

# 多元统计回归及地理加权回归方法在多尺度人口空间化研究中的应用

王珂靖<sup>1,2</sup>, 蔡红艳<sup>1\*</sup>, 杨小唤<sup>1</sup>

(1. 中国科学院地理科学与资源研究所资源与环境信息系统国家重点实验室, 北京 100101;

2. 浙江省测绘科学技术研究院, 杭州 310012)

**摘要:**对统计型人口数据进行格网形式的空间化可更直观地展示人口的空间分布,但不同的人口空间化建模方法和不同的格网尺度在表达人口空间化结果方面存在差异。本文在人口特征分区的基础上,引入DMSP/OLS夜间灯光对城镇用地进行再分类,采用多元统计回归和地理加权回归方法(GWR),开展人口统计数据空间化多尺度模型研究,生成1 km、5 km和10 km等3个尺度的2010年安徽省人口空间数据,并对3个尺度下2个模型结果进行精度评价与比较。结果表明:人口空间数据精度不仅与建模所用方法关系密切,还受到建模格网尺度大小的影响。基于多元统计回归方法的模型估计人口数与实际人口的平均相对误差值随着尺度的增加而降低,而基于GWR方法获得的人口空间数据误差值随着尺度的增加而升高。整体来看,基于GWR方法的1 km研究尺度的人口空间数据平均相对误差最低(22.31%)。区域地形地貌条件与人口空间数据误差有较强的关联,地貌类型复杂的山区人口空间数据误差较大。

**关键词:**人口分布;空间化;多尺度;多元统计回归;地理加权回归法;安徽省

## 1 引言

人口统计数据空间化可打破地域限制,实现以行政区域为单元的人口数据向规则格网形式的人口空间数据转换,从而模拟现实的人口分布情况,再现人口实际分布,对解决自然资源环境与人口耦合问题、制定国家宏观经济决策具有重要的意义(江东等, 2002)。

基于不同的人口分布影响因子及各种辅助数据,发展了多种人口统计数据空间化方法,主要包括:空间插值法(Linard et al, 2011)、多源数据融合法(王雪梅等, 2004; Yang, Huang et al, 2009; 杨续超

等, 2013)、遥感估算法(卓莉等, 2005; 曹丽琴等, 2009; Azar et al, 2013)、基于土地利用建模方法(杨小唤等, 2002; 田永中等, 2004; Yang, Ma, 2009)等。其中,土地利用/覆盖的空间格局与人口的空间分布关系紧密,基于土地利用类型与人口分布的关系,建立多元统计回归模型进行人口空间分布模拟的方法应用最为广泛。在此基础上,为体现同一土地利用类型内部人口空间分布的差异性,部分学者通过分析不同地理区位的同一土地利用类型人口分布特征的差异,对土地利用数据进行特征分类以提高原有模型精度(杨小唤等, 2006);有些学者引入夜间灯光数据或其他辅助数据,对土地利用数据进行

收稿日期:2016-01;修订日期:2016-09。

基金项目:国家自然科学基金项目(41271173, 41301155);国家科技支撑计划项目(2012BAI32B06) [Foundation: National Natural Science Foundation of China, No.41271173, No.41301155; National Science and Technology Support Program of China, No.2012BAI32B06]。

作者简介:王珂靖(1988-),女,山东文登人,硕士研究生,主要从事社会经济数据空间化建模研究,E-mail: wkj\_3210@163.com。

通讯作者:蔡红艳(1983-),女,黑龙江五常人,助理研究员,研究方向为人文要素建模、土地利用变化的人文因素分析,  
E-mail: caihy@igsrr.ac.cn。

引用格式:王珂靖, 蔡红艳, 杨小唤. 2016. 多元统计回归及地理加权回归方法在多尺度人口空间化研究中的应用[J]. 地理科学进展, 35(12): 1494-1505. [Wang K J, Cai H Y, Yang X H. 2016. Multiple scale spatialization of demographic data with multi-factor linear regression and geographically weighted regression models[J]. Progress in Geography, 35(12): 1494-1505.]. DOI: 10.18306/dlkxjz.2016.12.006

重分类或特征提取,优化原有模型方法(Zeng et al, 2011; 陈晴等, 2015; 王珂靖等, 2015)。考虑到传统多元统计回归建模方法是全局性建模分析,部分学者开始利用局部回归建模的方法进行人口数据空间化研究,如张建辰等(2014)利用地理加权回归(Geographically Weighted Regression, GWR)的局部分析方法对湖北鹤峰县村级人口统计数据进行人口空间分布模拟,取得了较好效果。

大量研究表明,尺度依赖性存在于各种地理学现象或过程中(李双成等, 2005)。因此,在进行人口统计数据空间化研究时,地理格网尺度问题同样十分重要且不可忽视。目前,在人口数据空间尺度问题方面多集中在格网形式的多尺度,常利用相关性分析和空间自相关方法进行尺度分析(杜国明等, 2007; 王培震等, 2012),模型的研究尺度多依据原始建模数据特征确定,往往是针对某种单一模型方法的多尺度研究(叶靖等, 2010; 王静等, 2012; 李月娇等, 2014)。因此,考虑到尺度效应对模型研究方法 & 结果的影响,还需针对不同建模方法进行多尺度的系统比较。

本文面向人口统计数据空间化模型方法的多尺度问题,针对2010年安徽省县级人口统计数据,在利用DMSP/OLS夜间灯光数据加强城镇居民地内部人口分布特征差异的基础上,结合农村居民地,建立多元统计回归和GWR等2种人口空间化模型,并设置1 km、5 km和10 km等3种格网尺度刻画人口空间分布,通过系统对比和误差分析,揭示不同模型方法在不同研究尺度上的差异。本文的模型方法对比和格网尺度分析,既可为人口空间数据的生产和应用提供科学依据,也可为今后其他类型的统计数据空间化研究提供方法借鉴。

## 2 研究区概况与数据处理

### 2.1 研究区概况

安徽省位于中国经济最发达的长江三角洲外围地区,经济发展较快,但省内自身经济发展不均衡,靠近江浙及铁路沿线交通便利地区经济水平较高,西北部平原地区经济发展较缓慢。全省地貌类型多样,海拔差异明显。主要地形特点为:山地、丘陵、平原各占1/3且相间排列,其中山地和丘陵多分布在南部和西部,如皖南丘陵山地和皖西丘陵山地,而平原地区多分布在长江和淮河流域。安徽省共有78个县市(16个地级市市辖区、62个县)。根据第六次全国人口普查数据,截至2010年底,安徽省常住总人口为5950万。全省城乡人口分布不均,城镇常住人口和乡村常住人口分别占总人口的43.01%和56.99%,且平原和丘陵地区人口分布密集,山地人口分布稀少。

