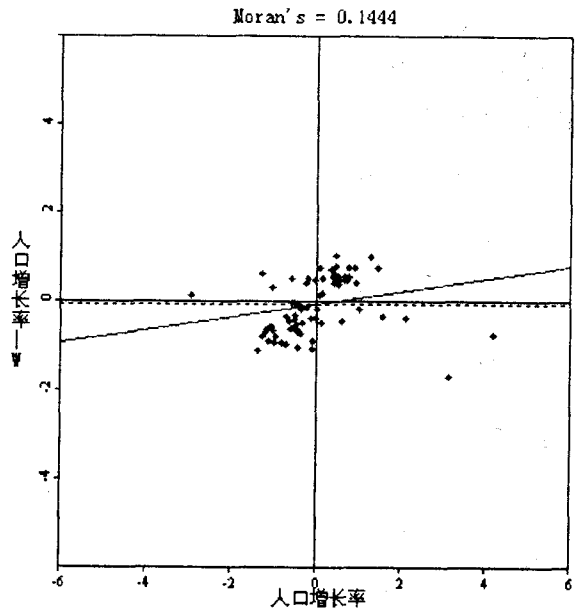

图 1　安徽县市人口增长率的 Moran 散点

人口增长是自然、历史、经济与社会综合作用的结果。将社会经济指标、安徽地貌与空间自相关分析结果综合分析,如图 2,研究表明地貌与经济发展水平是影响安徽省人口增长率的重要因子。分析得出整个聚集特征明显,安徽省县区人口增长率的高低分布,受经济发展水平与地貌影响明显,经济发展水平高与地势平坦的地区在安徽各地中最高,城市化、较适宜居住使得迁移性人口增长幅度高于其它地区,为最高人口增长率地区,与周围地区集聚偏小。皖南地区与皖西大别山区因地貌原因,人口稀少,多中山、低山、高丘,人口垂直分布变化大,低增长率区空间集聚。而淮北地带,多倾斜平原,出现高增长集聚,沿江平原多沉积盆地,人口稠密,沿主要江河(多沙洲、平原)与交通干线(经济发展水平较高)人口增长率也相对较高,城市面积大的地区人口增长率较高,部分中心城市国家边缘县区为低增长区,主要原因是土地被征用,农业人口转化为城市人口,如铜陵市与铜陵县。运用空间自相关

