文章编号:1671-4229(2024)04-0032-14

基于多源地理大数据我国县域人口估算方法研究

徐 勇,郑从威
(广州大学 地理科学与遥感学院,广东 广州 510006)

摘 要:实时人口数据对于城市规划、资源管理和社会的可持续发展等方面至关重要。为了有效提升现有基于地理大数据的人口估算方法,研究全面对比分析不同开放地理数据的人口模拟性能,并发展综合遥感与新兴社交媒体用户数据,以实现区县级人口的高精度快速估算。文章以中国各区县为试验区,运用多元线性回归及地理加权回归方法,全面评估各类地理遥感数据对我国人口模拟的性能,采用的数据包括腾讯位置服务(LBS)数据、高德兴趣点数据(POI)、夜间灯光遥感数据和基于遥感所得的土地利用/覆盖数据等。研究结果显示,在预估人口分布方面,腾讯定位数据与兴趣点数据比遥感所得的土地利用/覆盖数据和夜间灯光卫星数据都要好,人口模拟精度分别为81.6%、70.8%、68.8%和63.0%。文章进一步综合运用多源地理数据,可实现85.4%总体人口模拟精度,研究结果和发现可为我国人口相关政策提供数据和技术支撑。

关键词:人口;腾讯位置数据;POI数据;土地覆盖数据;夜间灯光
中图分类号:K909 **文献标志码**:A

Population mapping in China with multi-sourced geographical open data

XU Yong, ZHENG Cong-wei
(School of Geography and Remote Sensing, Guangzhou University, Guangzhou 510006, China)

Abstract: Real-time population data is crucial for urban planning, resource management, and the sustainable development of society. In order to effectively enhance existing population estimation methods based on geospatial big data, this study comprehensively compares and analyzes the population simulation performance of different open geospatial datasets, and develops a comprehensive approach integrating remote sensing and emerging social media user data to achieve high-precision rapid estimation of population at the county level. Taking Chinese counties as the experimental area, multiple linear regression and geographically weighted regression methods are employed to comprehensively evaluate the population modeling capability of various geospatial remote sensing data. The data utilized include Tencent Location-Based Service (LBS) data, Amap Point-of-Interest (POI) data, nighttime light remote sensing data, and land use/cover data derived from remote sensing. The research findings indicate that, in estimating population distribution, Tencent location data and POI data outperform remotely sensed land use/cover data and nighttime light satellite data, with population simulation accuracies of 81.6%, 70.8%, 68.8%, and 63.0%, respectively. Furthermore, the comprehensive use of multi-source geospatial data can achieve an overall population simulation accuracy of 85.4%. The research results and discoveries can provide data and technical support for population-related policies in China.

Key words: population; Tencent's social user location data; POI data; land cover data; nighttime light

收稿日期:2024-04-25;**修回日期**:2024-05-26
基金项目:国家自然科学基金面上资助项目(42271477);广东省哲学社会科学"十三五"规划资助项目(GD20CYJ40)
作者简介:徐勇(1982—),男,教授,博士. E-mail:xu1129@gzhu.edu.cn
引文格式:徐勇,郑从威. 基于多源地理大数据我国县域人口估算方法研究[J]. 广州大学学报(自然科学版),2024,23(4):32-45.

人口统计数据显示在特定时间和空间上的人口聚集和分布情况。最新的人口数据对于城市规划、环境保护和资源配置等方面至关重要。但人口统计数据传统上是通过实地调查获得的,且普查时间间隔较长。以中国人口普查数据为例,每 10 年进行一次全国人口普查,每 5 年进行一次 1% 的抽样调查。因此,其余年份的人口普查数据是根据调查年份的人口普查数据进行外推和平滑操作获得的[1]。由于获取真实人口数据所需的时间间隔较长,还要经过一段时间的数据汇总、处理和评估,且存在部分偏远地区数据缺失等问题,最快也要一两年后才能发布和开发利用,难以确定当前的人口规模及其空间分布。对最新人口数据的准确评估有利于实施前瞻性的城市规划和资源管理,特别是对中国这样的发展中国家而言,人口分布不均和流动性大会导致相当大的公共安全和资源约束问题[2],因此,通过其他方式获得实时、准确的人口数据是很有必要的。

随着遥感(RS)和地理信息系统(GIS)技术等地理空间技术的发展,使得非接触式遥感数据以其快速、广泛的覆盖范围在人口估算研究中得到普遍的应用[3-5]。目前用于人口估算研究的遥感数据主要有两类:夜光卫星数据和常规光学卫星数据。夜间灯光数据将人类社会经济活动的丰富程度表现为夜间灯光强度,由此可以推断人口信息[6]。一些广泛使用的夜光卫星数据集包括美国国家航空航天局(NASA)的国家极轨合作伙伴可见光红外成像辐射计套装(NPP/VIIRS)及其前身国防气象计划业务线扫描系统(DMSP/OLS)和中国的珞珈 1 号 夜光卫星数据[7]。DMSP/OLS 夜间灯光数据由于其空间分辨率较粗,辐射分辨率受限,主要被用于探究国家层面而非城市层面的人口规模。Elvidge 等[8]研究利用 DMSP/OLS 夜间灯光数据探测城市内部的人口差异,发现数据的过饱和亮度值难以表达城市中心内部人口的空间差异。Sutton 等[9]使用同样的数据来估计世界人口的分布,发现根据人口和发光强度之间的统计关系,可以将国家经济收入水平分为高、低、中 3 类。

夜光遥感技术的进步使得目前被广泛使用的 NPP/VIIRS 夜间灯光数据比其前身 DMSP/OLS 具有更高的空间和辐射分辨率。Chen 等[10]基于非洲城市的实验证实,NPP/VIIRS 夜间灯光数据比 DMSP/OLS 夜间灯光数据能够实现更精确的人口估计。Wang 等[4]利用 NPP/VIIRS 灯光数据和非线性模型探讨中国部分城市夜间灯光数据与人口之间的关系,发现总夜间灯光强度指标的表现远远优于其他指标,如灯光覆盖面积或平均夜间灯光强度。Zeng 等[11]比较夜间灯光和土地覆盖数据对人口的分解,发现土地利用/覆盖数据的表现略优于夜间灯光卫星数据。Li 等[12]利用 NPP/VIIRS 夜间灯光数据模拟北京市人口的空间分布,发现参照人口普查数据计算,利用夜间灯光数据实现的总体估算精度约为 62%。Guo 等[13]利用长时间序列的夜间灯光数据,结合 DMSP/OLS 和 NPP/VIIRS 数据的优缺点来估算中国人口密度。