### 2.2 数据源与数据处理

本文使用的基础数据主要包括:人口统计数据、行政区划数据、土地利用数据、DMSP/OLS夜间灯光数据、GDP统计数据、DEM及河流、坡度等基础地理信息数据。数据类型和数据来源见表1。

其中,DMSP/OLS夜间灯光数据采用美国国家地球物理数据中心(NGDC)的F182010平均稳定灯光强度数据,其灰度值范围为0~63,本文主要用于城镇用地再分类处理。土地利用数据为1 km栅格,每个栅格中记录了1 km<sup>2</sup>内某种土地利用类型的面积比例,该数据主要用于构建人口空间化模型。其余数据用于人口特征分区及误差分析比较。

数据处理主要包括数据整合校对、统计数据和空间数据匹配以及投影转换和重采样。各地级市市辖区人口数由该市所辖所有区合并得到。如合

表1 人口数据空间化数据源  
Tab.1 Data sources for spatialization of population distribution

数据名称	数据时相	数据类型	空间数据比例尺/分辨率	数据来源
县级人口统计数据	2010年	表格	/	2010年人口普查资料
县级GDP统计数据	2010年	表格	/	安徽省社会经济统计年鉴
县级行政边界数据	2010年	矢量	1:100万	中国科学院资源环境科学数据中心
河流、交通数据	2008年	矢量	1:100万	中国科学院资源环境科学数据中心
地貌数据	2009年	矢量	1:100万	中国科学院资源环境科学数据中心
ASTER-DEM	1999年	栅格	30 m	中国科学院资源环境科学数据中心
土地利用数据	2010年	栅格	1 km×1 km	中国科学院资源环境科学数据中心
DMSP/OLS(F182010)	2010年	栅格	1 km×1 km(重采样后)	<a href="http://www.ngdc.noaa.gov">http://www.ngdc.noaa.gov</a>

肥市市辖区包括瑶海区、庐阳区、蜀山区、包河区4区,本文将4个行政区划进行合并,整体作为合肥市辖区,再将空间数据属性与合肥市辖区的统计人口数进行关联。以Albers(双标准纬线等面积割圆锥投影)为投影标准,统一对矢量数据(行政区划图)和栅格数据(土地利用数据和夜间灯光数据)进行投影转换。所有栅格数据均裁剪为安徽省范围且采用最邻近重采样法采样成1 km×1 km分辨率数据。

本文还依据夜间灯光数据对城镇居民地内部进行差异划分处理。DMSP/OLS夜间灯光数据灯光强度值对人口空间分布具有指示作用,可作为分级标准对城镇居民地进行重分类,利用重分类后的数据再进行基于土地利用的人口数据空间化处理,结果精度也明显提高。参考王珂靖等(2015)的研究,针对各分区的夜间灯光数据特征设置分级阈值,分别进行夜间灯光数据分级提取,得到反映不同城镇地区特点的夜间灯光分级图。将各分区的夜间灯光分级图与城镇用地进行叠加,实现基于DMSP/OLS夜间灯光对城镇用地的再分类,得到灯光强度值较低、经济相对落后且人口密度较小的城镇用地第一分级,以及灯光强度值高、经济水平相对发达且人口密度大的城镇用地第二分级(图1)。

考虑到各地生态环境和经济发展水平不同,人口分布差异明显,采用同一个空间化模型进行人口空间建模,精度会受到限制,故本文在人口空间建模前预先开展区域人口特征一致性分区。依据王珂靖等(2015)的研究,以人口密度指标进行人口特

征第一次分区,并选取多种相关指标构建人口分布特征指数,对其进行第二次分区。安徽省人口特征分区结果如图2所示,各分区人口主要特征情况见表2。从分区结果可以看出,从分区1到分区4,整体呈现出经济发展水平由高到低、人口密度由高到低、地势由平原到山地的变化趋势。该分区结果能够体现安徽省不同地区区域特点,满足人口分区建模要求。

### 3 研究方法

在人口特征分区基础上,分别进行基于土地利用类型的统计分析建模;并引入夜间灯光数据(DMSP/OLS)进行分级,实现城镇用地再分类;并在此基础上分别采用多元统计回归和GWR等2种方法,针对各自模型特点建立了3种尺度的人口空间数据集,从而系统分析不同模型方法在各种尺度上的精度差异。本文总技术流程如图3所示。

#### 3.1 基于土地利用数据的多元统计回归方法

假设研究区内同一土地利用类型内部人口呈均匀分布,以不同土地利用面积为自变量,人口统计数据为因变量,建立多元统计回归模型,得到各土地利用类型的人口分布系数,据此模拟县级人口分布模式。模型的一般形式如式(1):

$$P_i = \sum_{j=1}^n a_j \times S_{ij} + b \quad (1)$$

式中:  $P_i$  为某分区下第  $i(i=1,2,3,\dots,m)$  县(市)的统

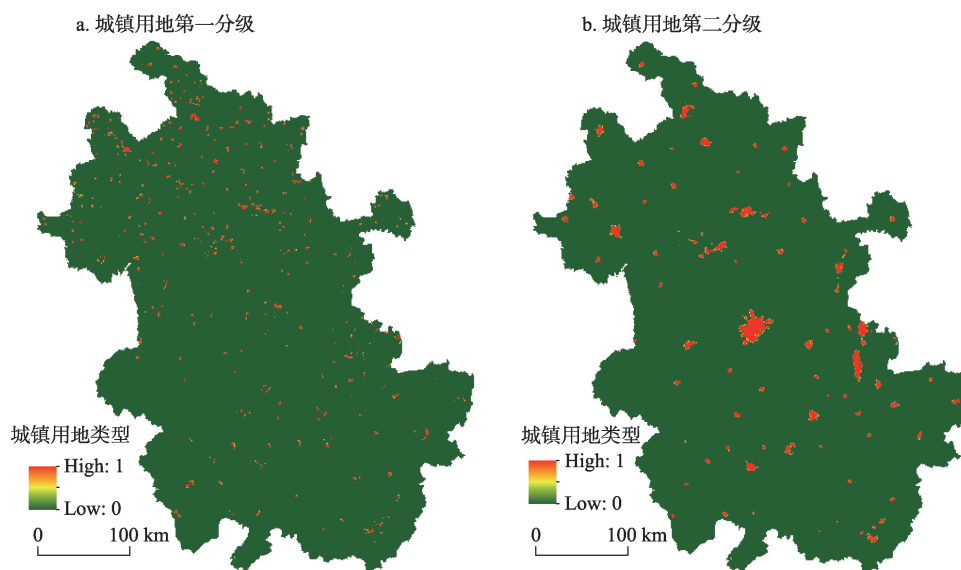


图1 安徽省城镇用地再分类

Fig.1 Reclassified urban land-use data of Anhui Province



计人口数;  $a_j$  为  $j(j=1, 2, 3, \dots, n)$  类土地利用下的人口分布系数/(人/km<sup>2</sup>);  $S_{ij}$  为第  $i$  县(市)  $j$  类土地利用的面积/km<sup>2</sup>。根据“无土地则无人口”的原则, 常数项  $b$  值为 0。多元统计回归建模在 SPSS 软件中实现, 具体以城镇用地第一分级、城镇用地第二分级和农村居民点用地面积为自变量(去掉常数项), 以各县(市)人口统计数据为因变量进行建模。

