doi: 10.6046/zrzyyg.2020297
引用格式: 肖东升,练洪. 顾及参数空间平稳性的地理加权人口空间化研究[J]. 自然资源遥感,2021,33(3):164 - 172. (Xiao D S, Lian H. Population spatialization based on geographically weighted regression model considering spatial stability of parameters[J]. Remote Sensing for Natural Resources, 2021, 33(3): 164 - 172.)

顾及参数空间平稳性的地理加权人口空间化研究

肖东升[1,2,3], 练 洪[1,2]
(1. 西南石油大学土木工程与测绘学院,成都 610500; 2. 西南石油大学测绘遥感地理信息防灾应急研究中心,成都 610500; 3. 四川师范大学公共安全与应急研究院,成都 610068)

摘要: 近年来,人口空间化的方法理论愈趋成熟,但对人口空间化建模中变量参数的空间平稳性处理却鲜有人关注。以土地利用数据、夜间灯光数据和人口统计数据为数据源,提出一种基于半参数地理加权回归模型(semi - parametric geographically weighted regression, S - GWR)的人口空间化方法,并利用该模型在县级尺度进行常住人口空间化建模,最后以四川省为研究区进行比较论证。在分析变量特征的同时,利用 S - GWR 模型处理参数变量的空间平稳性,以提高人口估计的精度,最后生成四川省 2010 年 1 km 分辨率的人口空间分布图(spatial distribution of population, SDP)。结果表明, S - GWR 模型的决定系数为 0.903,比传统回归模型表现更好,模型拟合的效果更优。精度验证方面,通过 2 个常用的人口数据集进行精度对比验证; 在县一级,研究区整体 SDP 的平均误差和每个区县的相对误差都接近于 0,比其他 2 个数据集有更高的精度; 在乡镇一级,SDP 的平均相对误差、平均绝对误差和均方根误差分别为 34.54%, 5 715.703 人和 12 085.932 人,均比其他 2 个数据集的误差更小,离散度效果更优; 从乡镇准确估计个数来看,SDP 准确估计的个数最多,达 185 个。因此,考虑参数的空间平稳性可以提高人口空间化的精度。

关键词: 半参数地理加权回归; 空间平稳性; 夜间灯光数据; 土地利用; 人口空间化
中图法分类号: TP 79　**文献标志码**: A　**文章编号**: 2097 - 034X(2021)03 - 0164 - 09

0 引言

人口是社会学、地理学、环境学等学科研究的重要基础,准确估计人口对许多国家都具有重要意义。精确的人口空间分布情况,不仅为政府制定合适的人口相关政策奠定重要基础,制定区域长远发展计划提供参考,还对人口分布与社会经济协调发展有着重要的参考价值,为资源配置和行政管理提供依据。目前,世界上大多数国家或地区实现人口调查的主要渠道是统计和分析,包括抽样调查和全体普查 2 种方式[1]。虽然人口调查和统计具有权威、系统、规范等优势,但是存在时间分辨率低、更新周期长、空间化精度低、不利于可视化和空间分析操作等问题,对人口空间分布的研究难以满足[2]。而人口空间化可以弥补人口统计数据的缺陷和不足,并且可以与其他更精细的空间数据集结合进行分析,以促进人口相关研究的发展。

DMSP/OLS 夜间灯光数据最初是用来探测云层对月光的反射以分析云层分布信息,后来被广泛用于获取地表夜间灯光以反映人类活动情况[3],并且证明有着极好的适用性[4]。但是夜间灯光数据的分辨率较低,且存在着像元饱和、溢出等现象,导致单一的夜间灯光数据只适用于大中尺度人口空间化的相关研究[5-6]。目前,基于人口统计数据和空间变量之间的关系来建立数学模型从而获取人口格网数据是研究人口空间化的热点。常用的方法主要有多源数据融合法[7-8]、夜间灯光与土地利用结合方法[9-10]、空间插值模型[11]等。此外,部分学者结合传统最小二乘线性(ordinary least square, OLS)全局模型将人口统计数据重新分配在地理空间上,默认

收稿日期: 2020 - 09 - 23; **修订日期**: 2020 - 12 - 01
基金项目: 国家自然科学基金项目"基于人类动力学的面向震后救援的人员在地理建筑空间的分布规律研究"(编号: 51774250)、四川省科技厅软学科项目"基于移动终端的室内定位技术的面向地震救援的人群在地理空间分布规律研究"(编号: 2019JDR0112)、四川省科技创新(苗子工程)培育项目"基于遥感的成都地区 PM2.5 的时空变化规律及应对措施"(编号: 2019089)和"基于人类动力学的地震应急救援决策辅助系统的研究"(编号: 2020120)共同资助。
第一作者: 肖东升(1974 -),男,博士,教授,主要从事空间信息技术与防灾减灾方面的研究。Email: 345083896@qq.com。
通信作者: 练 洪(1996 -),男,硕士研究生,主要从事地理空间信息与人员分布方面的研究。Email: 1845705004@qq.com。

模型所有参数都不随地理位置变化,即在空间上是平稳的,保持全局一致性,导致各变量在不同位置上的"平均行为"[12]。有些学者利用局部地理加权回归(geographically weighted regression,GWR)建模的方法进行人口数据空间化研究,默认所有参数在不同地理空间位置是不一样的,具有空间非平稳性[13-14],而实际上有的变量在不同地理空间位置的参数是相同的,即具有全局效应。也有学者使用分区建模对变量特征进行重分类,优化原有模型方法[15-16],尽管强调了分区间的差异,但是对分区内的差异仍然无法揭示[17]。因此,鉴于上述空间化方法的优缺点,本研究考虑变量的空间平稳性,采用变量的局部和全局模式进行混合地理加权回归,以提高人口空间化精度。

