

人口区划及其在人口空间化中的 GIS 分析应用

——以山东省为例

黄耀欢^{1,2}, 杨小唤¹, 刘业森¹

(1 中国科学院地理科学与资源研究所, 北京 100101; 2 中国科学院研究生院, 北京 100049)

摘要: 对基于土地利用/土地覆盖类型的人口空间化方法, 进行人口区划的 GIS 分析应用是一有效技术途径。本文在全国一级区划基础上, 通过数理统计分析提取二级区划各影响因子, 建立二级区划指标, 进而利用空间分析技术进行人口空间化的二级分区。同时以农村居民地、城镇居民地和耕地数据建立人口空间化模型, 在 GIS 的支持下实现山东省 2000 年人口统计数据空间化。并通过对比基于一级人口区划和二级人口区划的人口空间化数据, 选取郯城县进行结果精度检验。验证结果表明二级人口区划后的人口空间化数据在数值精度和空间精度上都有明显提高。
关键词: 人口空间化; 二级人口区划; 土地利用; 误差检验

1 引言

人口统计数据与所属区域的空间特征脱节, 将严重制约着各种信息, 如环境数据、自然资源数据、经济统计数据等的集成。解决这种脱节的一种方法就是进行人口空间化。人口统计数据空间化, 尤其是栅格化的人口数据对跨学科的研究与应用具有重要价值^[1]。

国内外对人口空间化已经有过研究, 国外如 1990 年美国人口普查后, 研制了 TIGER, 即为一种人口地理信息系统 (Topologically Integrated Geographic Encoding Referencing)。我国岳天祥等根据 NNP、海拔高度、城市分布和交通设施状况与人口分布相关关系, 采用格网构建人口分布表面模型 (SMPD), 模拟了我国自 1930~2000 年主要时段的人口分布状况, 并以此为基础分析了中国人口的分布规律和趋势^[2]。又如杨小唤等人在人口空间分布区划的基础上, 利用 TM 图像获取的 1 10 万的土地利用/覆盖数据, 建立了统计人口数据与土地利用类型的多元相关关系模型, 计算各种土地利用类型的人口系数, 在 GIS 支持下计算出全国 1km 格网人口空间分布, 并且用 DEM 等数据对结果进行了修正, 建立了全国分年度公里格网的人口空间化数据库

^[4], 建立了全国范围内公里格网人口数据^[5]。

本文通过分析国内外已有的人口空间化过程, 发现在不同尺度范围内, 人口空间分布密度均一度不同。本文将在小尺度范围内进行二级区划, 并以县(市)单元来探讨减少人口空间化误差的方法。

2 人口区划方法与分析方案

本文以山东省为研究区域, 重点以临沂市郯城县作为研究成果检验区域。郯城县位于山东省南部, 下辖 17 个乡镇, 属于冲积平原。分析所使用的数据: (1)2000 年的 1 10 万土地利用图及 20 世纪 90 年代末的 TM 影像; (2)“2002 年的中华人民共和国全国分县(市)人口统计资料”(中华人民共和国公安部编, 群众出版社, 2003 年北京); (3)“中国乡镇统计资料-2003”(国家统计局农村社会经济调查总队编, 中国统计出版社, 2004 年 2 月第 1 版); (4)1 5 万地形图; (5)1 40 万山东各县市政府网站和中国行政区划网(乡镇边界数据)。

2.1 二级人口区划影响的因子分析

在一级区划的基础上, 基于土地利用类型进行全国人口空间化建模, 基本体现了“区间差异明显、

收稿日期: 2006-03-19; 修回日期: 2006-07-03.