### 3.2 GWR 方法

GWR 模型是对普通线性回归模型的扩展, 即在回归参数中加入了数据的空间地理信息(Fotheringham et al, 2002)。通过加权最小二乘方法在局部范围内实现逐点参数估计, 根据地理空间位置不断发生变化的参数估计值进行回归分析, 进而直观地探测因地理位置不同而导致的变量之间关系或结构的差异, 即空间非平稳性(Spatial Nonstationarity)。该方法原理简单, 便于操作, 且估计分析结果清晰, 能够进行统计检验, 可与普通线性回归模型进行方法对比。

GWR 是建立局部回归, 在全局模型中加入地

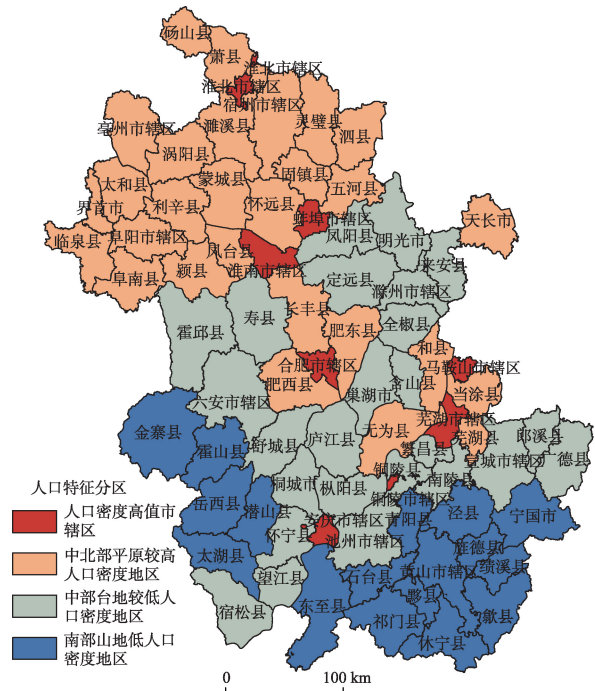


图2 安徽省人口特征分区图

Fig.2 Population regionalization in Anhui Province

表2 安徽省各分区因素特征统计

Tab.2 Statistics of regionalization factors in Anhui Province

分区 序号	分区名称	县(市、区) 个数	人口密度/ (人/km <sup>2</sup> )	GDP 密度/ (万元/km <sup>2</sup> )	河流密度/ (km/km <sup>2</sup> )	坡度/°	高程/m	耕地面积 比例/%	城镇建设用地 面积比例/%
1	人口密度高值市辖区	8	2414	13506	1.02	3	28	48	30
2	中北部平原较高人口密度地区	28	531	783	0.98	2	23	78	15
3	中部台地较低人口密度地区	25	334	607	0.87	4	58	59	7
4	南部山地低人口密度地区	17	155	291	0.58	15	345	19	2

理位置的权重函数(不同观测点处的权重不同, 一般与距离观测点的距离成反比, 即距离观测点越近的观测值权重越大, 反之越小), 使得模型参数在回归过程中不断变化。形式如式(2):

$$y_i = \beta_0(u_i, v_i) + \sum_{k=1}^p \beta_k(u_i, v_i) x_{ik} + \varepsilon_i \quad (2)$$

式中:  $(u_i, v_i)$  是第  $i$  个采样点的坐标;  $\beta_0(u_i, v_i)$  是第  $i$  个采样点统计回归的常数项;  $\beta_k(u_i, v_i)$  是第  $i$  个采样点上的第  $k$  个回归参数;  $x_{ik}$  为第  $i$  个采样点上的第  $k$  个变量;  $p$  为某一采样点上参与回归的变量个数;  $\varepsilon_i$  为误差项,  $\varepsilon_i \sim N(0, \sigma^2)$ ,  $Cov(\varepsilon_i, \varepsilon_j) = 0 (i \neq j)$ 。可简写为式(3):

$$y_i = \beta_{i0} + \sum_{k=1}^p \beta_{ik} x_{ik} + \varepsilon_i \quad (3)$$

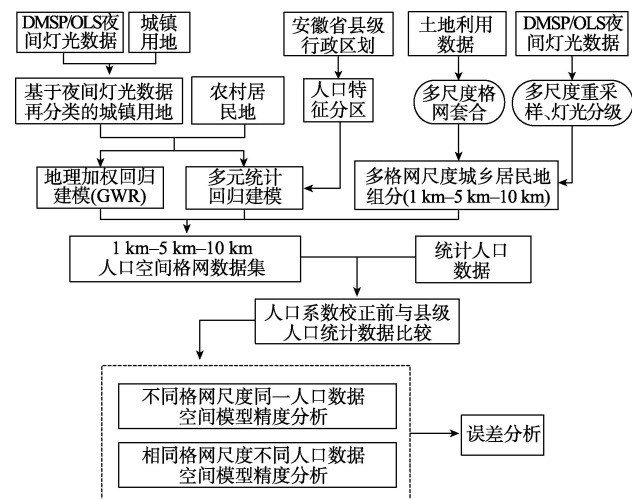


图3 本文技术流程

Fig.3 Flow chart of the study



若  $\beta_{1k} = \beta_{2k} = \dots = \beta_{ik}$ , 则地理加权回归模型就是普通线性回归模型。

本文利用 GWR4 软件进行安徽省人口空间化建模, 考虑到分区 1 的 8 个样本量难以满足建模要求, 而且 GWR 本身在建模中考虑局部邻域的特征, 因此, 不依据人口特征分区结果进行分区建模, 而是对全省各县的人口数据统一进行自动分区的局部回归 GWR 建模。GWR 模型参数选择如下: 选用自适应的二次平方自适应空间核函数(Bi-square)进行建模, 选择默认的黄金分割搜索程序进行带宽选取, 以赤池信息量准则 AIC(Akaike Information Criterion)作为信息评价准则。人口建模因子设置如下: 以行政区划代码为 id 索引值, 以各县(市)行政区划所在质心点坐标(x, y)作为地理位置坐标输入, 同样以城镇用地第一分级、城镇用地第二分级和农村居民点用地面积为自变量(去掉常数项), 以各县(市)人口统计数据为因变量进行建模。

