基于多重分类平滑法的人口数据格网化研究

王汶,付文静,鲁旭　(中国人民大学环境学院,北京 100872)

摘要　人口数据格网化是解决传统统计数据(以行政区划为基本统计单元)在数据精度、数据标准化和数据融合等方面不足的一种有力工具。综述了各种格网化方法和理论,在此基础上,提出了多重分类平滑法,在不引入过多变量的基础上提高了模拟精度。最后,以河南省人口数据为例,采用指数平滑法,得到了连续人口密度分布图。

关键词　地理信息系统;人口数据;格网化;多重分类平滑法
中图分类号　S92　　**文献标识码**　A　　**文章编号**　0517 - 6611(2009)25 - 12327 - 03

Study on Population Data Grid Transformation Based on Multiple Classification Smoothing Method

WANG Wen et al　(The School of Environment & Natural Resources, Renmin University of China, Beijing 100872)

Abstract　Population data grid transformation is a powerful tool to solve the deficiency of data accuracy, standardization and integration in the traditional statistical data (by administrative division as the basic statistical unit). Various methods and theories of grid transformation were summarized. On this basis, the multiple classification smoothing method was put forward, which increased the accuracy of a simulation without introducing too many variables. Finally, using the population data of Henan province as an example, a continuous distribution of population density was obtained by exponential smoothing method.

Key words　GIS; Population data; Grid transformation; Multiple classification smoothing method

随着遥感技术和地球空间技术的发展,人们对数据的空间性和连续性提出了更高的要求。这一方面促进了数据研究由定性化向定量化转变;另一方面也使得数值层面上的模型不能满足数据处理的需求,模型的发展必将过渡到空间层面。近年来不断进行的人口普查工作使得人口数据资料较易获取,但是,传统的人口数据是按行政单元来统计,按矢量型数据格式进行存储和管理。这种用失量数据格式表示的人口空间信息至少会存在以下 4 个问题:①基于行政单元的人口空间数据在行政单元内是均匀分布的,难以体现人口数据具体的空间分布特征;②行政单元边缘的数据跨度太大,数据间缺乏应有的连续性;③实际应用中的功能区划通常与行政分区的界线不一致,因此会增加研究中数据分析的难度和工作量,降低数据的实用性;④行政单元形状、大小没有统一的划分标准,不利于多源数据的融合和数学模型的建立,因此不便于数据的综合研究。因此,要实现人口信息的有效管理和利用,首要问题便是要解决人口分布的空间化问题,即人口数据格网化问题。人口数据格网化,即将以行政区为单元的人口统计数据按一定的数学模型展布到一定尺寸的格网中,实现统计区划由行政区划向格网的转换[1]。在格网数据中,格网是统计单元内部的细胞。通过格网化的人口数据梯度,能方便地表达出区域单元的人口空间分布与分异规律;能匹配、融合多源数据,实现人口数据空间模型的构建和表达;同时,还能在时间上形成以格网为基础的数据序列,以便人口变化动态规律分析[2]。

因此,将以行政区划为统计单元的人口数据格网化,将是今后存储和管理人口空间分布信息的主要手段之一。如何有效使用这一手段,也将成为人口数据地理空间格网化领域的一个研究热点。

基金项目　国家高技术研究发展计划(863 计划)"农村抽样调查空间化样本抽选和管理系统(2006AA120103)"和"大气复合污染的区域调控和决策支持技术(2006AA06A307)"。
作者简介　王汶(1966 -),男,天津人,博士,副教授,从事地理空间信息应用研究。
收稿日期　2009-05-04