综上,本文旨在利用夜间灯光数据、土地利用数据和人口统计数据,基于半参数地理加权回归模型(semi-parametric geographically weighted regression,S-GWR),提出了一种新的考虑参数平稳性的人口精确空间化方法,以四川省为研究区域进行比较和验证。本文以夜间灯光与土地利用数据为权重因子,建立人口模型;在分析变量特征的基础上,采用 S-GWR 模型处理变量的空间平稳性,减少区域误差。最后生成四川省 2010 年 1 km 分辨率的人口空间分布图(spatial distribution of population,SDP),并利用 2 个常用的数据集进行县乡分级精度验证。此外,本文通过 OLS,GWR 和 S-GWR 3 种回归模型进行比较和评价,分析不同模型的变量参数不同对人口空间化的影响。

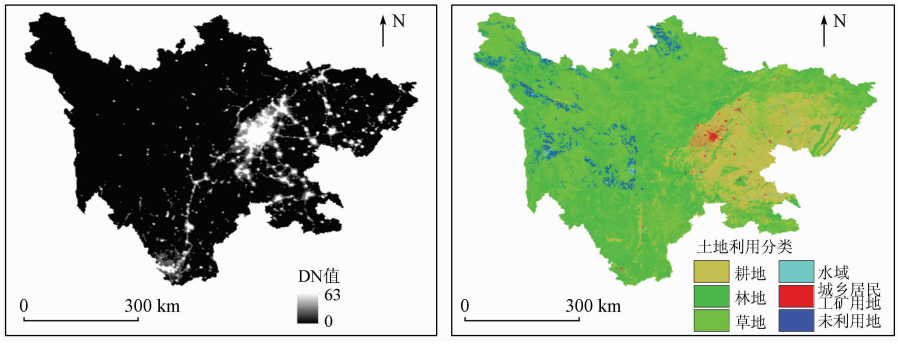
1　研究区概况及数据源

1.1　研究区概况

四川省位于中国大陆西南腹地,地处长江上游,青藏高原和长江中下游平原的过渡带,介于 E97°21′~108°33′和 N26°03′~34°19′之间,总辖区面积约 484 144.02 km^2。辖区有 21 个市级行政区,包括 18 个地级市和 3 个少数民族自治区,共计 181 个县级行政区[18]。四川省具有联动东西、带动南北的区位优势,是我国实施西部大开发战略的重点地区之一,是中国"一带一路"倡议下的丝绸之路的重要陆上出口区域[19]。四川省是中国西部人口重要的聚居地之一,2010 年常住人口 8 041.75 万人,其中城镇人口 3 231.2 万人,农村人口 4 810.55 万人。由于经济和地理上的差异,总体呈现川东地区人口密度高于川西地区的格局。四川省地震、洪涝和泥石流等自然灾害多发,加上其地形地貌复杂,所以研究四川省的人口空间分布可以为防灾减灾提供技术支持和维持区域平衡发展提供决策。

1.2　数据源及其预处理

1)夜间灯光数据。本研究使用的夜间灯光数据(图 1(a))来源于美国地球物理国家数据中心(National Geophysical Data Center, https://NGDC.noaa.gov/eog/DMSP/download_radcal.html),选取 2010 年发布第四版分辨率为 30″的 DMSP/OLS 夜间灯光稳定值数据,该数据通过了去云处理,并且消除了背景噪声及短时光数据如火山气体、森林火灾、极光等。数据的栅格范围值(DN 值)为 0~63,0 为黑暗无灯光区域,大于 0 为灯光区域。夜间灯光数据投影为 Krasovsky_1940_Alebers 坐标系,采用最近邻重采样算法将投影后的地图重采样到 1 km,然后根据研究区域行政边界对影像进行掩模提取,最后得到四川省范围的夜间灯光影像。



(a) 夜间灯光数据图　　(b) 土地利用分布图

图 1　四川省夜间灯光数据图和土地利用分布图

Fig. 1　Night light data map and land use distribution map of Sichuan Province

2)土地利用数据。本研究的土地利用数据来源于由中国科学院资源环境科学数据中心。采用 2010 年 1∶10 万的土地利用数据集,该数据集根据分级分类系统分为 6 个一级土地利用类别(耕地、林地、草地、水域、城乡工矿居民用地和未利用地)和 25 个二级类别(水田、灌木林、沙地、沼泽地等)

(图 1(b))[20]。为了数据后续使用,利用 ArcGIS 将土地利用分辨率转换为 1 km,并通过渔网工具将 25 个二级子类土地利用类型分别输出为 25 个栅格数据文件,每个栅格数据层代表了不同的土地利用类型。

3)人口统计数据。本研究的人口统计数据指的是常住人口数据,来源于四川省统计局的《四川省统计年鉴 2010》。由于行政单元边界与人口普查数不据不完全匹配,需要利用 ArcGIS 软件将属性数据与行政单位相应的空间数据进行关联,最终获得 181 个县有效数据。

4)行政区划数据。县乡两级行政区划数据来源于原国家测绘局。

5)其他辅助数据。本研究还采用中国科学院资源环境科学数据中心发布的中国格网人口分布数据集(grid population distribution of China, CGPD)和美国国际地球科学网络中心发布的第四版世界格网人口(grid population of world, GPWv4)。将上述数据集投影为 Krasovsky_1940_Alebers 坐标系,采用双线性重采样算法将分辨率重采样为 1 km,然后根据研究区域行政边界对影像进行提取。具体数据如表 1 所示。

表 1 数据类型及来源

Tab. 1 Data type and source

数据类型	数据年份	分辨率(比例尺)	数据来源
DMSP/OLS 数据	2010 年	30″	美国地球物理国家数据中心
土地利用数据	2010 年	1∶10 万	中国科学院资源环境科学数据中心
人口统计数据	2010 年	县、乡镇	四川省统计局
行政区划	2010 年	1∶10 万	原国家测绘局
CGPD	2010 年	1 km	中国国家资源环境科学数据中心
GPWv4	2010 年	30″	美国国际地球科学网络中心