### 3.3 多尺度数据处理方法

本文选取的研究尺度有 3 种: 分别为 1 km、5 km 和 10 km。在进行多尺度人口格网数据转换时, 为保证研究基础数据的一致性, 需要确保县级统计单元内的土地利用面积不发生变化。传统的尺度转换方法会对土地利用面积产生压缩或增加的变化, 而土地利用类型的种类数量对各类型的土地面积误差有相关影响(刘明亮等, 2001), 为避免此现象发生, 本文采用多尺度格网套合的方式: 生成不同尺度大小的规则格网单元, 并与土地利用数据进行套合, 利用区域统计方法统计格网范围内各种土地利用的面积, 进而利用数据转换的方法, 将格网型数据转成栅格数据, 由此得到不同尺度大小的土地利用数据。

### 3.4 模型精度验证和比较方法

不同尺度模型的精度主要采用平均相对误差

和相对误差百分比指标进行评价。具体方法为: 汇总研究区行政边界内的人口空间数据, 与对应的普查人口统计数据作对比, 即依据平差校正前的格网人口, 利用县级行政边界区域进行统计, 将得到的统计值与县市人口统计数据作比较, 从而进行模型对比和误差分析。其中, 平均相对误差值 MPE、相对误差百分比 RE 的计算式如式(4)-(5):

$$MPE = \frac{\sum_{i=1}^m |(RE)_i|}{m} \quad (4)$$

$$RE = \frac{POP_{模拟值} - POP_{统计值}}{POP_{统计值}} \times 100\% \quad (5)$$

式中:  $POP_{模拟值}$  为格网统计得到的县市人口模拟值;  $POP_{统计值}$  为县市的人口普查数据;  $i$  表示第  $i$  个县(市);  $m$  表示安徽省内县市个数。

为揭示不同尺度下模型的表现, 将开展 2 方面的比较研究: 一是基于同一模型生成不同大小格网的人口空间数据, 比较分析区域边界栅格化的影响, 探索同一模型的最优建模尺度; 二是基于不同模型生成同样大小的格网人口数据, 比较分析不同模型方法对人口空间数据精度的影响。

## 4 结果与分析

### 4.1 不同尺度下的人口空间分布

基于多元统计回归模型和 GWR 模型, 分别开展 1 km、5 km 和 10 km 等 3 种尺度的人口空间化建模。所有模型均通过了显著性检验( $p < 0.001$ ), 模型的可决系数( $R^2$ )均高于 0.8(表 3)。

图 4 展示了基于多元统计回归模型和 GWR 模型得到 3 种尺度的人口空间化结果。从图 4 可以明显看出, 不同格网尺度的人口空间数据所表达的信息特征不同: 1 km 尺度格网较细密, 人口空间化结

表 3 模型系数显著性检验结果表  
Tab.3 Significance test of coefficients

分区	基于多元统计回归方法的 $R^2$			Sig.	基于 GWR 方法的 $R^2$			Sig.
	1 km	5 km	10 km		1 km	5 km	10 km	
分区 1	0.98	0.98	0.98	0.00				
分区 2	0.93	0.94	0.93	0.00	0.82	0.83	0.81	0.00
分区 3	0.90	0.90	0.91	0.00				
分区 4	0.83	0.84	0.82	0.00				

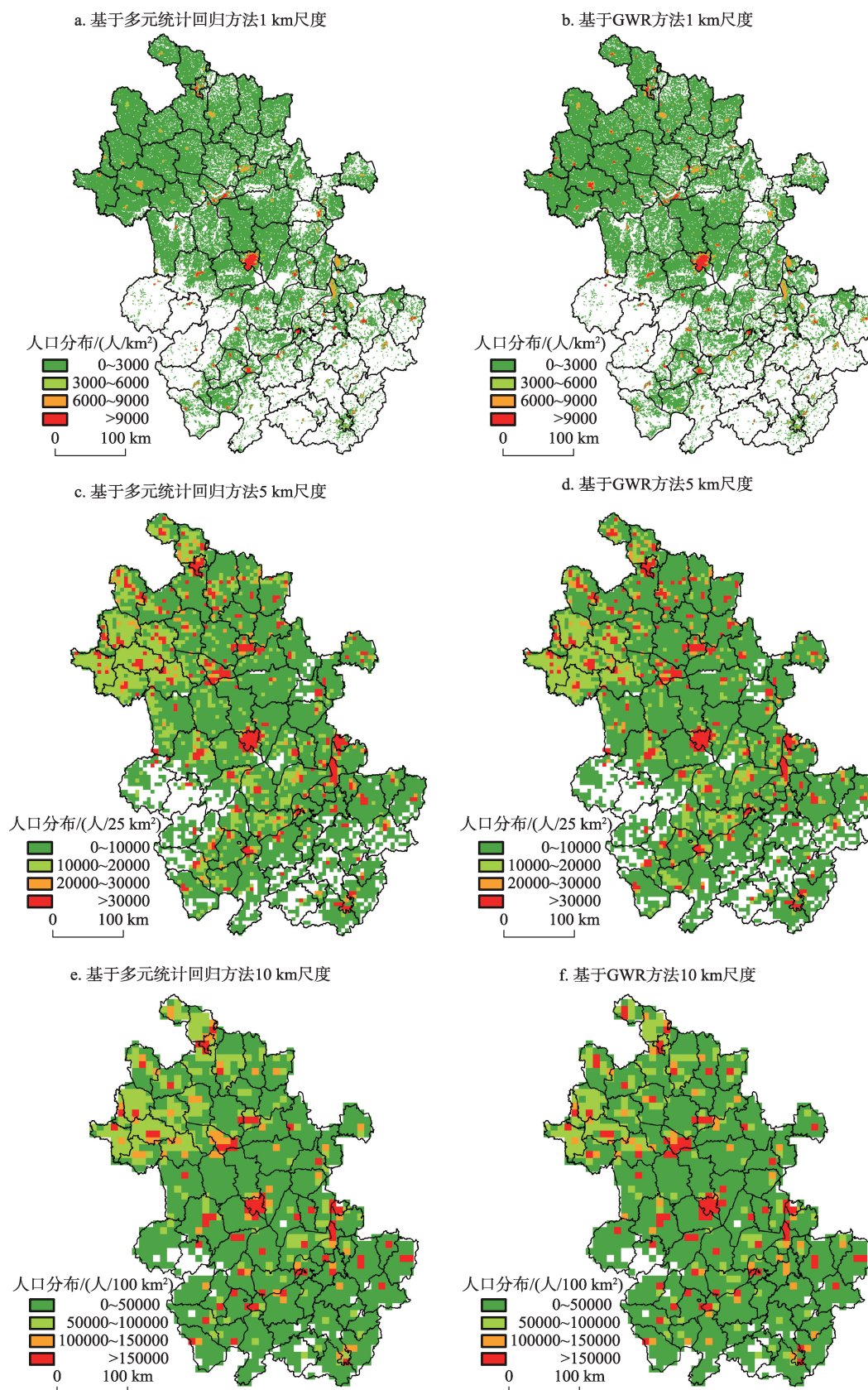


图4 安徽省多尺度人口空间化结果图

Fig.4 Spatialized population distribution in Anhui Province at three scales by multi-factor linear regression and geographically weighted regression methods