除夜光卫星数据外,中高分辨率的光学卫星数据也常被用于通过监测地面上的人居环境来估算人口分布[14]。与夜光卫星数据相比,土地覆盖数据中的城镇区域,或者不透水面等人类的物理住区,如住宅和建筑物,是用于估计人口分布的最重要特征[15-16]。也有的研究通过排除土地覆盖中的无人居住区域进行人口密度估算[17]。Stevens 等[18]以卫星遥感数据为主,开源地理信息数据为辅估算世界范围内的人口分布,也被称为 WorldPop 计划。同样,哥伦比亚大学发起的“世界人口网格化计划”整合遥感和开放地理信息数据,以生成全球范围的人口分布信息,并于 2015 年发布最新版本 GPWv4[19]。此外,Freire 等[20]探索使用中低分辨率卫星数据来获取全球人居环境图层,然后将卫星图像的光谱和纹理信息与全球人居环境图层数据相结合,生成全球人口分布图。

然而,目前基于遥感数据的人口估算方法的精度普遍不高。有研究证实,现有的 4 种基于卫星数据的中国人口估算产品的总体精度均在 60% 左右[12,21],一定程度上限制了其被广泛应用。为了提高遥感估算精度,可以结合其他数据进行人口估算研究,如结合典型的地理空间大数据——兴趣点(Point of Interest,POI)数据。POI 在地理信息系统中泛指各种点实体,如商铺、公交站或房屋等,每个 POI 包含名称、坐标和所属类别等信息,具有广泛覆盖、高准确性和实时性等优点[22-24],可用于反映人类社会活动。可将 POI 转换为不同网格大小的栅格图层,并结合遥感数据使用。不同类别的 POI 代表其内部和周围不同的人类活动,因此,与人口密度具有不同程度的相关性[25-26]。近年来,POI 数据常常与其他数据结合用于精细化的人口分布估算研究[27-30]。如淳锦等[27]利用 POI 数据和土地利用分类数据进行人口分布格网化方法

研究,取得较好的估算结果。赵鑫等[28]基于卫星遥感和POI数据的人口空间化研究中,证实引入能够反映空间微观细节信息的POI数据可以提高人口空间化结果的精度。

与传统的遥感方法相比,除了POI数据之外,还有来自社交媒体(如微信、QQ、Facebook、Twitter等)、众包服务(如美团)和在线地图等移动应用的基于位置服务(Location-based service,LBS)的数据可以从社会角度而非空间中的物理证据来反映城市内部个体的社会活动[31]。Patel等[32]利用Twitter的地理空间信息成功模拟印度尼西亚的人口空间分布。Deville等[33]基于手机数据的空间属性估计人口分布,并指出手机数据在人口稀疏和人口普查数据缺乏的地区更有效。Cheng等[34]利用手机定位数据估算我国2015年的月度人口分布及其变化。还有的研究以位置服务数据为主,其他地理空间数据为辅进行人口估算研究[35-36],如Chen等[37]利用腾讯定位数据和高德POI等地理空间大数据,生成我国100 m分辨率的人口网格,精度高于主要使用遥感数据生成的WorldPop和LandScan人口数据集。这些研究表明,LBS数据在估计城市内部人口分布方面具有很大的潜力。

总的来说,利用夜间灯光和土地利用等遥感数据,以及POI和腾讯LBS等地理空间数据进行人口估算的研究比较丰富,然而较少有研究对比分析这些数据估算人口空间分布的能力大小。另外,分析定位服务数据和POI数据,以及运用遥感数据在估算人口空间分布方面精度的研究也较少。因此,在利用多源地理空间数据进行人口空间分布预估的丰富研究基础上,本研究旨在对各种数据在人口空间分布估计方面的性能进行比较分析和评估,包括腾讯LBS数据、POI数据、夜光卫星数据和基于卫星的土地利用/覆盖数据。结合以上的遥感卫星数据和地理空间数据,评估各种人口空间化模拟方法和技术,实现对中国人口的精准估计,并验证其精度。研究结果和定量模拟工具可为缺乏人口数据的国家和地区的长远规划提供重要的人口数据和技术支持[38]。

1 试验区域与数据

1.1 试验区域

中国幅员辽阔,地形复杂多样,拥有56个民族,14多亿的广大人民生活在各个行政单位内部,人口分布不均。因此,粗略的行政单元(如省级和地市级行政单元)可能无法充分描述中国复杂的人口空间分布格局,需要选择更为精细的行政单元作为研究对象。因此,如图1,本研究选择了中国2 847个区县(除香港、台湾等)作为研究单元,分析多个来源的数据集,以估算2020年中国的县域人口。

图1显示基于国家统计局第七次人口普查数据的中国县域人口的空间分布,红色表示人口较多,蓝色表示人口较少。人口地图显示,2020年,胡焕庸线作为中国最重要的人口分界线仍基本保持稳定[39],中国东部地区尤其是东南沿海地区人口密度较大。人口最密集的地区包括珠江三角洲南部地区和长江三角洲东部地区。从人口数据来看,人口排名前五的省份分别是广东、山东、河南、江苏和四川,占比分别为8.93%、7.19%、7.04%、6.00%和5.93%,以上各省人口均在8 000万以上,广东和山东人口超过1亿。近30年来,中国人口从1990年的11.5亿增加到2020年的14.1亿,年均增长率约为0.4%。同时,近年来我国正在经历快速的城市化,城市人口不断增加,城乡人口流动频繁,因此,快速、准确地评估人口分布状况,有利于准确评估人地关系,实施适宜的空间规划和环境保护战略,实现社会的可持续发展。

1.2 试验数据与处理

收集的数据包括人口普查数据、中国区县行政区域矢量数据、腾讯定位记录数据、高德兴趣点POI数据、卫星获取的土地覆盖和夜间灯光影像,除腾讯定位记录数据收集时间为2018年和中国区县行政区域矢量数据收集时间为2019年外,其余所有数据收集时间均为2020年。

人口数据来源于第七次全国人口普查的县域统计数据。由于腾讯是中国最大的互联网服务提供商之一,本研究收集了腾讯活跃用户的基于位置的服务记录,并根据LBS记录的空间位置将其投影在地图上(如图2(a)所示)。高德兴趣点数据来源于高德开放平台,收集了23种POI大类,需要将带有经纬度信息的点,通过核密度分析投影到地图上(如图2(b)所示)。土地利用/覆盖数据是来自欧洲航天局利用卫星数据提供的标准化全球土地利用/覆盖产品。由于原始的土地利用/覆盖产品覆盖整个世界,因此,使用中国的边界数据来绘制其土地覆盖图,并获取土地利用/覆盖中的城镇区域总和作为研究数据(见图2(c))。夜

间卫星数据来自 NASA 提供的年平均 NPP/VIIRS 夜间灯光数据产品,其中,云的影响已消除,并获取县域夜间灯光强度总和作为研究数据(如图 2(d)所示)。为保证各种数据集具有一致性和可比性,将所有数据集投影到具有相同空间分辨率的统一地理坐标系中。因此,这张数据地图(图 2)可以反映中国人口的空间分布,具有在线临场感。