2 空间化方法与模型构建

在 SPSS 软件下,将土地利用和人口数据进行相关性分析,得出与人口分布显著正相关的土地利用类型。然后基于 ArcGIS 提取 DMSP/OLS 的亮元、暗元和灯光辐射区域,再与选定的土地利用类型进行叠加分析,得到各土地利用类型的灯光。通过行政区划分区统计后,将变量空间平稳性纳入人口空间化模型,利用 GWR4.0 软件对变量进行地理变异性检验,以区分变量的全局和局部模式,最后通过 S-GWR 模型生成研究区的像元人口数据。具体流程如图 2 所示。

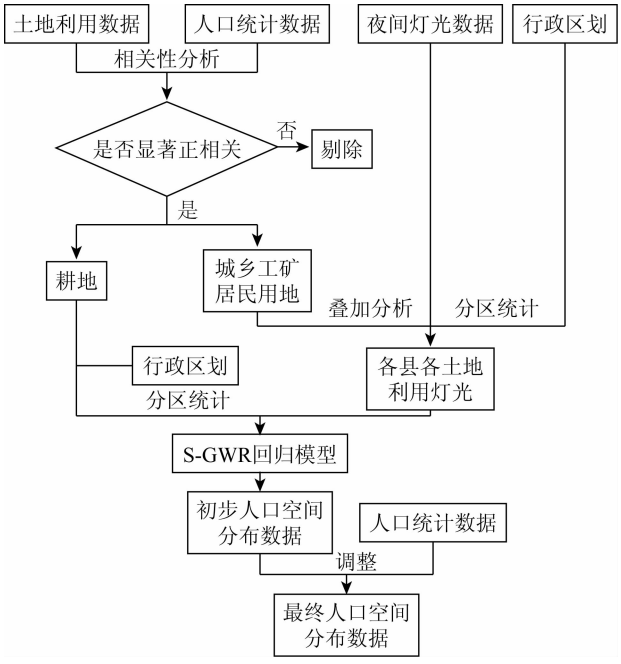


图 2 人口空间化流程图

Fig. 2 Flow chart of population

2.1 相关性检验与空间叠加分析

本研究利用皮尔逊相关系数(Pearson correlation coefficient, PCC)检验方法来获取与人口相关的土地利用类型。在统计学中,皮尔逊相关系数可简称为相关系数(R),是一个用来衡量变量 x 和 y 之间的线性相关关系的指标。计算公式为:

$$R = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - \left(\sum x_i\right)^2}\sqrt{n\sum y_i^2 - \left(\sum y_i\right)^2}}, \tag{1}$$

式中:R 为相关系数的值;x_i 为第 i 县的统计人口数据;y_i 为第 i 县的某一土地利用类型面积;n 为县的个数。

根据人口分布的实际情况,本研究在土地利用数据与人口统计数据叠加过程中,水域和未利用土地不参与空间化分析。利用 ArcGIS 将不同土地利用类型面积根据县界进行分区统计,基于 SPSS 软件对土地利用与人口进行相关性检验。然后通过 ArcGIS 提取 DMSP/OLS 数据的灯光区、无灯光区和灯光辐射区,选取与人口数据显著正相关的土地利用类型,采用空间分析工具中的叠加分析,将上述数据分别进行叠加统计,根据县级行政区划数据进行分区统计,最后得到各区县各类土地的灯光区面积像元数(the number of light pixels, NL)、无灯光区面积像元数(the number of unlit pixels, NU)和灯光辐射

总亮度值(light emission in pixels,LE)。在实际人口分布中,人口只存在于城乡及建设用地等建成区,而本研究考虑了耕地是由于卫星遥感对土地利用产品解译时的精度问题和像元混合问题,忽略了在林地、草地等都有可能存在零星分布农村居民点、农牧民独立房屋、帐篷、毡房等设施,这些分散零星但数量众多的居住设施在 1∶10 万的土地利用中是无法展现出来但又是确实存在的。因此,为了不影响对农村人口估计的低估和对城市人口的高估,将其他土地利用类型赋予一定的权重并纳入人口建模,并基于 ArcGIS 在县一级对其面积进行分区统计。

2.2　人口空间化模型

全局 OLS 模型是假定全部变量之间的空间关系都是稳定的,即得到的回归系数估计值就是整个研究区域内的平均值。而 GWR 模型是全局回归模型的扩展,即在计算回归参数时加入变量的空间地理位置信息,使得不同地理位置的回归参数值不同,因而提高人口空间化建模的精度。然而,由于生活环境和经济水平的不同,参数在不同地理位置有可能是会发生变化的,也有可能是固定的。因此,本研究利用混合固定系数和变化系数的 S - GWR 模型对人口空间化进行建模。与单纯性的全局或局部的方法相比,混合全局固定参数和局部变化参数实现了半参数空间平稳,而且模拟效果比其他模型表现得更好。在建立模型之前有必要对统计人口数进行空间自相关检验,采用 ArcGIS 软件中的空间统计工具分析空间自相关情况,通过 Moran' s I 指数值反映出研究区人口分布的集聚程度,取值范围介于[-1,1]之间。S - GWR 模型计算公式为:

$$p_i = \sum_{l=1}^{k} \alpha_l z_{il} + \sum_{j=1}^{m-k} \beta_j(u_i, v_i) x_{ij} + \varepsilon_i \quad , \quad (2)$$

式中: p_i 为第 i 县的估计人口数; m 为模型中变量的个数; k 为模型中全局变量的个数; α_l 为第 l 个全局变量 z_{il} 的固定系数; (u_i, v_i) 为第 i 县的质心坐标; x_{ij} 为第 i 县的第 j 个局部变量; $\beta_j(u_i, v_i)$ 为第 j 个局部变量 x_{ij} 的地理变化系数; ε_i 为满足球面摄动假设的随机误差。此外,当 $k = 0$ 时,式(2)就变成了局部 GWR 模型。

计算出像元级的估计人口数据后,对初步估计人口结果进行优化和校正,确保预测的 SDP 总人口等于县级行政单位的人口普查数据。计算公式为:

$$pop_i' = pop_i \times \left(\frac{\bar{p}_i}{p_i}\right) \quad , \quad (3)$$

式中: pop_i' 为第 i 县上调整后的像元级人口; pop_i 为初步估计的像元级人口; $\bar{p}_i$ 为第 i 县人口普查数据; p_i 为第 i 县的所有像元级人口之和,即估计人口数据。