资助项目: 本研究得到国家自然科学基金项目(40471112)和国家科技基础条件平台 2004DKA20180-07 课题资助。

作者简介: 黄耀欢(1982-), 男, 安徽黄山人, 硕士生。主要研究方向是遥感与 GIS 应用。Email: huangyh@reis.ac.cn

区内人口分布特征相似”的特点,但由于人口空间分布的复杂性,在一级区划内部,同一土地利用类型内部的人口密度是有差异的,这种差异在中小尺度上就会对精度造成较大的影响。为了体现人口空间分布的这种异质性,本文将采用在一级分区的基础上进行二级人口区划后建模,以期减少中小尺度上的人口空间数据误差。

影响人口空间分布异质性的因子是多方面的,包括各种自然因素和人文因素。土地利用的空间格局是影响人口空间分布的主要因素^[9],而且本文中

的人口空间数据是通过建立人口密度与各土地利用类型所占面积之间的关系获得的,因此在进行二级区划的过程中,土地利用/土地覆盖因子必不可少。另外人口分布与地形关系密切^[7],因此,研究中选取了平均海拔和平均坡度代表地形地貌信息。在县(市)尺度,选取山东省 108 个县市的各指标:平均海拔、平均坡度、居民地平均大小、居民地面积百分比(居民地%)、耕地面积百分比(耕地%)、林地面积百分比(林地%)、草地面积百分比(草地%)。对 7 个变量进行相关分析和主成分分析(表 1、表 2、表 3)。

表 1 各指标相关系数
Tab.1 The correlation of indices

	平均海拔	平均坡度	居民地平均大小	居民地%	耕地%	林地%	草地%
平均海拔	1.0000						
平均坡度	0.6806	1.0000					
居民地平均大小	-0.4968	-0.4263	1.0000				
居民地%	-0.2812	-0.1130	0.2259	1.0000			
耕地%	-0.2985	-0.1500	0.1612	0.8642	1.0000		
林地%	0.3381	0.5938	-0.1483	0.6071	0.5641	1.0000	
草地%	0.2789	0.4052	0.0163	0.3010	0.2973	0.6339	1.0000

表 2 主成分特征值及其贡献率
Tab.2 The eigenvalue of principal components and its contribution

主成分	特征值	贡献率/%	累积贡献率/%
1	2.773	39.614	39.614
2	2.492	35.599	75.213
3	0.807	11.523	86.736
4	0.415	5.936	92.672
5	0.305	4.353	97.025
6	0.129	1.842	98.867
7	0.079	1.133	100.000

从表 2 中可以看到,主成分 1 和主成分 2 方差累积贡献率达到 75%以上,基本已经能够代表大部分的信息,根据分析需要把其余主成分作为噪声来源处理。提取第 1 主成分和第 2 主成分进行分析(表 3)。分析表 3 可以选取在第一主成分和第二主成分中影响较大因子,同时参考相关系数表 1,考虑不同因子之间的相关性。本文作者选取平均海拔、居民

地(%)、林地(%)三个变量作为二级区划的主要参考因素。

表 3 各变量在主成分中的系数
Tab.3 The indices of variables in principal components

	1	2
平均海拔	0.377	-0.804
平均坡度	0.590	-0.685
居民地平均大小	-0.186	0.643
居民地%	0.631	0.694
耕地%	0.608	0.693
林地%	0.961	0.004
草地%	0.745	-0.050

2.2 基于空间分析的人口区划方案

(1) 区划指标计算及其标准化

对山东省每个县的三个指标按公式 1 分别进行标准化处理,得到 X_i

$$X_i = \frac{a_i - \min(a_i)}{\max(a_i) - \min(a_i)}, \quad i=1, 2, 3 \quad (1)$$

a_1 为平均海拔, a_2 为居民地%, a_3 为林地%
各县区划指数按公式 2 计算, 得到各县的综合
区划指数。

$$T_P = X_1 \times X_2 \times X_3 \quad (2)$$

将人口分布特征重心坐标(x, y)和省级人口分
布特征指数(T_P)根据归一化公式(3)进行处理

$$I_i = \frac{m_i - \min(m_i)}{\max(m_i) - \min(m_i)}, \quad i=1, 2, 3 \quad (3)$$