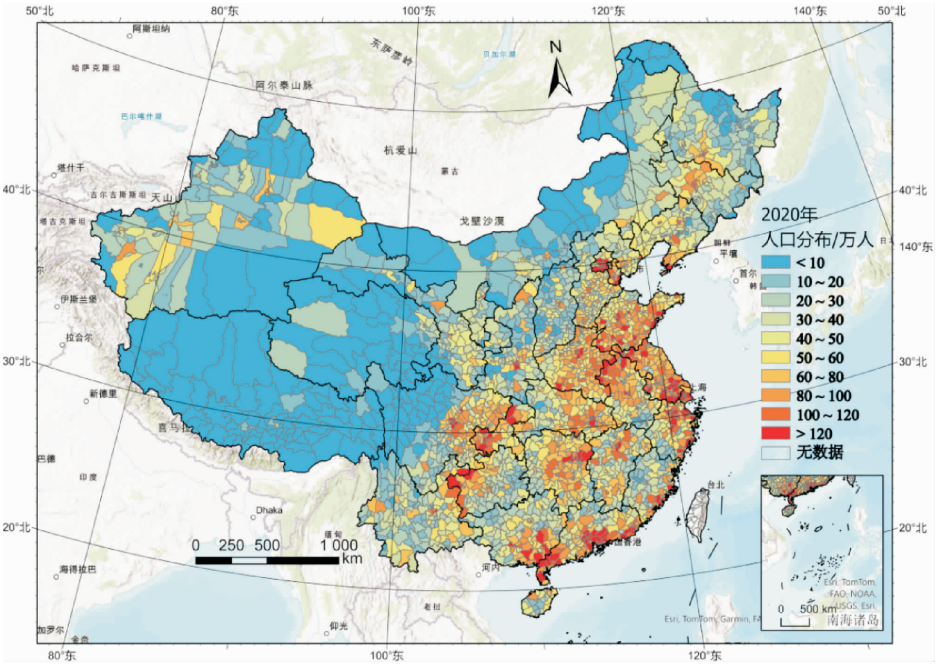

图 1　试验区域及区县人口统计数据

Fig. 1　Study area and relevant demographic data

注:基于自然资源部标准地图服务网站 GS(2019)1822 号标准地图制作,底图边界无修改。

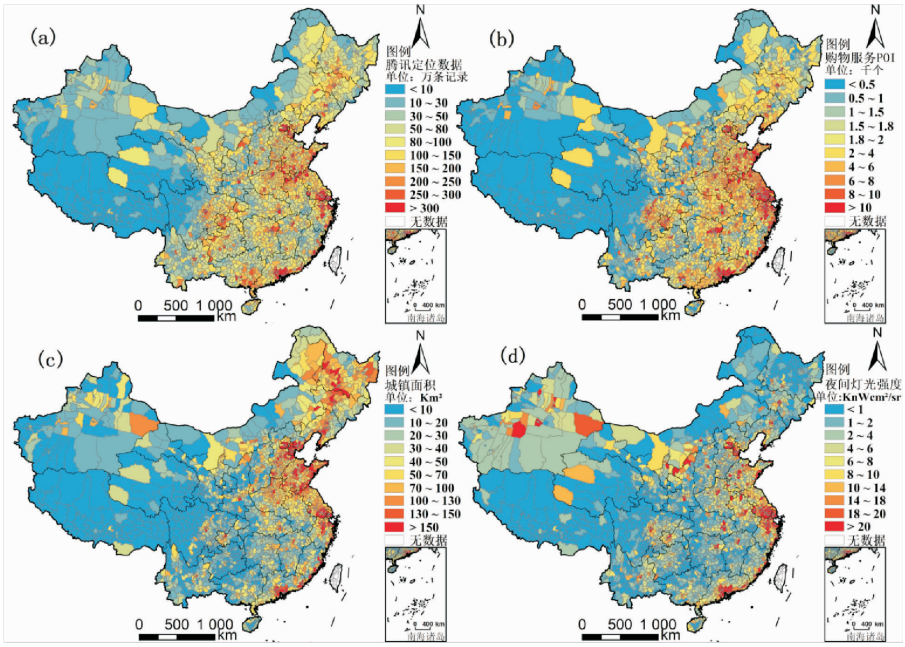

图 2　收集的数据集

Fig. 2　Collected datasets

注:1. 图中,(a)为腾讯定位服务县域记录;(b)为县级购物服务 POI;(c)为卫星遥感获取的县级城镇建设面积;(d)为县级 NPP/VIIRS 夜间灯光总强度。

2. 基于自然资源部标准地图服务网站 GS(2019)1822 号标准地图制作,底图边界无修改。

在本研究中,县级的人口普查数据被用作建立人口模型和评估精度的先验信息,而获得的遥感数据和地理空间数据被用作预测因子来模拟每个县的人口。人口普查数据是基于行政边界的矢量数据,不同于卫星传感器基于栅格的指标。为保证人口普查数据与各项指标保持一致,采用以面积为基础的统计方法,对各县各项指标之和进行加总。考虑到中国有2 874个县,如果在本研究中使用4个不同的指标(地表覆盖数据、夜间灯光数据、POI数据和LBS数据),则预测器大小可能是2 847×4。此外,还可以对人口与数据指标之间的关系进行建模。

1.2.1 腾讯位置服务(LBS)大数据

腾讯位置服务大数据展现的是某点及其周围一定范围内所有使用微信、QQ(社交讯息软件)、在线地图、京东和美团等生活应用软件的定位数量。实时数据应用程序可以在腾讯云(https://cloud.tencent.com/solution/lbs)访问,由于2020年后腾讯位置服务大数据的接口限制,其空间分辨率约为5 km,无法满足研究需要。因此,本研究选取2018年的腾讯用户位置服务大数据,获得自https://doi.org/10.6084/m9.figshare.20400717.v144。该数据集运用爬虫技术以0.01°的空间分辨率和GCS WGS84空间参考坐标系,每5 min获取一次,剔除春节和寒暑假等人口流动较大的情况,收集了2018年3—6月和9—12月,约8亿的在线用户位置信号数据[40]。经过相关性验证,证明本数据的质量优于2020年5 km空间分辨率的腾讯位置服务数据,因此,选用2018年的LBS数据进行研究。2018年LBS数据的空间分辨率约为1 km,为了使空间分辨率与本研究的其他数据的空间分辨率统一,使用ArcGIS软件将腾讯服务位置大数据重采样定为500 m空间分辨率。

1.2.2 高德兴趣点POI数据

高德地图是中国最多人用的在线地图之一,高德POI数据来自高德开放平台,本文获取了2020年全国各区县购物服务、生活服务、政府机构及社会团体和公司企业等23种大类的高德POI数据,主要的类别如表1所示。从表1可以看出,各POI大类与县域人口的相关系数,为了数据间对比分析的相对公平性,只选择与人口相关系数最好的POI类别进行分析。