果显示效果较平滑,栅格颗粒性不明显(图 4a-4b); 5 km 尺度格网较 1 km 尺度加大,可以较明显地表现出人口空间分布特点,同时栅格的颗粒性也凸显出来(图 4c-4d); 10 km 尺度格网较为粗大,对人口空间分布的展现比较粗略,栅格颗粒性十分显著(图 4e-4f)。对于同一个研究尺度,不同模型方法获得的人口空间化结果整体趋势大致相同,但在细微局部地区结果有差异。

4.2 不同尺度下的模型精度比较

为了获得最适宜安徽省人口空间化的模型方法和研究尺度,对由 2 种模型生成的不同格网尺度人口空间数据进行精度验证和误差分析。

在 3 种尺度下同一方法的结果对比表明(表 4): 对于多元统计回归方法,从全省整体均值相对误差来看,10 km 尺度的平均相对误差百分比最小,模型精度相对较高。此外,3 种尺度都基本呈现从分区 1 到分区 4 相对误差逐渐增大的情况,反映出在山区人口空间化精度明显低于平原地区以及市辖区,有可能是山区地貌类型复杂所致。对于利用 GWR 方法获取的格网人口数据,从全省整体均值相对误差来看,在 1 km 尺度的平均相对误差百分比最小,模型精度相对较高。同样,3 种尺度都基本呈现从分区 1 到分区 4 相对误差值逐渐增大的情况。2 种方法在同一尺度的模型精度对比表明:基于 GWR 方法的整体平均相对误差在 3 个尺度上均低于基于多元统计回归的方法。随着尺度的增加,基于多元统计回归方法的误差值逐渐减小,在 10 km 尺度达到最小值 23.38%;而基于 GWR 方法的误差值则逐渐增大,在 10 km 尺度达到最大值 23.35%。最终在 1 km 尺度上 2 种方法的误差值相差最大,相差达 1.89%;其次为 5 km 尺度,相差 1.35%;相差最小的是在 10 km,仅相差 0.03%。单从整体平均相对误差值角度来看,基于 GWR 模型的人口空间化方法

生成的人口空间化数据误差较低,模型精度较高,且在 1 km 尺度达到模型误差最低。

从误差分段统计表可知(表 5),1 km 尺度上,基于 GWR 方法的误差在<10%的范围区县分布明显高于基于多元统计回归方法,并且误差在>30%的范围区县分布明显降低,因此从整体误差来看基于 GWR 模型方法要小于基于多元统计回归方法;对于 5 km 尺度,虽然基于 GWR 方法的误差在<10%的范围区县分布仍高于基于多元统计回归方法,但误差在>30%的范围区县分布则较多,因此从整体误差来看基于 GWR 模型方法虽小于基于多元统计回归方法,但是二者差距与 1 km 尺度相比有所减少;对于 10 km 尺度,基于多元统计回归方法的误差在<10%的范围区县分布比例最多 33.33%,明显高于基于 GWR 方法,且误差在>30%的范围区县分布又明显降低,因此从整体误差来看 2 种方法差距最小,仅相差 0.03%。综上可知,对于 3 种尺度,基于 GWR 模型方法同多元统计回归方法相比精度较

表 4 安徽省各分区的平均相对误差百分比统计表  
Tab.4 Relative errors for each region in Anhui Province

分区	基于多元统计回归方法的 误差(RE)/%			基于 GWR 方法的 误差(RE)/%		
	1 km	5 km	10 km	1 km	5 km	10 km
分区 1	18.52	17.15	14.51	17.71	12.00	18.09
分区 2	19.94	19.40	22.10	16.89	16.43	18.30
分区 3	27.20	26.57	20.27	24.02	24.18	22.18
分区 4	29.46	29.92	29.67	30.86	34.55	35.86
全省均值	24.20	23.76	23.38	22.31	22.41	23.35

表 5 误差分布范围分段统计表  
Tab.5 County statistics for each error range

模型方法	尺度	误差范围/%	县个数	占总数比例/%
基于夜间 灯光数据 再分类的 多元统计 回归方法	1 km	<10	24	30.77
		10~20	21	26.92
		20~30	9	11.54
		>30	24	30.77
	5 km	<10	21	26.92
		10~20	24	30.77
		20~30	13	16.67
		>30	20	25.64
	10 km	<10	26	33.33
		10~20	22	28.21
		20~30	11	14.10
		>30	19	24.36
基于夜间 灯光数据 再分类的 GWR 方法	1 km	<10	26	33.33
		10~20	19	24.36
		20~30	13	16.67
		>30	20	25.64
	5 km	<10	24	30.76
		10~20	22	28.21
		20~30	10	12.82
		>30	22	28.21
	10 km	<10	22	28.21
		10~20	19	24.36
		20~30	16	20.51
		>30	21	26.92



高,且GWR模型在1 km尺度获取的人口空间化结果的误差值最低(22.31%)。

4.3 模型精度的影响因素

为进一步探讨地形地貌因素对于模型精度的影响,比较了不同高程、坡度及地貌类型分布下模型的误差分布情况。

2种模型方法都呈现由分区1(多市辖区、平原地区)到分区4(山区)误差值逐渐增大的趋势,说明市辖区以及平原地区的人口数据空间化精度要优于偏远山区,地貌类型因素很可能是导致人口空间化误差的原因。按照不同高程和坡度阈值范围,计算各阈值范围内所有区县的平均相对误差百分比,并依据模型方法和格网尺度进行汇总(表6)。

由表6可知,随着高程和坡度的增大,无论是基

于多元统计方法还是基于GWR模型方法,在相同尺度的误差值都基本呈现逐渐增大的趋势,但针对不同的高程和坡度范围,不同模型的误差值也有一定规律:在1 km和5 km尺度,高程值<150 m或坡度值<6°的区县,采用GWR模型方法的误差值均低于采用多元统计回归方法的误差值,说明此时GWR模型更具优势;而当高程值>150 m或坡度值>6°时,采用GWR模型方法的误差值均高于采用多元统计回归方法的误差值,说明此时多元统计回归方法更具优势。在10 km尺度,高程值<50 m或坡度值<3°的区县,采用GWR模型方法的误差值均低于采用多元统计回归方法的误差值,说明此时GWR模型更具优势;而高程值>50 m或坡度值>3°时,采用GWR模型方法的误差值均高于采用多元统计回归

表6 按照高程和坡度误差汇总表  
Tab.6 Error statistics by elevation and slope

	分级阈值 (m/°)	县数量 /个	多元统计方法误差百分比/%			GWR方法误差百分比/%		
			1 km	5 km	10 km	1 km	5 km	10 km
高程	0~50	49	18.78	17.24	19.83	16.60	15.15	16.98
	50~150	11	34.78	36.68	18.83	30.55	32.84	29.38
	>150	18	32.34	33.52	32.25	32.94	35.97	37.42
坡度	0~3	35	20.40	17.96	23.73	17.63	15.83	17.41
	3~6	19	20.77	21.75	15.06	17.15	16.80	18.58
	>6	24	32.44	33.85	26.23	33.27	36.45	35.79