为区分变量的全局和局部模式,基于 GWR4.0 软件对全部变量进行地理变异性测试。具体参数模型设置是选用自适应的二次平方空间核函数(Bi - square)进行建模,带宽选取采用默认的黄金分割搜索程序,以赤池信息量准则(Akaike information criterion,AIC)作为信息评价准则,决定系数 R^2 和调整决定系数 $adjR^2$ 对回归性能进行评价。其中,在样本小的情况下,AIC 转变为 AICc ,AICc 值可以反映模型的拟合优度和模型复杂度,在针对同一套因变量和自变量数据时,根据经验法则,当差值大于或等于 3,就表明模型有了明显改善。

2.3　精度评价

对得到的模拟结果有必要进行精度评估和误差分析,除了上述提到的相关系数 R、调整决定系数 $adjR^2$、赤池信息量准则 $AICc$ 等对模型进行评估外,本研究还选取平均绝对误差(mean absolute error, MAE)、平均相对误差(mean relative error,MRE)、均方根误差(root mean square error, RMSE)、相对误差(relative error,RE)、平均误差(mean error,ME)来对结果进行评价。计算公式如下:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |p_i - \bar{p}_i| \quad , \quad (4)$$

$$MRE = \frac{1}{n}\sum_{i=1}^{n} \frac{|p_i - \bar{p}_i|}{\bar{p}_i} \quad , \quad (5)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (p_i - \bar{p}_i)^2}{n}} \quad , \quad (6)$$

$$RE_i = \frac{p_i - \bar{p}_i}{\bar{p}_i} \quad , \quad (7)$$

$$ME = \frac{1}{n}\sum_{i=1}^{n} \frac{p_i - \bar{p}_i}{\bar{p}_i} \quad , \quad (8)$$

式中: n 为研究区域县的个数; p_i 为第 i 个县的人口估计数据; $\bar{p}_i$ 为第 i 个县人口普查数据。

3　结果与讨论

3.1　模型自变量参数

3.1.1　人口与土地利用相关性

通过 SPSS 软件计算各土地利用类型和人口之间的相关性,考虑到人口分布的实际情况,水域和未利用土地未参与相关性分析。结果表明,耕地下的 2 个二级子类(水田、旱地)和城乡工矿居民用地的

3 个子类(城镇用地、居民用地、其他建成区)同人口数据显著正相关,而林地、草地、水域和未利用土地均与人口显著负相关或不相关。其中,在双尾检测时,城乡居民工矿用地下的其他建成区检测结果显示为不相关,但在单尾检测时结果是显著正相关的。因此,为了提高对人口估计的精度,本研究将其作为一个变量纳入人口空间化模型。表 2 显示了土地利用与人口之间的相关性。

表 2 各土地利用类型与人口数据的相关系数

Tab. 2 The correlation coefficient between land use types and population data

土地利用类型	相关系数	土地利用类型	相关系数
水田	0.734 **①	中覆盖度草地	-0.448 **
旱地	0.555 **	低覆盖度草地	-0.323 **
有林地	-0.438 **	水域	—
灌木林	-0.488 **	城镇用地	0.554 **
疏林地	-0.108	农村居民用地	0.387 **
其他林地	-0.115	其他建成区	0.144 *
高覆盖度草地	-0.356 **	未利用地	—

①** 表示在 0.01 级别(双尾)相关性显著; * 表示在 0.05 级别(单尾)相关性显著。

3.1.2 空间模型参数

对人口做自相关检验,得到县级人口 Moran's I 指数值为 0.358,z 值为 21.95,表示人口数据在 0.01 水平上显著自相关,说明 181 个县域的人口分布具有明显的集聚性。在分析土地利用与人口数据的相关性后,选取城镇用地、农村居民用地、其他建成区与 DMSP/OLS 灯光数据进行叠加分析,得到 3 个子类的灯光区面积像元数(NL)、无灯光区面积像元数(NU)、灯光辐射总亮度值(LE)。然后对水田和旱地赋予一定的权重,将上述 11 个参数作为人口空间化模型的变量。基于 GWR4.0 软件对全部变量进行参数估计及参数平稳性检验,利用参数在没有空间变异性的情况下,参数的 F 统计量就遵循一定自由度的 F 分布,最后通过"DIFF of Criterion"结果以区分全局变量和局部变量(表 3)。结果表明,城镇用地 NU 和其他建成区的 LE,NL,NU 的"DIFF of Criterion"大于 2,说明在空间上不具备空间非平稳性,故将其作为全局变量,而将其余 7 个变量作为 S-GWR 模型的局部变量。此外,可以通过 AICc 值来选取最优带宽值,本研究最佳带宽值为 62。基于 GWR4.0 软件进行地理变异性测试结果如表 4,该表显示了全局 OLS、局部 GWR 和半参数混合 S-GWR 模型的性能及拟合优度,评价标准包括 R^2,$adjR^2$ 和 $AICc$ 值。当所有变量都作为全局变量的时候,OLS 回归模型的解释力达到 0.798;当把所有变量作为局部变量时,考虑到变量的局部影响,解释力进一步增加到 0.877,而 $AICc$ 值从 4 846 降到了 4 810,模型得到显著提升;而当采用变量的混合模式时,S-GWR 模型的解释力增加为 0.903,同时 $AICc$ 值下降到 4 786。虽然全局 OLS 模型和局部 GWR 模型都能得到较好的人口空间化结果,但是 S-GWR 模型进一步提高了人口空间化的解释力,并且提高了人口空间化的精度。因此,考虑参数的空间平稳性,能够使得模型拟合得更好。

表 3 地理加权模型参数估计及参数平稳性检验

Tab. 3 Parameter estimation and parameter stationarity test of geographically weighted model