表1 高德POI数据类型表

Table 1 Amap POI Data Type Table

POI大类	POI中类	点个数/个	相关系数
体育休闲活动	休闲场所、娱乐场所、度假疗养场所、运动场馆、体育休闲服务场所、高尔夫相关、影剧院	748 217	0.851
汽车维修	各品牌汽车维修	444 066	0.840
商务住宅	产业园区、商务住宅相关、住宅区、楼宇	991 405	0.805
政府机构及社会团体	政府机关、政府及社会团体相关、社会团体、工商税务机构、民主党派、公检法机关、外国机构、交通车辆管理	1 623 712	0.852
金融保险服务	银行相关、自动提款机、金融保险服务机构、保险公司、财务公司、证券公司、银行	749 977	0.831
购物服务	花鸟鱼虫市场、便民商店/便利店、家电电子卖场、体育用品店、特色商业街、个人用品/化妆品店、文化用品店、超级市场、特殊买卖场所、服装鞋帽皮具店、购物相关场所、综合市场、商场、家居建材市场、专卖店	12 016 875	0.887
公共设施	紧急避难场所、报刊亭、公共设施、公用电话、公用厕所	314 916	0.725
生活服务	洗浴推拿场所、信息咨询中心、搬家公司、电力营业厅、丧葬设施、售票处、旅行社、电讯营业厅、自来水营业厅、邮局、共享设备、维修站点、人才市场、婴儿服务场所、生活服务场所、彩票彩券销售点、摄影冲印店、物流速递、洗衣店、中介机构、美容美发、各类事务所	5 022 625	0.881

(续表 1)

POI 大类	POI 中类	点个数/个	相关系数
地名地址信息	交通地名、自然地名、门牌信息、普通地名、标志性建筑物、热点地名、市中心	9 522 772	0.763
科教文化服务	科研机构、培训机构、档案馆、学校、美术馆、博物馆、传媒机构、图书馆、展览馆、天文馆、文化宫、科教文化场所、驾校、会展中心、文艺团体、科技馆	1 476 702	0.815
交通服务设施	地铁站、停车场、出租车、上下客区、轻轨站、港口码头、机场相关、公交车站、班车站、长途汽车站、轮渡站、交通服务相关、索道站、火车站、过境口岸	1 689 772	0.814
餐饮服务	中餐厅、冷饮店、糕饼店、甜品店、餐饮相关场所、休闲餐饮相关场所、咖啡厅、快餐厅、茶艺馆、外国餐厅	4 619 791	0.868
汽车服务	汽车租赁、洗车场、汽车养护/装饰、加油站、二手车交易、汽车服务相关、汽车救援、加气站、汽车配件销售、其他能源站、汽车俱乐部、充电站	931 042	0.833
公司企业	公司、工厂、知名企业、农林牧渔基地、公司企业	3 486 136	0.825
住宿服务	旅馆招待所、宾馆酒店、住宿服务相关	766 662	0.724
医疗保健服务	医疗保健服务场所、诊所、急救中心、医药保健销售店、动物医疗场所、专科医院、综合医院、疾病预防机构	1 488 530	0.881
风景名胜	风景名胜相关、公园广场、风景名胜	254 422	0.647
汽车销售	各品牌汽车销售	154 588	0.803

核密度分析的不同搜索半径得到的结果也不同,本研究将搜索半径从 100 m 逐次增加 100 m 直至 1 000 m 以获取最佳核密度搜索半径。当搜索半径在等于或大于 400 m 时,POI 与县域人口的相关性基本稳定且基本涵盖所有 POI 小类别,这也与已有的研究结论相似[41]。经过分析,为了尽量减小共线性的影响,最终将搜索半径确定为 300 m,并以 100 m 为输出像元大小,为了使 POI 数据与其他数据有对比性,以及考虑到对比分析的合理性,最后重采样至统一的空间分辨率 500 m,方便后续使用。

1.2.3　*土地覆盖数据*

本研究所采用的是来源于欧空局(https://maps.elie.ucl.ac.be/CCI/viewer/)2020 年 300 m 空间分辨率的全球土地覆盖/利用遥感监测数据。该数据分为耕地、灌木丛、草原、苔藓地、乔木覆盖区域、树木覆盖区域、裸露区域、水体、城镇区域(Urban areas)和永久冰雪区域等共 37 类(用 0~220 的数值表示,如城市区域用数值 190 表示),其中,有 23 个一级分类,这 23 个一级指标中又有 7 个细分成 3 类。本文城镇建设面积总和提取自土地利用类型中与人口关联性最大的城镇区域,原始数据的空间分辨率为 300 m,且为 netCDF 格式,需转换成 tiff 格式后裁剪出中国区域,并重采样为本研究统一的 500 m 空间分辨率,最后统计城镇建设面积总和作为自变量。

1.2.4　*夜间灯光数据*

本文的夜间灯光数据下载自中国科学院资源环境科学与数据中心(http://www.resdc.cn)处理好的 2020 年度 NPP-VIIRS 数据。NPP-VIIRS 数据——黑色大理石(Black Marble)是美国宇航局(National Aeronautics and Space Administration,NASA)开发的新产品,原始数据(VNP46A4)可从地球数据(https://ladsweb.modaps.eosdis.nasa.gov/)网址下载,进行处理应用。

1.2.5　*人口数据*

本研究使用的人口数据资料是来自国家统计局的 2020 年第七次全国人口普查数据,以及《中国人口普查分县资料—2020》,有行政区划变更的区县级区域则按照实际区域调整人口数据。

2 研究方法

检验方法包括相关分析、逐步回归分析和地理回归分析。相关性分析用于检测人口与一些潜在指标之间的关联。在此基础上,从各种潜在变量中筛选出4个重要指标——LBS记录总量、购物POI数量、城镇建设面积和夜间灯光强度之和,包括但不限于一些可能的指标,如平均夜间灯光强度和其他种类用地的总量。人口与所得4种指标的相关系数分别为0.92、0.89、0.78和0.81。因此,本研究选择LBS记录总和、购物POI数量、城镇建设面积和夜间灯光强度总和这4个指标作为预测指标。

2.1 普通最小二乘法

普通最小二乘法(Ordinary Least Square,OLS)是一种用于空间统计和地理数据建模的统计技术,也是一种最常用于确定自变量与因变量之间线性关系的回归分析方法。本文用普通最小二乘法量化腾讯定位大数据、高德POI、欧空局土地覆盖的城镇面积和夜间灯光等地理大数据和县域常住人口的关系。OLS模型如下:

$$y_i = \beta_0 + \sum_{i=1}^{n} X_i\beta_i + \varepsilon_i, \tag{1}$$

其中,y是因变量,表示第i个区县的人口数量;β_0为模型的截距;X_i对应模型的第i个区县的解释变量;β_i是回归系数,可以反映出每种变量对因变量的影响程度;ε_i为随机误差项。