上式中 m_i 分别代表 x, y, T_P 从而分别得到 I_1 、
 I_2 、 I_3 。

(2) 空间聚类分析

将 I_1 、 I_2 、 I_3 作为变量的三个属性, 在 SPSS 软
件中使用 K 均值聚类方法进行聚类分析。山东省共
108 个县(市), 按照 5 类的分区方案进行区划, 结果
如图 1。

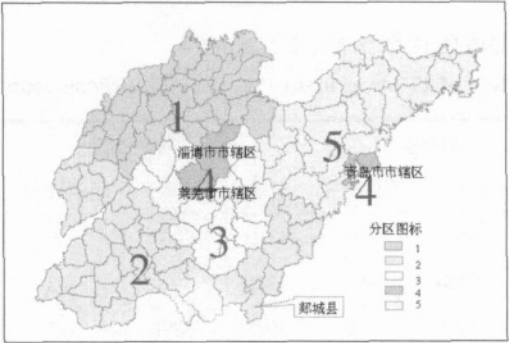


图 1 5 个二级人口区划图
Fig.1 The map of five classes by class-2
population regionalization



图 2 4 个二级人口区划图
Fig.2 The map of four classes by class-2
population regionalization

分析图 1 发现四区仅有 3 个县, 在模型建立中
基本无意义, 根据地理学第一定理, 将此 3 县(市)划
分到相邻的两个区中, 其中将莱芜市市辖区和淄博
市市辖区划分到 3 区中, 将青岛市市辖区与原五区
合并为 4 区, 即为图 2。将图 2 作为山东省人口区划
二级分区最后的结果, 即将山东省划分为四个二级
分区。

3 人口空间化分析计算与应用

在对山东省进行二级人口分区后, 应用 ArcGIS
软件和统计软件 SPSS 进行人口空间化计算。

3.1 人口空间化分析计算

在 ArcGIS 中从 1 10 万土地利用矢量图中提取
水田(代码 11)、旱地(代码 12)、城镇居民地(代码
51)、农村居民地(代码 52)得到矢量土地利用图。

用 2000 年山东省县界图(sclbnd)与土地利用
图进行相交(intersect)操作, 获得 108 个县内土地
利用类型图(xjlanduse), 并应用 SPSS 进行统计计
算获得各县水田(11)、旱地(12)、城镇居民地(51)、
农村居民地(52)的面积。胡焕庸(1936)曾指出, 获得
高精度人口密度图需 2 个必要条件——小范围统
计人口, 居民地数据^[9]。因此本文也认为人口主要分
布在居民地。然而中国耕地面积较大, 在耕地中存
在有少量居民地, 但是由于居民地面积太小有可能
被忽略, 因此将水田(11)与旱地(12)合并成为耕地
面积。并且在模型中考虑耕地(1)面积, 减少因为居
民地丢失造成的误差。

在每一个二级区划内, 将区划中的各县(市)建
立方程组公式(4):

$$P_i = \sum_{j=1}^m D_j \times A_{ij} + b_i \quad i=1, 2, \dots, n \quad (4)$$

P_i 为第 i 县(市)的人口, D_j 为第 j 类居民地
的人口居住密度, A_{ij} 为第 i 县第 j 类居民地的面积, b
为方程截距, n 为研究区中的县(市)数目。山东省共
108 个县, 故 $n=108$ 。假设人口全部居住在三种土地
利用类型中, 因而在分析时将 b 设置为 0。

在 SPSS 中分别对 4 个分区中县(市)3 个指标:
农村居民地、城镇居民地和耕地进行回归。获得不
同分区的不同土地利用类型的人口密度

$$D_{ij} \quad i=1, 2, 3, 4 \quad j=1, 2, 3 \quad (5)$$

式中 i 为分区, j 为土地利用类型, 如 D_{12} 表示分区一城镇居民地人口密度。

3.2 人口空间化制图

在 ArcGIS 中根据获得 4 个分区的 D_{ij} 与土地利用图进行计算, 获得各个土地利用类型的人口数, 并进行栅格化, 生成山东省 2000 年栅格人口分布图(图 4)。同时用相同的方法进行一级区划的人口空间化(图 3)。