变量	F 统计量	F 检验自由度	DIFF of Criterion
水田	2.755 253	2.266 149.509	-0.721 729
旱地	6.148 736	2.850 149.509	-11.687 572
城镇用地 LE	22.965 443	2.377 149.509	-49.330 235
城镇用地 NL	11.587 845	2.045 149.509	-20.590 824
城镇用地 NU	1.382 923	1.882 149.509	2.441 081
农村居民用地 LE	1.663 454	0.495 149.509	0.483 673
农村居民用地 NL	0.761 547	0.615 149.509	1.267 335
农村居民用地 NU	5.406 091	3.707 149.509	-11.939 404
其他建成区 LE	0.089 571	2.341 149.509	6.646 620
其他建成区 NL	0.666 850	2.706 149.509	5.785 913
其他建成区 NU	0.627 823	1.342 149.509	2.966 391
最优带宽	62.000	最小 AICc	4 786.266

表 4 3 种模型的拟合优度评价

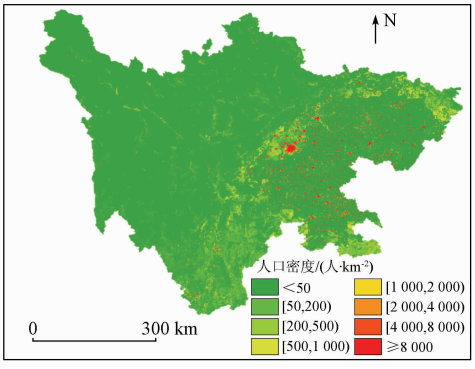
Tab. 4 Evaluation of goodness of fit of three models

评价指标	全局 OLS	局部 GWR	半参数 S-GWR
R^2	0.798	0.877	0.903
$adjR^2$	0.785	0.843	0.867
$AICc$	4 846.167	4 810.764	4 786.267

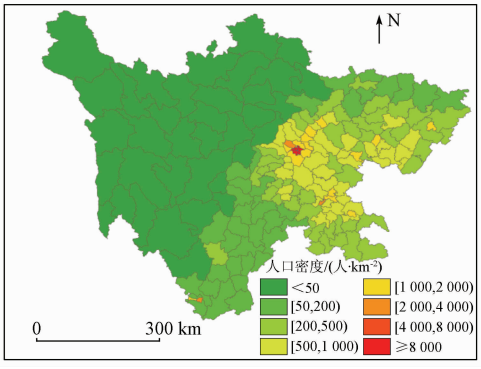
3.2 人口空间化结果

基于土地利用和 DMSP/OLS 数据,利用 S-GWR 模型生成了四川省 2010 年的 SDP(图 3(a)),和人口统计数据的人口密度分布图相比较(图 3(b)),两者有相同的人口分布趋势,但是前者更突出了人口分布的细节。为了可以更清晰地看到两者的区别,提取了成都市部分区县 SDP(图 3(c)),并与县级统计数据人口密度图进行对比(图 3(d)),可以看出人口空间分布情况大致相同,但是 SDP 可以提供更小的像元人口密度,将人口分配到了更细致的空间尺度上,更符合实际人口的分布情况。人口主要集中在居民地和城镇建设用地上,各区县的人口密度高值区主要集中在县城所在地,同时,人口空间分布图显示的中心城区与周边城区人口密度变化更加自然,印证了当代中国人口分布的实际情况。而稀疏零散的农村人口则被分配到耕地上,大多是无光或者光值很低的农村地区。当与夜间灯光数据(图 1(a))比较时,灯光越亮的地方,人口密度越高,人口密度低的地方,灯光亮度也相应较低。因

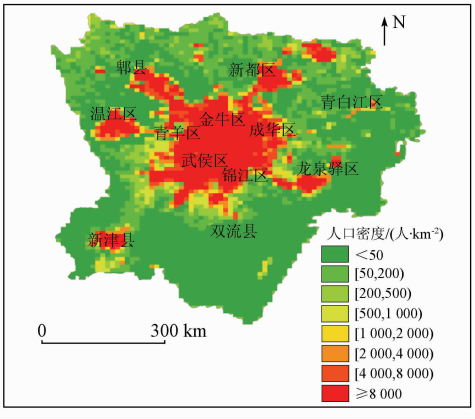
此,利用 S－GWR 模型来生成人口空间分布图在很大程度上符合人口实际分布。



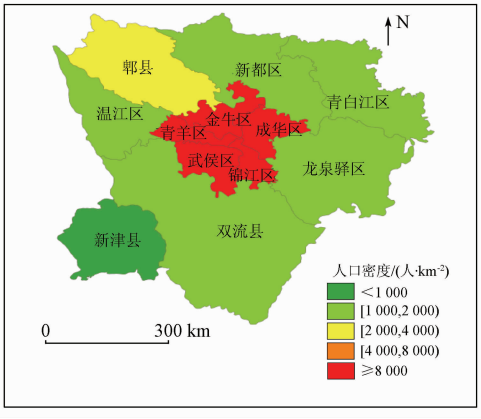
(a) S－GWR 生成 SDP



(b) 区县统计人口密度图



(c) 成都市部分区县 SDP



(d) 成都市部分区县统计人口密度图

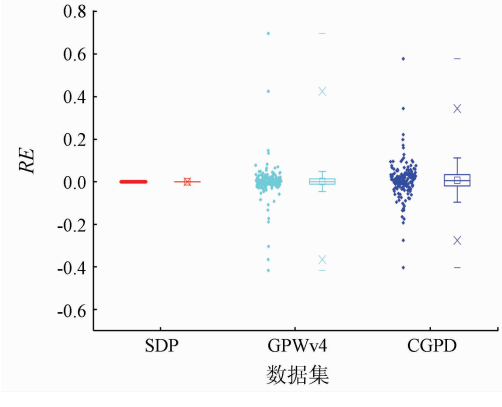
图 3　2010 年人口空间分布密度图

Fig. 3　Spatial distribution of population in 2010

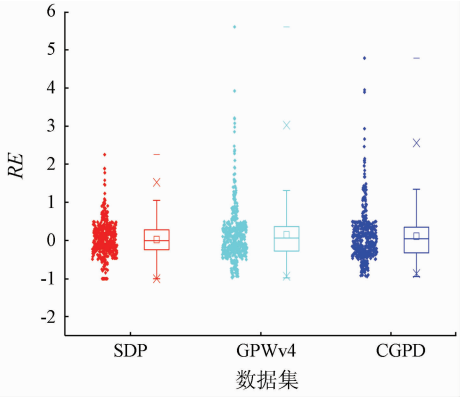
3.3　分级验证评估

精度评估是人口空间化研究的重点也是难点,基于前人的经验和方法,本研究 2010 年世界格网人口第四版 GPWv4 和中国格网人口分布数据 CGPD,分别在县乡两级进行对比验证。此外,县乡人口统计数据默认为真实人口数据。