2.2 地理加权回归分析

地理加权回归(Geographically Weighted Regression,GWR)是一种用于探索空间非平稳性的局部空间统计技术。OLS是假定全局的参数是稳定的理想状态,没有考虑空间的局部差异性,即空间异质性,仅仅是解释变量在所有区县平均意义上的参数估计值。因此,需要引入GWR模型,一种用于空间变化关系建模的线性回归的局部形式,可以反映出各种空间数据与人口之间的局部关系。模型的公式如下:

$$y_i = \beta_{i0}(u_i,v_i) + \sum_{j=1}^{n}\beta_{ij}(u_i,v_i)x_{ij} + \varepsilon_i, \tag{2}$$

其中,y是因变量,表示第i个区县的人口数量;β_{i0}(u_i,v_i)为该区县的截距;x_{ij}对应模型的第i个区县的解释变量;β_{ij}是回归系数;ε_i为随机误差项。

2.3 精度验证和评估

以真实的人口数据作为标准参考值,对各类数据进行精度评价。评价指标包括相关系数(Correlation Coefficient, CC)、决定系数R^2(又称拟合优度)、平均绝对误差(Mean Absolute Error, MAE)、平均相对误差(Mean Relative Error, MRE)和均方根误差(Root Mean Square Error, RMSE)。利用上述5种指标,分别对比分析不同人口数据集的精度,并制作误差空间分布图。各评价指标公式如下:

$$MAE = \frac{1}{N}\sum |f_i - r_i|, \tag{3}$$

$$MRE = \frac{1}{N}\sum \frac{|f_i - r_i|}{r_i} \times 100\%, \tag{4}$$

$$RMSE = \sqrt{\frac{1}{N}\sum (f_i - r_i)^2}, \tag{5}$$

$$CC = \frac{cov(f,r)}{\sigma_f\sigma_r}, \tag{6}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(f_i - r_i)^2}{\sum_{i=1}^{N}(\bar{f} - r_i)^2}, \tag{7}$$

其中,r_i表示第i个区县的真实人口统计数据;f_i表示第i个区县估算所得人口数值;$\bar{f}$表示各区县估算所得人口数值的平均值;N表示区县的个数,$cov(f,r)$表示估算人口数与人口普查数的协方差;σ_f表示估算人口数的标准差;σ_r表示人口普查人口数的标准差。

3 结果与讨论

将中国区县人口与4种数据进行皮尔逊相关系数分析,得出腾讯位置大数据、购物服务POI、城镇建设面积、夜间灯光的相关系数分别为0.918、0.887、0.775和0.807,置信水平均在0.01级别,相关性显著。相关系数通常用于衡量两个变量之间的相关程度,其取值范围为[-1,1],越接近1或-1,则认为两个变量的相关程度越高。本文将单独分析4种数据因子对人口的空间拟合能力。

同时,对4种因子进行逐步回归分析,得出腾讯位置大数据、购物服务POI、城镇建设面积、夜间

灯光的方差膨胀因子(VIF)分别为 11.85、10.48、3.58 和 4.41,VIF 值大于 7.5 表明解释变量(因子)存在冗余,即存在共线性问题,因此,腾讯位置大数据和购物服务 POI 存在共线性问题,除去相关性较低的购物服务 POI,再做逐步回归分析,得出得出腾讯位置大数据、城镇建设面积、夜间灯光的 VIF 值分别为 3.48、3.58 和 4.34,符合条件。最后,综合利用这 3 种因子进行 OLS 和 GWR 回归分析得出最优结果。

3.1　不同数据因子的拟合能力

将 4 种建模因子分别单独进行 OLS 和 GWR 回归分析,利用不同的评价指标探讨其中最优的人口拟合因子。再逐步回归分析选出最优的综合建模因子,然后利用综合因子进行回归分析,得出最优结果。从表 2 可以看出,利用腾讯位置大数据、购物服务 POI、土地利用/覆盖中的城镇面积和夜间灯光强度 4 种数据拟合人口的能力,以及 OLS 模型和 GWR 模型拟合能力的对比。

表 2　不同因子的精度评价

Table 2　Accuracy evaluation of different factors

模型	数据	评价指标				
		MAE/人	MRE	RMSE/人	AICc	R^2
OLS 模型	腾讯位置大数据	130 688	0.700	185 231	77 150	0.843
	购物服务 POI	151 456	0.801	213 555	77 960	0.791
	城镇建设面积	205 060	0.895	295 166	79 813	0.600
	夜间灯光	194 038	0.993	276 066	79 422	0.651
	最优结果	128 161	0.667	179 196	76 965	0.853
GWR 模型	腾讯位置大数据	64 703	0.184	102 727	74 904	0.952
	购物服务 POI	96 570	0.292	147 177	76 959	0.901
	城镇建设面积	101 717	0.312	159 066	77 403	0.885
	夜间灯光	113 288	0.370	167 766	77 689	0.872
	最优结果	49 016	0.146	77 744	73 847	0.972

利用不同评价指标分析 4 种因子的人口拟合能力。从统计学角度来说,通常认为平均绝对误差(MAE)、平均相对误差(MRE)和均方根误差(RMSE)越接近 0 越好,同时从模型优度来看,决定系数(R^2)和阿凯克信息准则(AICc)都是模型拟合度/性能的测量,一个模型的 AICc 比另一个模型小于 3 以上,则认为 AICc 低的模型较好,而 R^2 取值范围为[0,1],越接近 1,说明回归拟合效果越好。因此,从表 2 可以看出,GWR 模型腾讯位置大数据的 MAE 是 64 703 人、MRE 是 0.184、RMSE 是 102 727 人、AICc 是 74 904 和 R^2 是 0.952,均优于其他 3 种因子,说明其对人口的拟合能力最好。其次是购物服务 POI 和城镇建设面积,拟合效果排第二和第三,夜间灯光数据的人口拟合能力相对较弱。同时,从各种评价指标均可以看出,用 GWR 模型拟合人口的效果优于 OLS 模型,从拟合优度 R^2 来看:腾讯位置大数据、购物服务 POI 数据、城镇建设面积数据和夜间灯光数据的优度各提升了 10.9%、11%、28.5% 和 22.1%,综合利用多种数据得出的最优结果提升了 11.9%。

相对误差图可以探究不同数据在不同地区的人口拟合性能。利用 4 种建模因子分别单独进行 OLS 和 GWR 回归分析得到结果后,使用回归估算的人口数和第七次普查的真实人口数经过计算得到相对误差(指利用数据回归估算的人口数与真实人口数所造成的绝对误差与真实人口数之比),并将相对误差划分为 (0,0.2],(0.2,0.4],(0.4~0.6],(0.6~0.8],(0.8, +∞] 5 种精度等级,然后映射到地图上,相对误差越小,证明估算的结果越好,如图 3 所示。左列(图 3(a)、(c)、(e)、(g))是基于 4 种数据并各自使用 OLS 方法得出的县域人口相对误差图;右列(图 3(b)、(d)、(f)、(h))是基于 4 种数据并各自使用 GWR 方法得出的县域人口相对误差图。