方法的误差值,说明此时多元统计回归方法更具优势。由此可见,不同的高程和坡度对模型方法产生的人口误差影响不同。

为进一步研究人口误差与地貌类型之间的关系,绘制人口误差空间分布图,并与安徽省地貌类型图进行对比。参考周成虎等(2010)对中国地貌图集的研究成果,绘制安徽省地貌类型(图5),以及基于2种方法在不同尺度的误差空间分布图(图6)。

对比图5与图6可知,安徽省人口误差空间分布同该省的地貌类型关系密切:人口误差值空间上呈现北部低南部高的趋势,从地貌类型复杂程度上看,北部地貌类型相对单一,多为平原或台地,而南部地貌类型十分复杂,多为地面起伏度变化巨大的山地。不难看出,人口误差值大的地区其地貌类型复杂程度也较高。具体而言,对于多元统计回归方法,误差值大于40%的区县在整个安徽省分布分散,多数分布在南部山区(如金寨县、石台县等地),但在部分低海拔台地地区也有出现(如定远县);对于GWR方法,误差值大于40%的区县在整个安徽

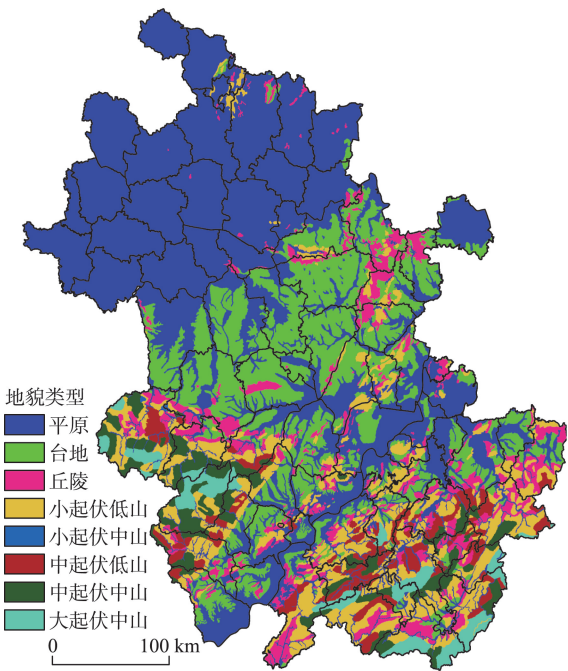


图5 安徽省地貌类型图  
Fig.5 Map of landforms in Anhui province

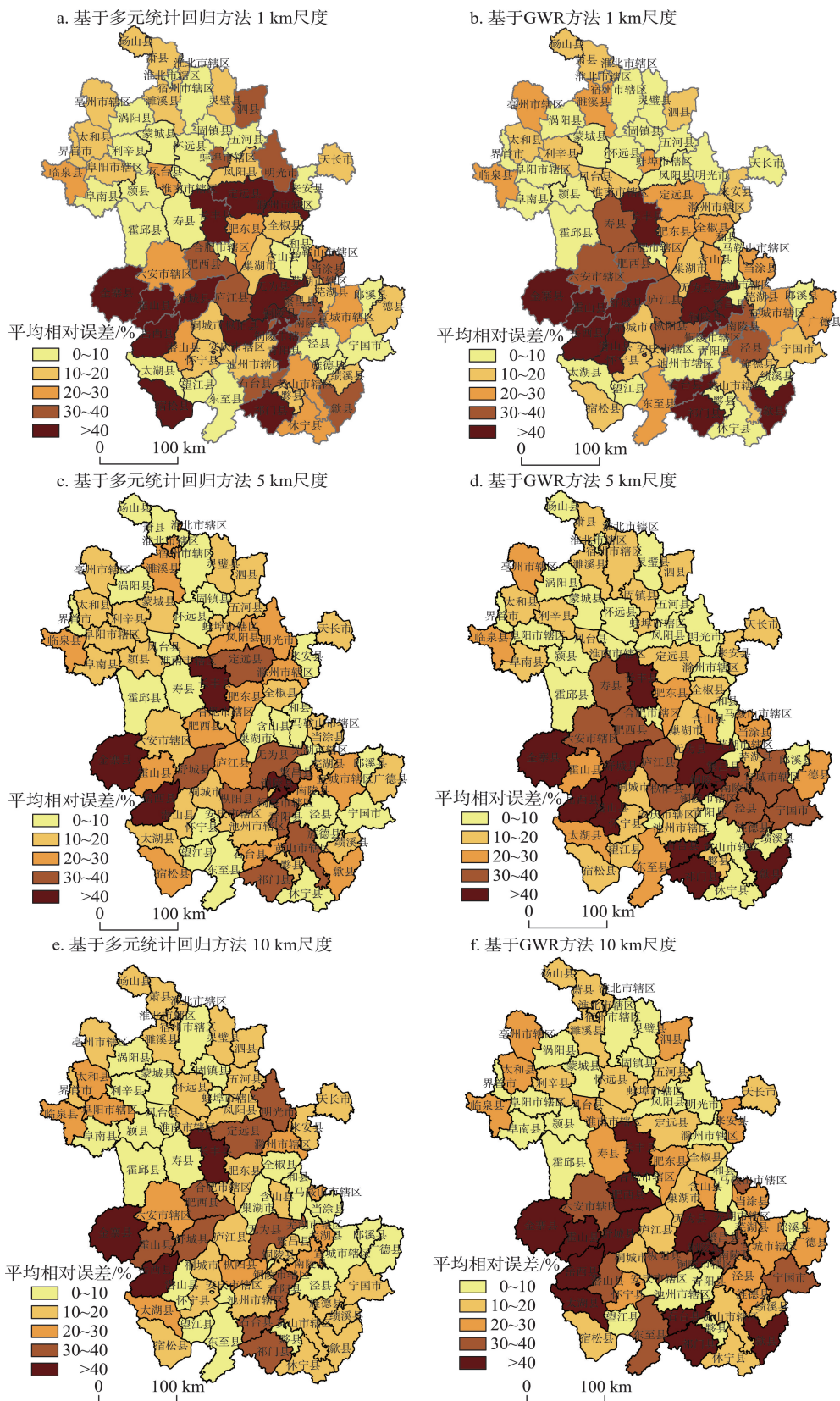


图6 安徽省1 km、5 km和10 km尺度误差空间分布图

Fig.6 Spatial distribution of error at three scales by multi-factor linear regression and geographically weighted regression methods in Anhui Province