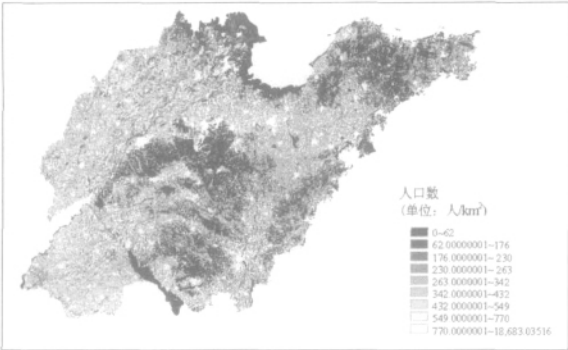


图 3 基于一级区划的人口空间分布
Fig.3 Population distribution based on class-1 regionalization

4 人口空间化结果分析

对上一节中得出两图进行对比, 可以发现有一定的差别。为了验证结果, 选取山东省南部郯城县进行误差检验。郯城县位于山东省二级分区中的第 2 区并隶属于临沂市。临沂市地形较复杂, 这种复杂地形区域的误差分析能充分检验人口空间化方法和结果。

同时我们考虑到, 由于县内的乡界对土地利用图的分割作用, 如直接从空间化图中提取各乡镇人口, 会造成一定的误差^[9]: 因此应用 ArcGIS 对 1 10 万土地利用矢量图进行 intersect 处理, 以乡为单位进行人口计算。

(1) 人口空间化后, 进行误差数值检验。误差计算公式如下:

$$EP = \frac{(popj - popt)}{popt} \times 100\% \quad (6)$$

其中 EP 为人口误差, $popj$ 为按照模型估算人口数, $popt$ 为统计人口数。在 SPSS 中对郯城县 17 个乡镇进行误差计算, 得表(4)。

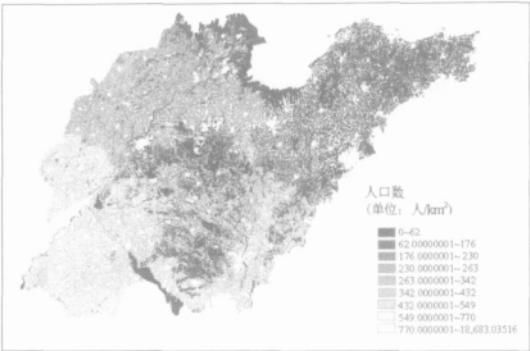


图 4 基于二级区划的人口空间分布
Fig.4 Population distribution based on class-2 regionalization

表 4 郯城县 17 个乡镇人口误差情况

Tab.4 The error of population of 17 towns and countries in Tancheng county

乡镇名	统计值	一级区划			二级区划		
		计算值	误差	百分比(%)	计算值	误差	百分比(%)
李庄	49664	44567	-5097	-10	47196	-2468	-5
黄山	46965	22511	-24454	-52	27023	-19942	-42
沙墩	46964	33694	-13270	-28	39973	-6991	-15
褚墩	55078	34652	-20426	-37	43071	-12007	-22
泉源	47389	43275	-4114	-9	51657	4268	9
庙山	43933	47754	3821	9	51058	7125	16
胜利	45645	26676	-18969	-42	32622	-13023	-29
郯城	186051	200405	14354	8	180116	-5935	-3
马头	78785	64440	-14345	-18	66668	-12117	-15
重坊	53543	36261	-17282	-32	38161	-15382	-29
港上	39835	30807	-9028	-23	30998	-8837	-22
新村	35176	18735	-16441	-47	22670	-12506	-36
高峰头	47396	32539	-14857	-31	40699	-6697	-14
归昌	36798	30877	-5921	-16	38094	1296	4
花园	48083	33725	-14358	-30	42767	-5316	-11
红花	61590	51619	-9971	-16	65270	3680	6
杨集	49373	50820	1447	3	56601	7228	15