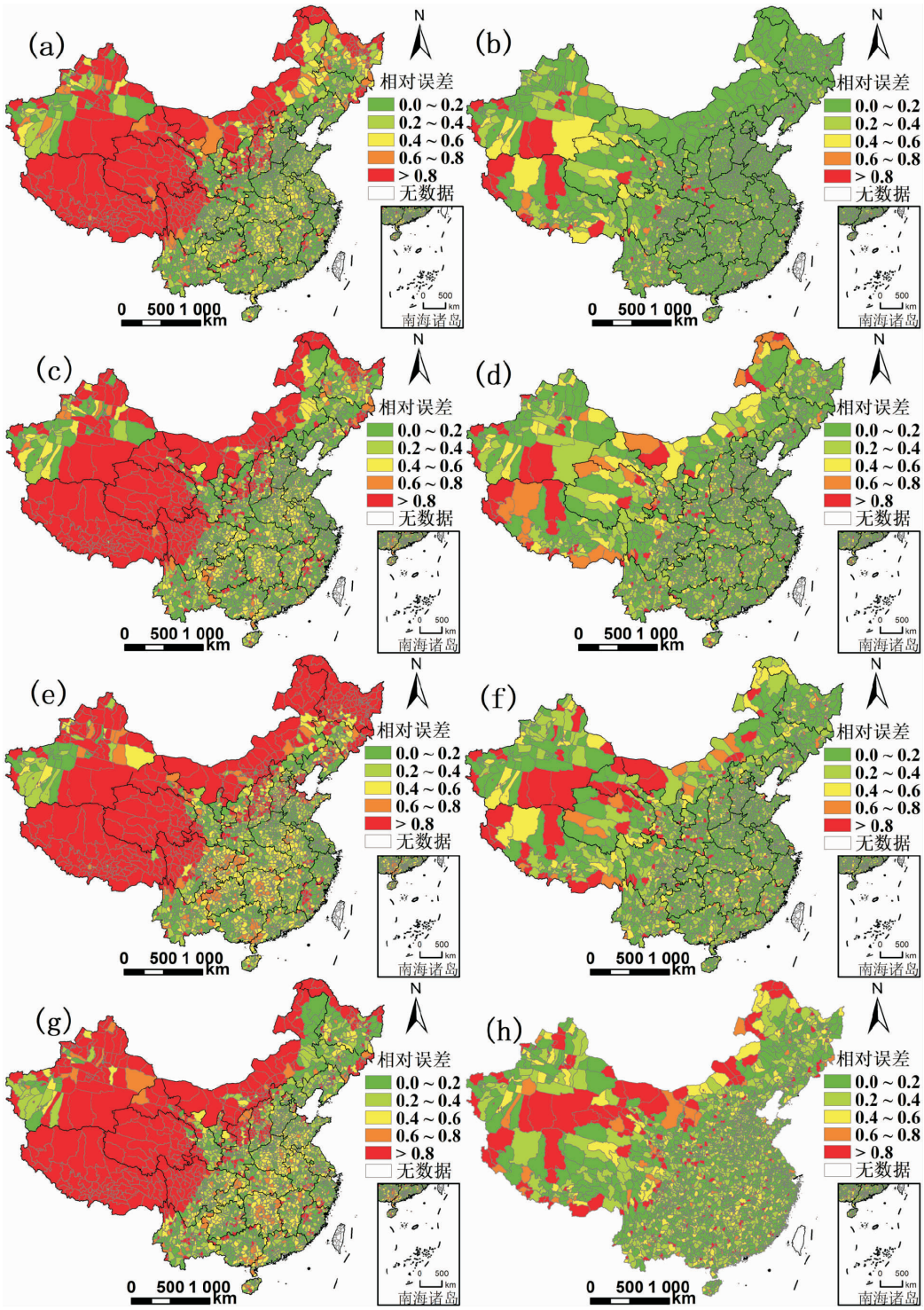

图 3 不同建模因子的拟合相对误差图

Fig. 3 Plot of relative fitting errors for different modeling factors

注:1. 图中,(a) ~ (b)为腾讯位置大数据;(c) ~ (d)为购物服务 POI;(e) ~ (f)为城镇建设面积;(g) ~ (h)为夜间灯光强度。

2. 基于自然资源部标准地图服务网站 GS(2019)1822 号标准地图制作,底图边界无修改。

相对误差图可以从空间视觉效果展现出利用各种数据估算人口数量的局部精度,从图 3 同一列来看,人口数相对误差区域最少的是腾讯位置大数据,并总体由上往下递增,也就是说腾讯位置大数据精度最高,然后按购物服务 POI 数据、城镇建设面积、夜间灯光强度排列,符合表 2 评价指标

中平均相对误差(MRE)的规律。同时,从视觉效果看来,图 3 右列相对误差较小的区域远远多于左边,证明无论哪种数据,使用 GWR 方法的效果都优于使用 OLS 方法,全国大部分区域都有明显的改善,特别是胡焕庸线西部从视觉效果看来是更明显优化的。

3.2　不同建模数据因子的空间分异

探讨不同建模数据因子之间的空间异质性可以为预测人口提供数据源选择参考,而 GWR 具有处理空间差异的能力,GWR 回归系数可以用来反映各数据因子对人口估算(因变量)的影响力,用 Z-Score 标准化方法对各数据的 GWR 回归系数标准化后消除了各数据单位的影响,将不同量级的数据转化为统一量度,进而使不同量级的数据有了可比性。图 4 展示了单个建模数据因子的 GWR 标准化回归系数。