省分布较为集中,主要在南部山区(如金寨县),在北部几乎没有分布,原因很可能是由于南方地貌类型复杂,在土地利用数据解译过程中难免出现漏分和错分的情况,从而造成土地利用数据精度低于平原地区,这对基于土地利用数据进行多元统计回归方法或者GWR方法的人口数据空间化模型精度产生影响。

需要指出的是,在利用多元统计回归方法时,长丰县地貌类型仅有台地和平原2种类型,虽同南部山区相比相对简单,但误差值在3个尺度上都异常偏高。可见,除了地貌类型外,地理位置也可能导致误差产生:因为长丰县地处合肥市辖区和淮南市辖区之间,虽然人口特征分区时将长丰县划分为分区2,但其真实的人口分布难免会受到2个分区1的市辖区影响,所以当同其他分区2的区县一起进行人口空间化建模后,其人口模拟值就与真实值出现较大差异。可见,本文所采用的人口特征分区能将研究区域按照相似的社会经济发展水平以及地理环境划分,与按照全省统一建模相比可在整体精度方面有明显提升;但对于某些情况复杂的区域可能描述不够准确,导致个别误差异常值的出现。

## 5 结论与讨论

本文在人口特征分区的基础上,以DMSP/OLS夜间灯光数据灯光强度值作为分级标准对城镇居民地进行重分类,并基于多元统计回归和地理加权回归(GWR)2种方法进行人口统计数据空间化建模处理,生成1 km、5 km和10 km等3种尺度的人口空间数据,针对结果进行模型精度比较和误差分析。研究表明:

(1) 人口空间化数据精度不仅与建模所用的方法关系密切,还会受到生成的格网尺度大小影响。在本文中,随着尺度的增加,基于多元统计回归方法的人口空间数据误差值逐渐减小,而基于GWR方法的误差值则逐渐增大。从整体平均相对误差来看,基于GWR模型方法同多元统计回归方法相比精度较高,且在1 km尺度达到模型误差最小值22.31%。由此可见,加入地理位置的GWR模型方法更能够体现局部特征,尤其是在1 km精细尺度下,相对于普通多元统计回归方法模型精度有进一步的提升。

(2) 2种模型在不同高程和坡度组合的地理环

境下表现有所差异。多元统计回归模型在10 km尺度、高程值<50 m或坡度值<3°的研究区域更具优势,而GWR模型方法在1 km和5 km尺度、高程值<150 m或坡度值<6°的研究区域更具优势。这种差异说明研究区域的高程和坡度特点也是影响人口空间化建模方法选择的一个重要因素,尤其是地理差异性显著的地区,应考虑综合多种模型方法,有针对性地实现空间化。

区域地形地貌条件与人口空间数据误差有较强的关联,地貌类型复杂的山区相对平原地区人口误差较大。本文主要是基于土地利用数据进行人口统计数据空间化处理,因此土地利用数据的精度对人口空间分布结果影响显著。在土地利用数据生产过程中,对于地形相对复杂的山区及平原山地过渡地区,遥感影像的解译精度会受到地形地貌的影响,而土地利用数据精度没有平原地区高,所以造成基于同一种模型方法不同地貌类型的人口精度差异显著。可见,数据源质量对于人口空间化精度十分重要,在当今科学技术迅速发展时代,选用高分辨率的遥感数据对于提高人口空间化精度是一个不错的选择。在未来的研究中,将考虑集成更高分辨率的土地利用数据及其他多源数据开展多尺度人口空间化研究。

## 参考文献(References)

- 曹丽琴,李平湘,张良培. 2009. 基于DMSP/OLS夜间灯光数据的城市人口估算:以湖北省各县市为例[J]. 遥感信息, (1): 83-87. [Cao L Q, Li P X, Zhang L P. 2009. Urban population estimation based on the DMSP/OLS night-time satellite data: A case of Hubei Province[J]. Remote Sensing Information, (1): 83-87.]
- 陈晴,侯西勇. 2015. 集成土地利用数据和夜间灯光数据优化人口空间化模型[J]. 地球信息科学学报, 17(11): 1370-1377. [Chen Q, Hou X Y. 2015. An improved population spatialization model by combining land use data and DMSP/OLS data[J]. Journal of Geo-Information Science, 17(11): 1370-1377.]
- 杜国明,张树文,张有全. 2007. 城市人口分布的空间自相关分析:以沈阳市为例[J]. 地理研究, 26(2): 383-390. [Du G M, Zhang S W, Zhang Y Q. 2007. Analyzing spatial autocorrelation of population distribution: A case of Shenyang City[J]. Geographical Research, 26(2): 383-390.]
- 江东,杨小唤,王乃斌,等. 2002. 基于RS、GIS的人口空间分布研究[J]. 地球科学进展, 17(5): 734-738. [Jiang D, Yang X H, Wang N B, et al. 2002. Study on spatial distribution