在县一级,分别计算了 3 种数据结果在研究区内的所有区县的 *RE*。为了揭露误差的细节和总体情况,将 3 种数据的相对误差用箱线图表示出来(图 4(a)),图中散点代表每个区县的相对误差值,两端的短横线代表最大值和最小值,而 1% ~99% 之间的误差显示在交叉线中。可以看出,GPWv4 的 *RE* 最大是 0.7,最小是 -0.42,*ME* 为 1%;CGPD 的 *RE* 最大是 0.58,最小是 -0.4,CGPD 的 *ME* 为 7%。



(a) 县级 *RE* 散点箱线图



(b) 500 乡镇 *RE* 散点箱线图

图 4　3 种数据集 RE 箱线图

Fig. 4　Relative error box chart of three kinds of datasets

而由于 SDP 人口经过式(3)的系数调整,其 *RE* 和 *ME* 都接近于0。另外2种数据集对区县不同程度的高估或低估,可能是由于这些县的人口密度与其他县的人口密度不一致,影响人口分布的因素不一样,不能很好地从基于回归模型中得出。上述3种数据都分别经过不同方法的调整,但通过上述分析可以知道,通过县级人口统计数据来调整 SDP 是有必要的。在乡镇一级,根据随机数的生成,随机选取500个乡镇进行精度评价。将500个乡镇的人口统计数据视为真实人口值,分别计算估计人口与统计人口之间的 *RE*,并分级统计分析,再分别计算整体的 *MAE*,*MRE*,*RMSE*。

表5统计了3种数据集的误差指标,可以看出 SDP 的3种误差均小于其他两种数据集,GPWv4 和 CGPD 的 *MRE* 分别为47.48%和45.43%,而用 S-GWR 得到的 SDP 仅为34.54%;在 *MAE* 方面,GPWv4 和 CGPD 分别为7 997.774人和7 256.342人,而 SDP 为5 715.703人;RMSE 可以反映预测结果与实际数据的偏差,GPWv4 和 CGPD 分别为18 846.285人和16 997.919人,两者有相似的离散度,而且均高于 SDP 的12 085.932人。由此可以看出,SDP 比其他两种数据得到的结果更好,精度更高,说明 SDP 预测人口更接近于人口普查数据,具有更高的可信度。

表5 3种数据集精度对比

Tab.5 Accuracy comparison of three datasets

误差指标	SDP	GPWv4	CGPD
MAE/人	5 715.703	7 997.774	7 256.342
MRE/%	34.54	47.48	45.43
RMSE/人	12 085.932	18 846.285	16 997.919

为了可以直观地看出3种数据的在局部乡镇上的差异和细节,同样将乡镇误差显示在箱线图中(图4(b))。可以看出,GPWv4 的相对误差最大是5.61,最小是-0.97,CGPD 的相对误差最大是4.79,最小是-0.94,SDP 的相对误差最大是2.26,最小是-0.88。异常值分布在高值区域,低值区域无较大差别,且大多都是由于对人口的高估所导致,说明 GPWv4 和 CGPD 这2种全球性数据集不适合在局部进行回归,而 SDP 由于考虑了回归变量的非平稳性,在局部获得了较好的结果。SDP 比另外2种数据的散点分布更集聚一些,其相对误差更集中在0附近,与真实人口数据比较接近。

为了得到3种数据结果的误差结构,将500个乡镇进行分级统计(表6),根据 RE 值分成5个范围,分别是严重低估(≤-50%)、一般低估((-50%,-20%])、准确估计((-20%,20%])、一般高估((20%,50%])、严重高估(>50%)。图5显示了500个乡镇 RE 各级别的相对占比情况。

表6 500个乡镇相对误差分级统计表

Tab.6 Statistical table of relative error classification in 500 villages and towns (个)

相对误差分级	SDP	GPWv4	CGPD
严重低估	48	51	56
一般低估	97	101	114
准确估计	185	151	158
一般高估	107	107	97
严重高估	63	90	75

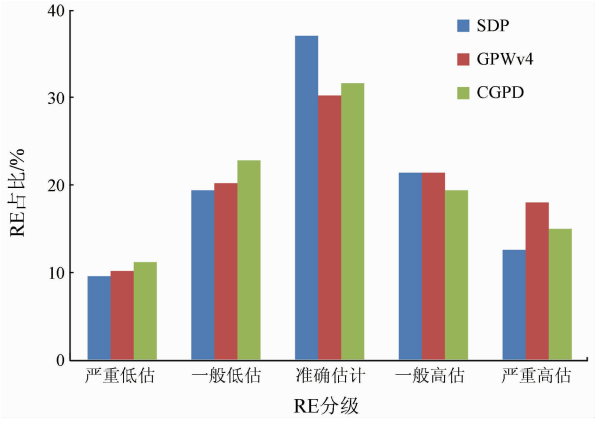


图5 500乡镇 RE 占比统计图

Fig.5 Relative error ratio of villages and towns

SDP 的乡镇误差分级统计个数分别是48,97,185,107和63个,误差占比为9.6%,19.4%,37%,21.4%和12.6%;GPWv4 分别为51,101,151,107和90个,误差占比为10.2%,20.2%,30.2%,21.4%和18%;CGPD 分别为56,114,158,97和75个,误差占比为11.2%,22.8%,31.6%,19.4%和15%。可以看出,3种结果均存在不同程度的高估,而人口高估的乡镇大多位于青藏高原东部和邛崃山脉以西的川西高原。此外,分析出现明显高估和明显低估的原因可能与该地区的气候、海拔等其他影响人类分布的因素有关。在3种数据结果中,SDP 准确估计的乡镇最多,多达185个,占比达到了37%,出现低估和高估的乡镇个数比另外两个数据集要少,而且分布更为分散。因此,考虑参数的空间平稳性可以较好的提高人口空间化的精度和减少对乡镇人口的高估。