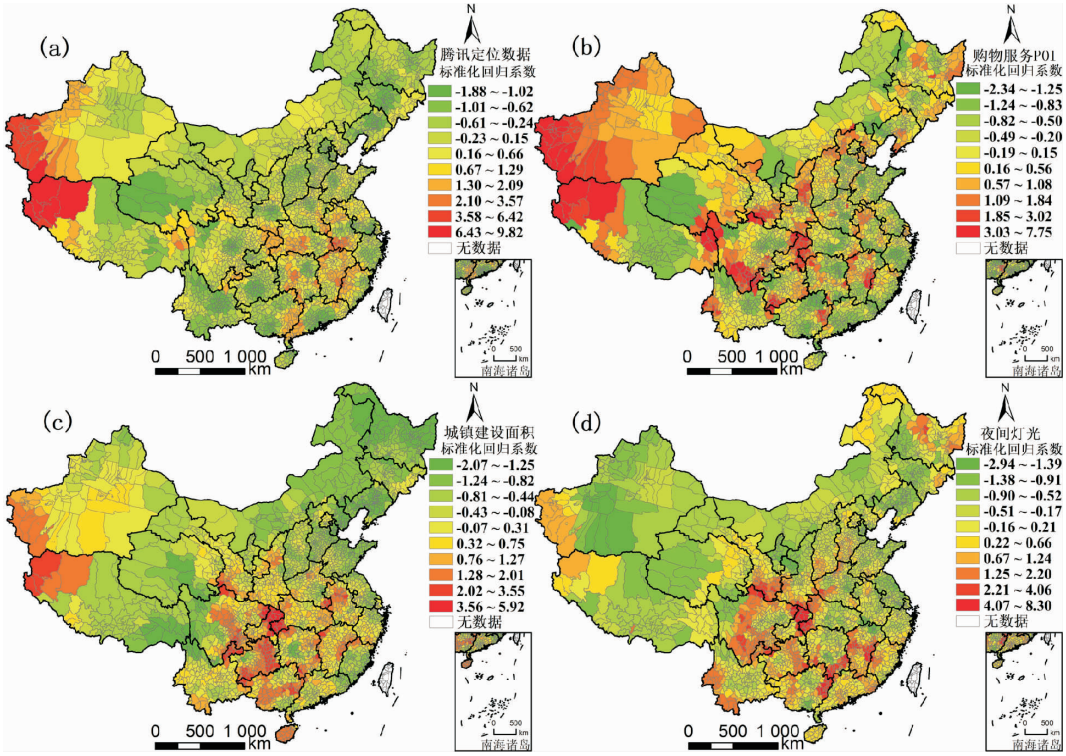

图 4　不同因子的 GWR 标准化回归系数图

Fig. 4　The standard coefficient of GWR model using different indicators

注:1. 图中,(a)为腾讯定位数据;(b)为购物服务 POI;(c)为城镇建设面积;(d)为夜间灯光强度。

2. 基于自然资源部标准地图服务网站 GS(2019)1822 号标准地图制作,底图边界无修改。

利用 GWR 方法的标准化回归系数可以反映数据的局部效果,标准化回归系数越大,该数据对人口数的影响力就越大。如图 4 所示,绿色的区域(回归系数较小)代表此类数据比较能反映此区域的人口,对人口估计的效果较好,相反,越红的区域(回归系数较大)则代表此类数据反映人口的能力较差,对人口估计的影响会较大。结果表明,在中国中部省份交界和新疆、西藏西部,几乎所有得到的回归系数都趋向于较大的回归系数(红色部分),意味着这些因素可能对上述地区的人口估计有较大的影响。如图 4(a)所示,腾讯定位数据在大部分区域表现优良,除了在新疆、西藏西部和中部省界山地之外,或许是这些区域信号传输设备相对缺少,社交媒体使用率较低;如图 4(b)所示,购物服务 POI 回归系数较大的区域分布在我国西北和西南部分区域,以及部分省份交界处和山区,或许是这些区域购物服务设施相对不足,与人口数不匹配;如图 4(c)所示,对于土地覆盖的城镇面积来说,回归系数较大的区域,即红色分布的区域与腾讯定位数据相似,意味着红色的区域城镇化率相对较低;如图 4(d)所示,对于夜间灯光强度数据,除了东北和部分边境地区,大部分与城镇建设面积相似,或许是这些区域夜间灯光强度(包括工业和家庭用电量)与区域的人口数不匹

配。用各种指标得到的结果表明,不同数据的空间差异对人口估算精度的贡献是不同的。因此,选择适当的指标可以有效改善中国西部、中部和东北等地区因区域差异而导致的人口估算空间不确定性。

除了局部回归系数分析之外,图5展示了单个建模数据因子的GWR局部R^2(拟合优度)。从图5可以看出,各数据的局部拟合效果,颜色越红的区域拟合效果越好。从整体来看,除东北三省外,靠近省会城市的区域大部分拟合效果是比较好的,西藏、新疆和珠三角、长三角区域表现尤为明显。从数据对比来看,如图5(a)所示,腾讯定位数据对县域人口的拟合度从整体来说明显是最优的,即拟合效果好的区域多,其次是POI数据(图5(b)),后两位是土地覆盖中的城镇建设面积(图5(c))和夜间灯光强度数据(图5(d))。但腾讯定位数据在东北部分区域的拟合效果相对没这么好,可以用POI和土地覆盖数据较好的部分补充,说明了数据间在使用时的互补性。

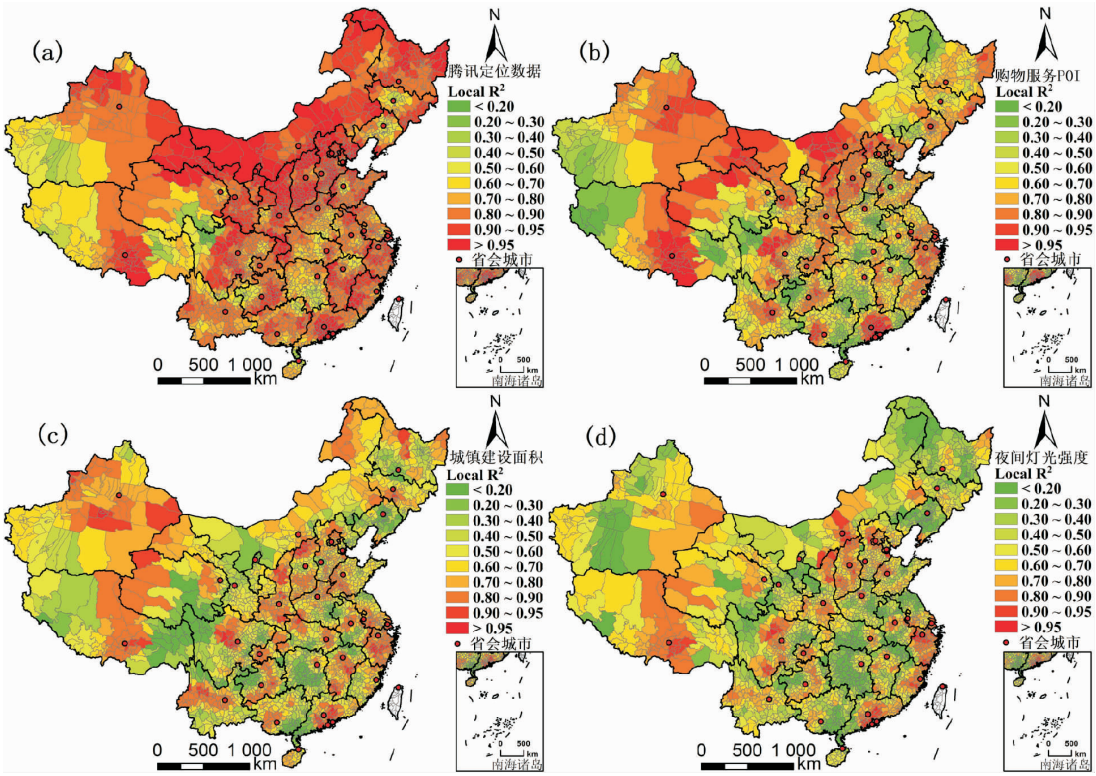

图5 局部R^2图

Fig. 5 Local R^2 plots

注:1. 图中,(a)为腾讯定位数据;(b)为购物服务POI;(c)为城镇建设面积;(d)为夜间灯光强度。

2. 基于自然资源部标准地图服务网站GS(2019)1822号标准地图制作,底图边界无修改。

3.3 中国区县人口空间拟合结果

利用回归预测人口与第七次全国人口普查的真实人口,作出的散点图,可以从整体展示出中国区县人口的拟合结果。图上的散点越是聚集于对角线上拟合效果越好,表明利用此数据或方法估计的人口数与真实的普查人口数越接近,散点越是远离对角线则误差越大。