1　国内外研究概述

1.1　国外研究进展　1929 年,芬兰地理学家 Graneau 利用 1 km × 1 km 分辨率格网分析自然与社会现象,开创了以各级尺寸格网为基本研究单元进行地理分析的先河。目前,该分析方法已逐渐发展成为一种地学分析方法[3]。经济、社会和环境的研究中,常常涉及到面积内插,这主要是由于研究区域上相应的数据往往不能直接得到,只能由已知区域的数据求得,即统计数据需要空间转换。根据行政单元统计数据生成不同尺寸格网数据实质上是一种特殊的空间内插法[4]。继 1973 年 Markoff 和 Shapiro 系统地论述了面积内插法的标准方法之后,Goodchild 和 Lam 又于 1980 年对该方法进行了深入的研究[5-6]。1992 年,Flowerdew 和 Green、1993 年 Goodchild、Anselin 和 Deichmann 等人相继提出面积内插法、面积权重法[7-8],即通过行政单元在格网中的面积和行政单元平均人口密度获得每个格网内的人口数据。然而,通过该方法获得的人口格网化空间数据在实际应用上有一个缺点,即仅是数据格式的转换,未将人口空间分布的影响因素考虑在内。1993 年,Langfore 和 Unwin 利用卫星影像识别英国爱尔兰东部郡居住区和非居住区像素,并将统计后的人口空间数据分配到每个居住区格网单元中,实现了该模型在合理性方面的一大跨越[9]。

随着人们对高分辨率人口密度数据的需求日益扩大以及地理空间信息技术的引入,更多的栅格数字模拟技术得到了开发和利用。1936 年,Wright 使用美国地质勘探局地形图估计了不同居民点类型的人口密度,首次将分区密度制图技术引入了人口数据的空间连续分布研究中[10],这也成为人口数据地理空间格网化方法的主要理论依据。美国在 1980 年的人口普查中采用自动化技术发展了一套地理基础文件/独立坐标地图编码系统,但存在数据不能匹配、对比,错误难以检测等问题,故又在 1990 年人口普查后研制了人口地理信息系统(Topological Integrated Geographical Encoding Reference System,简称 TIGER),这套系统将人口普查与具体的地理空间位置结合起来,为空间数据的融合和分析提供了平台[11]。随着遥感与地理信息系统技术的发展,土地利用/覆

盖数据逐渐成为人口数据格网化模型的主要建模指标。美国社会经济应用中心利用 1990 年度 NOAA AVHRR 数据提取出土地覆盖空间特征，建立了人口密度与土地利用空间格局之间的关系，从而得到了 1990 年美国的 1 km×1 km 分辨率格网数据库。2001 年，Eicher 和 Brewer 运用 3 级土地利用分级类型，完成县级人口数据格网化[12]。2003 年，Jeremy 综合 White，Eicher，Brewer 等的方法，将城市发展水平分成高、中、低 3 级，利用 TM 数据，通过城市发展程度与人口密度之间的关系，建立了 100 m 人口格网化数据模型[4]。此外，在一些国家（如日本、芬兰、挪威和瑞士等）已建立格网统计数据相关标准。日渐繁多的模拟方法和合理完善的标准体系的建立，将会促进格网统计数据的进一步完善。

1.2 国内研究进展 尽管国内在这一领域的研究远落后于国外，但近 20 年来，许多国内学者在人口数据与地理空间的结合方面做了大量的尝试。例如，1990 年国内第 4 次人口普查结束后，中国人口情报研究中心用 POPMAP 软件直观地显示了全国 30 个省、自治区、直辖市的人口分布、出生、死亡以及社会经济水平等方面的信息[11]。