- of population based on remote sensing and GIS[J]. *Advances in Earth Sciences*, 17(5): 734-738.]
- 李双成, 蔡运龙. 2005. 地理尺度的转换若干问题的初步探讨[J]. *地理研究*, 24(1): 11-18. [Li S C, Cai Y L. 2005. Some scaling issues of geography[J]. *Geographical Research*, 24(1): 11-18.]
- 李月娇, 杨小唤, 王静. 2014. 基于景观生态学的人口空间数据适宜格网尺度研究: 以山东省为例[J]. *地理与地理信息科学*, 30(1): 97-100. [Li Y J, Yang X H, Wang J. 2014. Grid size suitability of population spatial distribution in Shandong Province based on landscape ecology[J]. *Geography and Geo-Information Science*, 30(1): 97-100.]
- 刘明亮, 唐先明, 刘纪远, 等. 2001. 基于1 km格网的空间数据尺度效应研究[J]. *遥感学报*, 5(3): 183-190. [Liu M L, Tang X M, Liu J Y, et al. 2001. Research on scaling effect based on 1 km grid cell data[J]. *Journal of Remote Sensing*, 5(3): 183-190.]
- 田永中, 陈述彭, 岳天祥, 等. 2004. 基于土地利用的中国人人口密度模拟[J]. *地理学报*, 59(2): 283-292. [Tian Y Z, Chen S P, Yue T X, et al. 2004. Simulation of Chinese population density based on land use[J]. *Acta Geographica Sinica*, 59(2): 283-292.]
- 王静, 杨小唤, 石瑞香. 2012. 山东省人口空间分布格局的多尺度分析[J]. *地理科学进展*, 31(2): 176-182. [Wang J, Yang X H, Shi R X. 2012. Spatial distribution of the population in Shandong Province at multi-scales[J]. *Progress in Geography*, 31(2): 176-182.]
- 王珂靖, 蔡红艳, 杨小唤, 等. 2015. 基于城镇居民用地再分类的人口数据空间化方法研究: 以长江中游4省为例[J]. *遥感技术与应用*, 30(5): 987-995. [Wang K J, Cai H Y, Yang X H, et al. 2015. Spatialization method for census data based on reclassifying residential land use in urban areas: A case study in the middle reaches of the Yangtze River Watershed[J]. *Remote Sensing Technology and Application*, 30(5): 987-995.]
- 王培震, 石培基, 魏伟, 等. 2012. 基于空间自相关特征的人口密度格网尺度效应与空间化研究: 以石羊河流域为例[J]. *地球科学进展*, 27(12): 1363-1372. [Wang P Z, Shi P J, Wei W, et al. 2012. Grid scale effect and spatialization of population density based on the characteristics of spatial autocorrelation in Shiyang River Basin[J]. *Advances in Earth Science*, 27(12): 1363-1372.]
- 王雪梅, 李新, 马明国. 2004. 基于遥感和GIS的人口数据空间化研究进展及案例分析[J]. *遥感技术与应用*, 19(5): 320-327. [Wang X M, Li X, Ma M G. 2004. Advance and case analysis in population spatial distribution based on remote sensing and GIS[J]. *Remote Sensing Technology and Application*, 19(5): 320-327.]
- 杨小唤, 江东, 王乃斌, 等. 2002. 人口数据空间化的处理方法[J]. *地理学报*, 57(S1): 70-75. [Yang X H, Jiang D, Wang N B, et al. 2002. Method of pixelizing population data[J]. *Acta Geographica Sinica*, 57(S1): 70-75.]
- 杨小唤, 刘业森, 江东, 等. 2006. 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*, 25(3): 62-69. [Yang X H, Liu Y S, Jiang D, et al. 2006. An enhanced method for spatial distributing census data: Re-classifying of rural residential[J]. *Progress in Geography*, 25(3): 62-69.]
- 杨续超, 高大伟, 丁明军, 等. 2013. 基于多源遥感数据及DEM的人口统计数据空间化: 以浙江省为例[J]. *长江流域资源与环境*, 22(6): 729-734. [Yang X C, Gao D W, Ding M J, et al. 2013. Modeling population density using multi-sensor remote sensing data and DEM: A case study of Zhejiang Province[J]. *Resources and Environment in the Yangtze Basin*, 22(6): 729-734.]
- 叶靖, 杨小唤, 江东. 2010. 乡镇级人口统计数据空间化的格网尺度效应分析: 以义乌市为例[J]. *地球信息科学学报*, 12(1): 40-47. [Ye J, Yang X H, Jiang D. 2010. The grid scale effect analysis on town leveled population statistical data spatialization[J]. *Journal of Geo-Information Science*, 12(1): 40-47.]
- 张建辰, 王艳慧. 2014. 基于土地利用类型的村级人口空间分布模拟: 以湖北鹤峰县为例[J]. *地球信息科学学报*, 16(3): 435-442. [Zhang J C, Wang Y H. 2014. Simulation of village-level population distribution based on land use: A case study of Hefeng County in Hubei Province[J]. *Journal of Geo-Information Science*, 16(3): 435-442.]
- 周成虎, 程维明. 2010. 《中华人民共和国地貌图集》的研究与编制[J]. *地理研究*, 29(6): 970-979. [Zhou C H, Cheng W M. 2010. Research and compilation of the geomorphological atlas of the People's Republic of China[J]. *Geographical Research*, 29(6): 970-979.]
- 卓莉, 陈晋, 史培军, 等. 2005. 基于夜间灯光数据的中国人人口密度模拟[J]. *地理学报*, 60(2): 266-276. [Zhuo L, Chen J, Shi P J, et al. 2005. Modeling population density of China in 1998 based on DMSP/OLS nighttime light image[J]. *Acta Geographica Sinica*, 60(2): 266-276.]
- Azar D, Engstrom R, Graesser J, et al. 2013. Generation of fine-scale population layers using multi-resolution satellite imagery and geospatial data[J]. *Remote Sensing of Environment*, 130: 219-232.
- Fotheringham A S, Brunson C, Charlton M. 2002. Geographically weighted regression: The analysis of spatially varying relationships[M]. New York: Wiley.
- Linard C, Gilbert M, Tatem A J. 2011. Assessing the use of global land cover data for guiding large area population distribution modelling[J]. *GeoJournal*, 76(5): 525-538.

- Yang X H, Huang Y H, Dong P L, et al. 2009. An updating system for the gridded population database of China based on remote sensing, GIS and spatial database technologies[J]. *Sensors*, 9(2): 1128-1140.
- Yang X H, Ma H Q. 2009. Natural environment suitability of China and its relationship with population distributions[J]. *International Journal of Environmental Research and Public Health*, 6(12): 3025-3039.
- Zeng C Q, Zhou Y, Wang S X, et al. 2011. Population spatialization in China based on night-time imagery and land use data[J]. *International Journal of Remote Sensing*, 32(24): 9599-9620.

## Multiple scale spatialization of demographic data with multi-factor linear regression and geographically weighted regression models

WANG Kejing<sup>1,2</sup>, CAI Hongyan<sup>1\*</sup>, YANG Xiaohuan<sup>1</sup>

(1. State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, CAS, Beijing 100101, China;

2. Zhejiang Academy of Surveying & Mapping, Hangzhou 310012, China)

**Abstract:** Population distribution data are essential for socioeconomic and environmental studies, such as population estimation, spread of disease, natural disaster relief, and environmental protection. Existing research has proved that spatialized population grid data can precisely delineate the spatial pattern of population distribution, while model selection and size of grids may influence the accuracy of population distribution modeling. It is therefore important to estimate population distribution using appropriate models and at a proper spatial scale. This study mainly focused on the spatialization modeling of Anhui Province county-level population census data in 2010 at three grid scales. Anhui Province was selected for the study due to its complex landforms and significant difference of population distribution within its area. Population regionalization was carried out as a preprocessing step: 78 counties in Anhui Province were divided into four groups. Combining with land-use data and nighttime light (DMSP/OLS), urban residential areas were reclassified to reflect regional differences. Based on the population regionalization, multi-factor linear regression (MFLR) and geographically weighted regression (GWR) models were employed to integrate the reclassified urban residential land-use data with the rural residential land-use data. This study established three population spatial datasets at 1 km, 5 km, and 10 km grid scales. Comparing the two models' precision at each scale, the results show that the modeling and grid scale have much influence on the accuracy of the spatialization result, which increased with the grid scale by using the MFLR model and the highest accuracy was achieved in the 10 km grid datasets. For the GWR model, the accuracy decreased as the grid scale increased, and the highest model accuracy was obtained at the 1 km scale. Overall, the GWR model had a higher accuracy (22.31%) than the MFLR model when taking into account the geographic location and local modeling. This study may provide a scientific basis for the production and application of population spatial data and provide a reference of spatialization for other types of statistical data in the future.

**Key words:** population distribution; spatialization; multi-scales; multi-factor linear regression; Geographically Weighted Regression (GWR); Anhui Province