分析表 4, 可以得到二级分区后的人口误差明显小于一级分区后得到人口空间化结果, 误差数值分布相对也比较随机。具体定量分析统计图如下。

表 5 乡镇尺度精度检验结果
Tab.5 The result of precision checked-out on county scale

误差(%)	分区	乡镇个数	
		一级分区	二级分区
误差范围 (绝对值)	0~9	4	5
	10~19	4	6
	20~29	2	4
	≥30	7	2
误差平均值 (%)		-21.824	-11.353
误差绝对值平均值 (%)		24.176	17.235

一级分区误差 30% 的乡镇占 41.2%, 而二级分区后的误差 30% 的乡镇仅占 11.8%。误差平均值也由 -21.824% 降低至 -11.353%, 降低了约 48%。误差绝对值平均值也由 24.176% 下降到 17.235%, 精度提高了约 30%。由此可见, 在数值精度上, 二级分区后的空间化精度明显高于一级分区模型下的人口空间化(表 5)。

(2) 由于人口空间化是一个将统计数据分布到其相应的地理空间上的过程, 所以空间化后的空间误差在误差检验中也是不可忽视的。因此把郯城县在一级分区和二级分区后的人口空间化误差作图表示如下(图 5、图 6)。

由图 5 可直观地发现二次区划后乡镇尺度上的人口数值大误差明显减小。

在误差的空间分布上, 也有显著提高。图 5 中误差 < -30% 的 7 个乡镇基本位于郯城县西部, 图 6 的误差较一级分区后结果, 不仅在数值上分布随机, 而且在空间上也属于随机分布, 没有图 5 中的误差相近区域团簇现象出现。

对郯城县 17 个乡镇分析, 得出二级分区在数值和空间两方面都能够减少误差。在数值方面, 误差数值分布随机, 且精度有了很大提高。而在空间方面误差的分布更加随机, 很大程度消除了由于地形等因素引起的系统误差, 基本消除了误差团簇现象。

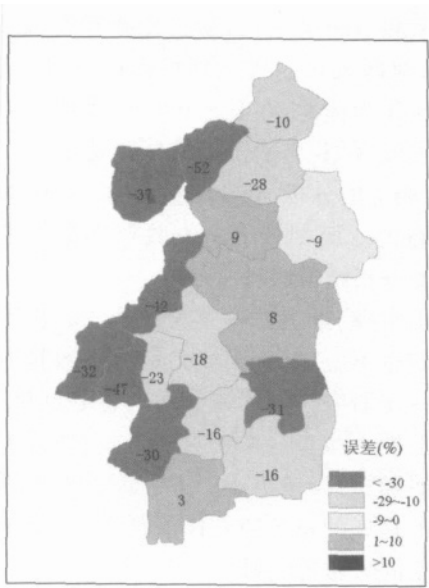


图 5 一级分区误差空间分布
Fig.5 The spatial distribution of error based class-1 regionalization

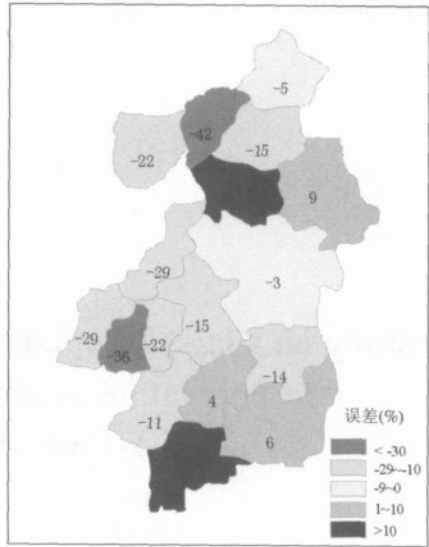


图 6 二级分区误差空间分布
Fig.6 The spatial distribution of error based class-2 regionalization