4 结论

1)Pearson 相关检验结果显示了土地利用类型与人口分布之间的相关性。研究选取了与人口显著正相关的土地利用类型作为模型变量,根据建模结果表明,考虑人口分布建模的时候不应该只考虑与人口正相关的土地类型,其他土地类型林地、草地甚

至水域都可能有人口分布。

2)该模型与传统的全局模型和局部模型相比,其考虑了空间变量的平稳性,将全局变量和局部变量混合起来,通过局部变量在不同空间地理位置上的系数不同来提高人口空间化精度。基于 GWR4.0 软件得出 3 种模型拟合优度,结果表明, S - GWR 模型的拟合效果最优,决定系数 R^2 和 *AICc* 值分别为 0.903 和 4 786.263,较其他 2 个传统模型均有明显提升,进一步提高了对人口空间化的解释力。

3)本研究对 SDP 进行了分级精度评估。在县一级,GPWv4 和 CGPD 这 2 种数据集的 ME 分别为 1% 和 7%,而由于人口系数的调整,SDP 的 ME 接近于 0。在乡镇一级,随机生成的 500 个乡镇中,与 GPWv4 和 CGPD 相比,SDP 准确估计的乡镇个数最多,达 37%,极端乡镇(严重低估和严重高估)数量较少,低估和高估乡镇个数都分别比另外 2 个数据集要少。在 RE 方面,SDP 的 RE 最大是 2.26,最小是 -0.88,比另外 2 种数据集的范围要小; 在 MAE 方面,SDP, GPWv4 和 CGPD 的误差分别为 5 715.703 人,7 997.774 人和 7 256.342 人; 在 MRE 方面, SDP, GPWv4 和 CGPD 的误差分别为 34.54%, 47.48% 和 45.43%; 在 RMSE 方面,SDP, GPWv4 和 CGPD 的误差分别为 12 085.932 人,18 846.285 人和 16 997.919 人。总的来说,SDP 在人口预测方面比另外 2 种数据表现得更好,证明了 S - GWR 模型生成的 SPD 在准确重新分配人口方面优于其他数据集。

本研究使用 S - GWR 模型方法,可用于在区域尺度上产生地理空间细节不同的网格人口,其人口估计结果比传统模型精度更高、效果更好,对生态学、灾害评价等相关研究具有重要意义。但夜间光照和土地利用数据在全球范围内都是免费提供的,因此更适合缺乏详细数据的大规模人口空间化。因此,在未来可以利用更高分辨率和更高精度的数据进行研究,也可以从影响人口分布因素方面以进一步提高人口空间化的精度。

参考文献(References):

[1] 谭 敏,刘 凯,柳 林,等. 基于随机森林模型的珠江三角洲 30 m 格网人口空间化[J]. 地理科学进展,2017,36(10):1304 - 1312.
Tan M, Liu K, Liu L, et al. Spatialization of 30 m grid population in Pearl River Delta based on stochastic forest model[J]. Progress in Geography, 2017, 36(10): 1304 - 1312.

[2] 柏中强,王卷乐,杨 飞. 人口数据空间化研究综述[J]. 地理科学进展,2013,32(11):1692 - 1702.
Bai Z Q, Wang J L, Yang F. Research progress in spatialization of populationdata[J]. Progress in Geography, 2013, 32(11): 1692 - 1702.

[3] 肖东升,杨 松. 基于夜间灯光数据的人口空间分布研究综述[J]. 国土资源遥感,2019,31(3):10 - 19. doi:10.6046/gtzyyg.2019.03.02.
Xiao D S, Yang S. A survey of population spatial distribution based on night light data[J]. Remote Sensing for Land and Resources, 2019, 31(3): 10 - 19. doi: 10.6046/gtzyyg.2019.03.02.

[4] Elvidge C D, Baugh K E, Dietz J B, et al. Radiance calibration of DMSP - OLS low - light imaging data of human settlements[J]. Remote Sensing Enviroment, 1999, 68, 77 - 88.

[5] Zhang Q L, Seto K C. Mapping urbanization dynamics at regional and global scales using multi - temporal DMSP/OLS nighttime light data[J]. Remote Sensing of Environment, 2011, 115(9): 2320 - 2329.

[6] 陈 晴,侯西勇,吴 莉. 基于土地利用数据和夜间灯光数据的人口空间化模型对比分析——以黄河三角洲高效生态经济区为例[J]. 人文地理,2014,29(5):94 - 100.
Chen Q, Hou X Y, Wu L. Comparison of population spatialization models based on land use data and DMSP/OLS data respectively: A case study in the efficient ecological economic zone of the Yellow River Delta[J]. Human Geography, 2014, 29(5): 94 - 100.

[7] 杨续超,高大伟,丁明军,等. 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,2013,22(6):729 - 734.
Yang X C, Gao D W, Ding M J, et al. Spatialization of population statistics based on multi - source remote sensing data and DEM: A case study of Zhejiang Province[J]. Resources and Environment in the Yangtze Basin, 2013, 22(6): 729 - 734.

[8] 赵利利,孟 芬,马才学. 基于多源遥感数据的武汉市人口空间分布格局演化[J]. 地域研究与开发,2016,35(3):165 - 169.
Zhao L L, Meng F, Ma C X. Spatial distribution pattern evolution of Wuhan population based on multi - source remote sensing data[J]. Areal Research and Development, 2016, 35(3): 165 - 169.