理论还可分析人均收入、受教育程度等其它社会学特性的空间分布特征。

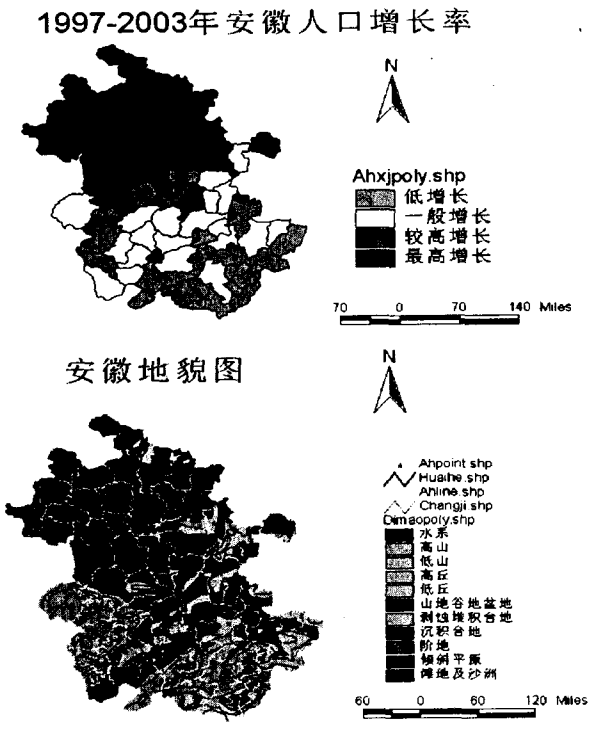

图 2 安徽人口增长率与地貌组合图

3 结束语

运用 GM(1.1)模型对时态人口信息进行预测，具有要求样本数据较少，短期预测精度高的特点，但不太适合长期预测，将三次指数平滑与一元线性回归模型加入，可消除单一模型的局限性；运用空间自相关理论，可以发现人口增长率在空间上的聚集与相似性，发现非直观的规律与异常数据。文中仅使用了 1997～2003 年两个时间点的横截面数据，如果做多年份空间自相关分析，还可以发现研究要素的变化情况，做空间趋势分析。综合使用人口 GIS 中的时空集成信息，可以分析人口信息的时空演变机制和规律。通过时空域数据挖掘技术研究海量人口信息中隐藏的知识与特征，提高人口 GIS 的应用效率，可为科学的人口管理与决策提供更好的服务。

参考文献：

[1] Longley P A, Goodchild M F. 地理信息系统(上卷)：原理与技术[M]. 第 2 版. 唐中实等译. 北京：电子工业出版社，2004.

[2] Martin D. Geography for the 2001 Census in England and Wales[R]. United Kingdom: Department of Geography, University of Southampton, 2002.

[3] 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[J]. 武汉大学学报：信息科学版，2001(3)：226－231.

[4] Broome F R, Meixler D B. The TIGER database structur[M] // In Marx R W. The census Burean's TIGER system. Bthesda: ACSM, 1990: 39－47.

[5] 李成名，印 洁，王继周，等. 人口地理信息系统[M]. 北京：科学出版社，2005：11－12；122－147.

[6] 吴春阳，何友全. 数据挖掘技术及其在旅游线路规划系统的应用[J]. 计算机技术与发展，2008，18(9)：234－237.

[7] 安鸿志. 时间序列的分析与应用[M]. 北京：科学出版社，1983.

[9] 徐建华. 现代地理学中的数学方法[M]. 北京：高等教育出版社，2002：338－343.

[10] 邓聚龙. 多维灰色规划[M]. 武汉：华中理工大学出版社，1989：20－22.

[11] 易德生，郭 萍. 灰色理论与方法－提要、题解、程序、应用[M]. 北京：石油工业出版社，1992：227－230.

[12] Tobler W R. Cellular geography[M] // In Gale S, Olsson G. Philosopy in geography. Dordrect, Reidel: [s. n.], 1979: 379－386.

[13] 刘湘南，黄 方，王 平，等. GIS 空间分析与方法[M]. 北京：科学出版社，2005：189－194.

[14] Anselin L. Local indicators of spatial association－lisa[J]. Geographical Analysis, 1995, 27: 115－116.

(上接第 212 页)

[2] Mack S. 流媒体宝典[M]. 邢栩嘉，王 佟，赵 峪，等译. 北京：电子工业出版社，2003.

[3] Microsoft Corporation and RealNetworks, Inc. Advanced Streaming Format (ASF) Specification[S]. Public Specification Version 1.0. [s.l.]: [s.n.], 1998.

[4] 何晓鹏. ASF 媒体格式基于 RTSP 协议流化与流媒体服务器中 VCR 功能的实现的研究[D]. 北京：北京邮电大学，2006.

[5] 陶洪久，柳 键，田金文. Windows Media 的流媒体格式 ASF 的分析[J]. 交通与计算机，2001(6)：52－55.

[6] 赵志敏. ASF 流媒体课件编辑技术的研究与实现[D]. 重庆：重庆大学，2007.

[7] 杨 征，王 晖，吴玲达. 高级流格式 ASF 的剖析与应用研究[J]. 计算机应用研究，2001(6)：65－68.

[8] 沈秀红. 基于 Web 的流媒体同步多媒体课件的制作与应用[J]. 广东技术师范学院学报，2007(3)：65－67.

[9] 毕 野. 基于 Windows Media Encoder 实现流媒体同步控制[J]. 淮海工学院学报，2008，17(4)：24－27.