国内格网化技术研究比较早，但有关人口统计数据格网化模型的研究近几年才出现。目前，国内的人口数据格网化模型主要有两类：一类是应用面积内插与统计分析原理——即数据的空间转换——进行人口数据格网化。2002 年，吕安民、李成名等人在简单面积内插法（面积权重法）的基础上，提出了采用人口密度递归算法模拟人口空间分布的面积内插法，并将统计分析中的核心估计法作为一种比传统面积内插法更优越的模拟方法引入格网化研究[13]；2004 年，范一大、史培军等以 ARC/INFO 为平台，采用面积权重法，通过叠置分析对 1997 年中国北方 13 省的人口数据进行了格网化计算[14]。然而，以上这些模型都仅是单纯地探讨了数据格网化的方法和其技术的实现，虽然操作简单，易与 GIS 结合，但却未考虑影响人口空间分布的各种自然和人文因素。另一类模型则是通过建立人口空间分布影响因素（包括自然因素和人文因素）与人口数据之间的函数，进行人口数据格网化。2003 年，金君、李成名等克服已有模型的不足，将居住区建筑物结构与分布对城镇人口疏密程度的影响纳入考虑范围，建立适合城镇人口分布的格网化人口模型（DPM），这一建模方法增强了模型的实用性和人口空间数据的连续性[11]；2004 年，田永中、陈述彭等人总结了已有基于土地利用的人口分布模拟方法的不足，根据分县控制、分城乡、分区域建模的思路，建立了基于土地利用的中国 1 km×1 km 分辨率格网人口栅格模拟模型[15]；2006 年，叶宇、刘高焕等人通过多源信息融合，从乡镇行政区划的尺度上对人口数据空间化模拟进行了尝试[16]；2006 年，杨小唤、刘业森等在已有人口空间化方法的基础上，依据遥感影像中所获取的其比重对农村居民地进行重新分级，从而进一步提高了人口空间化模拟的精度[17]。2007 年，黄耀欢、杨小唤等人通过数理统计分析，把平均海拔、居民地、林地等作为主要影响因素，对山东省人口进行空间化，并对其结果进行了分析[18]。

2 多重分类平滑法研究

2.1 三级分类处理 以河南省为例，利用河南省 2005 年 1∶10万土地利用数据和《2005 年中华人民共和国全国分县市人口统计资料》中的人口统计数据，对其进行分类处理。本研究所用的分类共包括 3 级：

（1）划分居住区与非居住区：从土地利用类型图中可明显看出，河南省约有 90% 的地区为非居住地，如果不提前剔除这类区域，必将严重影响模拟精度。在 ArcGIS 系统的支持下，利用 2005 年河南省土地利用类型图将河南省划分为居住区与非居住区两类，并将非居住区单元内的人口密度属性值直接设置为 0。

（2）划分城镇居住区和农村居住区：根据 2005 年河南省土地利用类型图可将居住区划分为城镇居民点与农村居民点两类。由于城镇与农村的人口密度差异极大，故分别计算各县市的城镇人口密度和农村人口密度。

（3）居住区重分类：参考 2003 年廖顺宝等人在青藏高原人口统计数据空间化的研究中可知，各地区人口密度与其居民点密度有很大关系[19]；而 2006 年从杨小唤等人所做的山东省人口数据空间化的研究可知，各地区居住密度与其居住区比重有极大关系。故可检验这一规律是否也适合河南省 2005 年的情况，其中河南省各县市的居住密度、城镇居住区比重和农村居住区比重等数据可取自 2005 年河南省 1∶10 万土地利用数据。从河南省 100 多个县市中抽取 80 个分别进行分析，发现用某一地区的城镇居住区比重、农村居住区比重、城镇居住密度以及农村居住密度分别拟合该地区的城镇居住密度、农村居住密度、城镇人口密度和农村人口密度，所得的决定系数均在 0.6 左右（相当于相关系数的绝对值不低于 0.75）。因此，可依据各县市的城镇居住区比重和农村居住区比重将各城镇或农村居住区重分类（研究中重分类的分类数为 4，分类界限分别是 0.03、0.06 和 0.09），分别计算重地区各城镇或农村居住区平均比重，然后依据拟合函数计算重分类后各单元内的平均人口密度，从而提高模拟精度。

2.2 格网化处理 三级分类处理后，河南省 2005 年人口分布精度已经得到很大提高，但为了方便与其他数据融合，还需将这种以不规则多边形为基本单元的人口分布数据转化成以格网为基本单元的格网化数据。具体算法采用面积权重法及指数平滑法：

（1）首先，依据传统的面积权重法进行初步格网化处理，其算法大致如下：

$$P_{ij}^{0}=\sum_{k=1}^{n}P_k\cdot\frac{S_{ijk}}{S_{ij}} \quad (1)$$

式中，P_{ij}^{0}表示第 i 行、第 j 列格网单元的人口密度属性初值；S_{ij}表示第 i 行、第 j 列格网单元的面积；P_k 表示重分类后的 n 类区域中第 k 类区域的人口密度属性值；S_{ijk}表示第 i 行、第 j 列格网单元与第 k 类区域交叉的区域面积。