[9] 胡云锋,赵冠华,张千力. 基于夜间灯光与 LUC 数据的川渝地区人口空间化研究[J]. 地球信息科学学报,2018,20(1):68 - 78.
Hu Y F, Zhao G H, Zhang Q L. Population spatialization in Sichuan and Chongqing based on night lighting and LUC data[J]. Journal of Geo - Information Science, 2018, 20(1): 68 - 78.

[10] 黄 杰,闫庆武,刘永伟. 基于 DMSP/OLS 与土地利用的江苏省人口数据空间化研究[J]. 长江流域资源与环境,2015,24(5):735 - 741.
Huang J, Yan Q W, Liu Y W. Spatial analysis of population data in Jiangsu Province based on DMSP/OLS and land use[J]. Resources and Environment in the Yangtze Basin, 2015, 24(5): 735 - 741.

[11] 丁文秀,赵 伟,左德霖,等. 基于土地利用分类模型和重力模型耦合的人口分布模拟——以武汉市人口数据为例[J]. 大地测量与地球动力学,2011,31(s1):127 - 131.
Ding W X, Zhao W, Zuo D L, et al. Population distribution simulation based on coupling of land use classification model and gravity model: A case study of Wuhan population data[J]. Geodesy and Geodynamics, 2011, 31(s1): 127 - 131.

[12] Fotheringham A S, Brunsdon C. Local forms of spatial analysis[J]. Geographical Analysis, 2010, 31, 340 - 358.

[13] 王珂靖,蔡红艳,杨小唤. 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,2016,35(12):1494 - 1505.

Wang K J, Cai H Y, Yang X H. Application of multivariate statistical regression and geographically weighted regression in the study of multi - scale population spatialization[J]. Progress in Geography, 2016, 35(12): 1494 - 1505.

[14] 张建辰,王艳慧. 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, 2014, 16(3): 435 - 442.
Zhang J C, Wang Y H. Simulation of rural population spatial distribution based on land use classification: A case study of Hefeng County, Hubei Province[J]. Journal of Geo - Information Science, 2014, 16(3): 435 - 442.

[15] 陈 晴,侯西勇. 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, 2015, 17(11): 1370 - 1377.
Chen Q, Hou X Y. Integrating land use data and night light data to optimize population spatialization model[J]. Journal of Geo - Information Science, 2015, 17(11): 1370 - 1377.

[16] 王明明,王卷乐. 基于夜间灯光与土地利用数据的山东省乡镇级人口数据空间化[J]. 地球信息科学学报, 2019, 21(5): 699 - 709.
Wang M M, Wang J L. Spatialization of township population data in Shandong Province based on night lighting and land use data[J]. Journal of Geo - Information Science, 2019, 21(5): 699 - 709.

[17] Dong N, Yang X H, Cai H Y. Research progress and perspective on the spatialization of population data[J]. Journal of Geo - Information Science, 2016, 18: 1295 - 1304.

[18] 四川统计局. 四川统计年鉴[M]. 北京: 中国统计出版社, 2010.
Statistical Bureau of Sichuan Province. Sichuan statistical yearbook[M]. Beijing: China Statistics Press, 2010.

[19] 杨继瑞,李月起,汪 锐. 川渝地区:"一带一路"和长江经济带的战略支点[J]. 经济体制改革, 2015(4): 58 - 64.
Yang J R, Li Y Q, Wang R. Sichuan and Chongqing region: Strategic fulcrum of the Belt and Road initiatives and Yangtze River economic zone[J]. Reform of Economic System, 2015(4): 58 - 64.

[20] 刘纪远,宁 佳,匡文慧,等. 2010—2015 年中国土地利用变化的时空格局与新特征[J]. 地理学报, 2018, 73(5): 789 - 802.
Liu J Y, Ning J, Kuang W H, et al. Spatial and temporal patterns and new characteristics of land use change in China from 2010 to 2015[J]. Acta Geographica Sinica, 2018, 73(5): 789 - 802.

Population spatialization based on geographically weighted regression model considering spatial stability of parameters

XIAO Dongsheng[1,2,3], LIAN Hong[1,2]

(1. *School of Civil Engineering and Surveying and Mapping, Southwest Petroleum University, Chengdu* 610500, *China*; 2. *Disaster Prevention and Emergency Research Center of Mapping and Remote Sensing Geographic Information of Southwest Petroleum University, Chengdu* 610500, *China*; 3. *Public Security and Emergency Research Institute, Sichuan Normal University, Chengdu* 610068, *China*)

Abstract: The theories on population spatialization tend to be mature in recent years. However, the spatial stability of the variables and parameters used in population spatialization modeling has been scarcely focused on. With the land use data, night - time light data, and demographic data as the data sources, this study proposed a novel precise population spatialization method based on a semi - parametric geographically weighted regression model (S - GWR). Then a permanent population spatialization model on a county scale was built using the method proposed in this study and then was verified using the Sichuan Province as the study area. In this study, the spatial stability of parameters and variables were obtained using the S - GWR model while the characteristics of the variables were analyzed, in order to improve the accuracy of population estimation. Finally, the population spatial distribution map (SDP) with a resolution of 1 km of Sichuan Province in 2010 was formed. The results show that the coefficient of determination coefficient of the S - GWR model was 0.903, which is higher than that of traditional regression models and indicates better fitting effects. The S - GWR model was verified using two commonly used population datasets, and the verification results are as follows. At a county level, the overall average error of the study area and the relative error of each district and county in the SDP all approximated to 0, and thus the SDP was more precise than the other two datasets. At a township level, the mean relative error, mean absolute error, and root mean square error of SDP were 34.54%, 5 715.703, and 12 085.932, respectively, which were all lower than those of the other two datasets. Meanwhile, the SDP showed more favorable dispersion effects than the other datasets. Furthermore, the number of the towns whose population was accurately estimated was 185 in the SDP, which was higher than that in the other two datasets. Therefore, the accuracy of population spatialization can be improved by considering the spatial stability of parameters.

Keywords: semi - parametric geographically weighted regression; spatial stability; night - time light data; land use; population spatialization

(责任编辑: 张 仙)