5 结语

本文通过主成分分析获取人口二级区划的主要指标, 并通过 ArcGIS 软件进行指标提取。根据各指标进行多元回归分析, 最后在 ArcGIS 中获取一级区划及二级区划的人口空间化栅格图。研究表明二级区划对中尺度人口空间化具有明显效果, 能极

大减小数值和空间误差。二级区划指标提取过程中发现, 山东地区人口二级区划可以以海拔、坡度和耕地面积比作为标准, 在其他区域可以地形为主要因子进行二级区划, 当然不同地区可能主导因子不同, 应进行相应的分析, 获取划分指标。在一级区划基础上进行的二级划分的聚类数在本文中主要依据多次聚类分析结果获得。

分析山东省, 得出以 4 类为最佳。然而不同分类数的确定也不应相同, 因此二级分区数量的确定仍有待进一步研究。以郯城县作为检验区域发现, 二级区划后能够减少一般的数值误差, 使误差在数值分布上更随机, 误差更小。并且能够减少因为地形等因素所造成的系统空间团簇误差, 使空间化后人口误差空间分布更随机。

综上所述, 在人口空间化过程中进行二级区划能够明显地提高人口空间化的数值精度以及空间精度。但是在模型的建立过程中, 今后仍然有一些问题需要深入研究。

参考文献

- [1] 刘业森硕士毕业论文. 人口空间数据误差分析研究. 北京, 中国科学院地理科学与资源研究所, 2005.
- [2] Deichmann U we. A review of spatial population database design and modeling. NCGIA Technical Report 96-3, 1996.
- [3] T X Yue, Y A Wang, S P Chen et al. Numerical simulation of population distribution in China. *Population and Environment*, 2003, (2): 141~163.
- [4] 杨小唤, 江 东, 王乃斌, 熊利亚, 刘红辉. 给予空间分析方法的人口空间分布区划. *地理学报*, 2002, 57 (增刊): 76~81.
- [5] 江 东, 杨小唤, 王乃斌, 刘红辉. 基于 RS、GIS 的人口空间分布研究. *地球科学进展*, 2002, 5 (17): 734~738.
- [6] 高志强, 刘纪远, 庄大方. 基于遥感和 GIS 的中国土地资源生态环境质量同人口分布的关系研究. *遥感学报*, 1999, 3(1): 66~70.
- [7] 叶文振. 江西人口的地形区域分布: 1953~1993. *南方人口*, 1999, (1): 47~52.
- [8] 胡焕庸. 句容县人口之分布. *地理学报*, 1936, (3): 621~627.
- [9] 杨小唤, 江 东, 王乃斌, 刘红辉. 人口数据空间化的处理方法. *地理学报*, 2002, (57)增刊.

A Study on Class-2 Population Regionalization and Its Application to Population Spatial Distribution Based on GIS: A Case of Shandong Province

HUANG Yaohuan^{1,2}, YANG Xiaohuan¹, LIU Yesen¹

(1 Institute of Geographic Sciences and Natural Resources Research, CAS, Beijing 100101, China;

2 Graduate School of the Chinese Academy of Sciences, Beijing 100049, China)

Abstract: The ways of regionalization affect the precision of population distribution based land-use and land-cover directly. In this paper, the authors made use of math-statistics to choose the factors to design the index of class-2 regionalization, and then did class-2 population regionalization using the technique of spatial analysis. Based on it, population distribution was modeled according to data of the rural residential area, city residential area and cultivated land area, and then distributed the statistic population of Shandong province in 2000 in aid of GIS. Eventually, the precisions of the population distribution data were compared between class-1 and class-2 regionalization in Tancheng county. The result of comparison indicated that the precision of population distribution based on class-2 regionalization was improved both in numerical and spatial aspects.

Key words: population spatial distribution; class-2 population regionalization; land-use; error check-out