（2）经初步格网化处理所得的人口密度图相邻栅格间数值跳跃度通常很大，因此，还需进一步利用平滑法进行平滑处理，使所得的数据更加符合人口分布的连续性特征。研究中采用的是指数平滑法，其理论算法如下：

$$P_{ij}=a\cdot P_{ij}^{0}+(1-a)\cdot\overline{P_{ij}} \quad (2)$$

式中，P_{ij}表示第 i 行、第 j 列格网单元修正后的人口密度属性值；$\overline{P_{ij}}$表示与该格网单元相邻的 9 个格网单元修正后的人口

密度属性值的平均值；a 称为平滑系数，其取值在 (0，1) 上，可通过限定性方程 $\sum_{i=1}^{u}\sum_{j=1}^{v_i}P_{ij}\cdot a^2 = N$ (u 表示人口密度图中格网单元的总行数；v_i 表示第 i 行的列数；N 为河南省的总人口数) 计算而得。

指数平滑法考虑到了格网间的相互影响与格网间距离成反相关这一规律，其平滑效果优于简单平滑法，因此实际应用更广泛。

3　结论与讨论

在 ArcGIS 9.2 的支持下，经过一系列的数据处理，生成了基于地理空间信息技术的河南省 2005 年 1 km ×1 km 分辨率格网人口密度数据。从图 1 中可以看出，从居住区人口分布来看，河南省城镇居住区人口密度明显高于农村居住区人口密度；从行政区区划来看，各市市辖区人口都比同市其他县市人口稠密；从地理位置来看，人口稠密地区主要分布在河南省中部偏北地区；居住区人口密度超过 1.5 万人/km^2 的市主要有郑州市、洛阳市、信阳市、南阳市等城市。

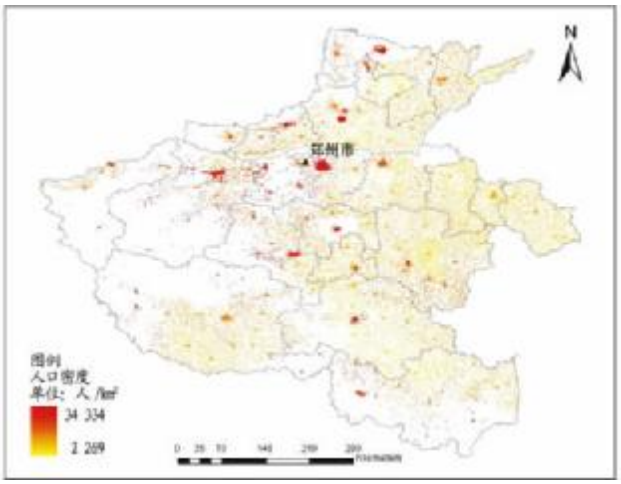

图 1　河南省 2005 年 1 km ×1 km 格网人口密度

Fig. 1　The population density of Henan at 1 km ×1 km resolution in 2005

多重分类平滑法方法除了解决以行政区划为基本单元的数据在数据精度、数据标准化、数据融合等方面存在的问题外，还在现有的人口数据格网化方法的基础上，突出地解决了以下几个问题：①采用多重分类法，在区分居住区与非居住区、城镇居住区与农村居住区的基础上，利用与人口密度分布相关性最强的要素 (如居住密度、居住区比重) 来提高估计精度，避免了因引入过多变量而导致误差过度扩大、多重共线性及操作复杂化等一系列问题；②将城镇人口与农村人口分开处理，提高了数据的实用性，如可进一步应用于城乡二元结构研究或城市化水平测度；③将指数平滑法用于空间问题的分析，并利用限定性方程计算平滑系数，既解决了空间数据的连续性问题，又保持了数据总量的稳定。

多重分类平滑法在根据不同地区特征做适当修改的基础上可推广至全国、区域乃至世界人口分布的研究，还可进一步形成时间序列数据，分析人口分布的动态变化规律。尽管目前所得的人口密度图在精度和连续性上都已有了很大的进步，但要适应不同实际应用的需要，仅采用单一分辨率的人口密度图往往是低效率的。因此，还需对实际问题进行具体的需求分析，通过适当的重采样操作，进一步生成多级格网数据，以满足不同应用类型对数据精度的需求。由于模拟数据的自身特征，通过人口数据格网化的数据密度图主要适用于地区、国家、区域乃至世界的宏观经济管理与规划，如可用于商业点、学校、医院的服务网点和配送中心的选址，物流 (如粮食、能源等) 的交通规划，电路、通讯等网络经济布局。如果对该方法做适当改进，可应用于国家或地区的受教育程度、性别结构、年龄结构和老龄化问题等研究。

参考文献

[1] 符海月，李满春，赵军，等. 人口数据格网化模型研究进展综述[J]. 人文地理，2006，89 (3)：115 －119.
[2] 李德仁，邵振峰. 空间信息多级网格及其功能[J]. 地理空间信息，2005 (4)：1 －3，5.
[3] 张超，万庆，张继权. 基于格网数据的洪水灾害风险评估方法——以日本新川洪灾为例[J]. 地球信息科学，2003 (4)：69 －73.
[4] MENNIS J. Generating surface models of population using dasymetric mapping [J]. Professional Geographer，2003，55 (1)：31 －42.
[5] MARKOFF J，SHAPIRO G. The linkage of data describing overlapping geo －graphicalunits[J]. Historical Methods Newsletter，1973 (7)：34 －46.
[6] GOODCHILD M F，LAM W. Area interpolation：a variant of the traditional spatial problem [J]. Geo-Processing，1980 (1)：297 －312.
[7] FLOWERDEW，ROBERT，GREEN M. Developments in areal interpolation methods and GIS [J]. Annals of Regional Science，1992，26：67 －78.
[8] GOODCHILD M F，ANSELIN L，DEICHMANN U. A framework for the areal interpolation of socioeconomic data [J]. Environment and Planning A，1993，25 (3)：383 －397 .
[9] LANGFORD M，DAVID J，UNWIN D J. Generating and mapping population density surfaces within a geographical information system [J]. The Cartographic Journal，1994，31：21 －26.
[10] WRIGHT JOHN K. A method of mapping densities of population with Cape Cod as an example[J]. Geographical Review 1936，26：103 －110.
[11] 金君，李成名，印洁，等. 人口数据空间分布化模型研究[J]. 测绘学报，2003，32 (2)：278 －282.
[12] EICHER CORY L，CYNTHIA A BREWER. Dasymetric mapping and areal interpolation：Implementation and evaluation [J]. Cartography and Geographic Information Science，2001，28 (2)：125 －138.
[13] 吕安民，李成名，林宗坚，等. 人口统计数据的空间分布化研究[J]. 武汉大学学报：信息科学版，2002，27 (6)：301 －305.
[14] 范一大，史培军，　智慧，等. 行政单元数据向网格单元转化的技术方法[J]. 地理科学，2004，24 (1)：105 －108.
[15] 田永中，陈述彭，岳天详，等. 基于土地利用的中国人口密度模拟[J]. 地理学报，2004，59 (2)：283 －292.
[16] 叶宇，刘高焕，冯险峰. 人口数据空间化表达与应用[J]. 地球信息科学，2006 (2)：59 －65.
[17] 杨小唤，刘业森，江东，等. 一种改进人口数据空间化的方法：农村居住地重分类[J]. 地理科学进展，2006 (3)：62 －69.
[18] 黄耀欢，杨小唤，刘业森. 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[J]. 地球信息科学，2007 (2)：49 －54.
[19] 廖顺宝，孙九林. 基于 GIS 的青藏高原人口统计数据空间化[J]. 地理学报，2003，58 (1)：25 －33.