图6(a)~图6(d)是使用GWR方法,并利用单一数据预估的人口与真实人口作出的散点图,可以看出,图6(a)利用腾讯定位数据回归估算的人口与真实人口的散点聚集效果最好,即拟合效果最好,随后是POI、城镇建设面积和夜光强度。图6(e)和图6(f)则是利用多种数据和OLS、GWR两种方法得出最优回归人口数,并与真实人口数作出的散点图,可以看出,图6(f)的聚集效果优于图6(e),证明在估算县域人口方面,GWR方法优于OLS方法。从散点图可以用看出,散点图展现的结果与表2的结果相呼应。

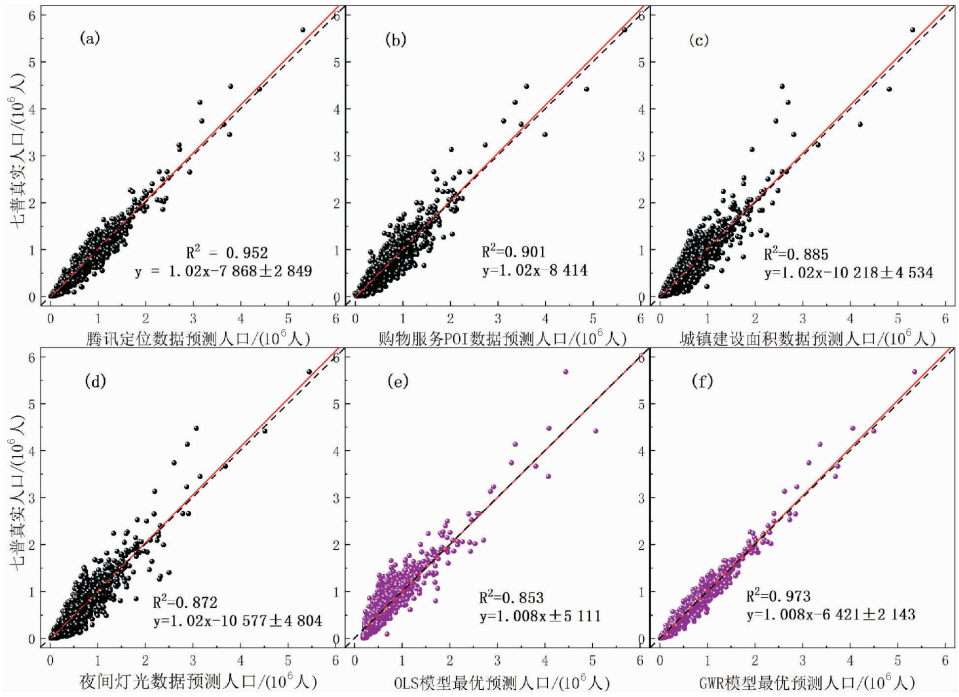

图 6　预测人口与真实人口散点图

Fig. 6　Scatter plots of Predicted Population vs. Actual Population

注:图中,(a)为腾讯位置大数据;(b)为购物服务 POI;(c)为城镇建设面积;(d)为夜间灯光强度;(e)为 OLS 模型;(f)为 GWR 模型。

最后综合使用多种数据,并利用相关性分析、逐步回归和地理加权回归分析(GWR)得到最优的拟合人口数,如图 7。由图 7 可以看出,最后使用 GWR 模型拟合的人口分布与真实的人口分布(图 1)相吻合,2020 年人口空间分布符合我国人口密度分界线“黑河 - 腾冲线”,使用此方法可为获取实时的人口情况提供参考。

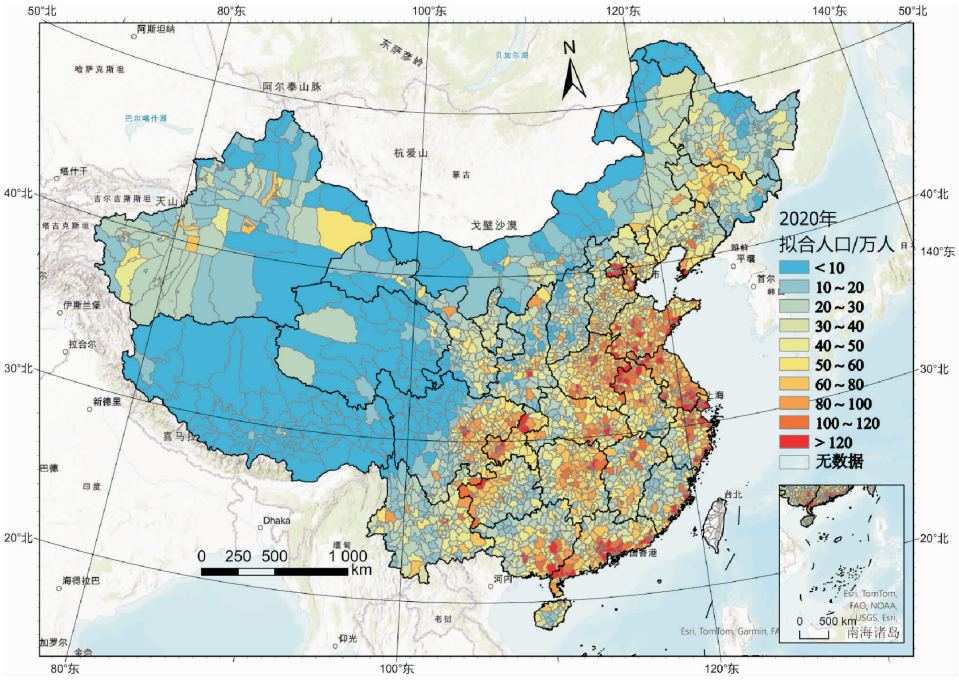

图 7　GWR 模型拟合人口分布图

Fig. 7　Map of Population Distribution Fitted by GWR Model

注:基于自然资源部标准地图服务网站 GS(2019)1822 号标准地图制作,底图边界无修改。

4 结 论

利用遥感(RS)和地理信息系统(GIS)技术对人口空间分布的研究,快速获取精细的人口数据,对于城市规划、环境保护、资源配置和社会的可持续发展等方面的研究至关重要。本文通过对比分析腾讯位置大数据、高德POI数据、土地利用/覆盖数据和夜间灯光数据等多种数据,并分别采用普通最小二乘法和地理加权回归两种统计模型,对县城人口估算的效能进行了深入的探讨,主要结论如下:

(1)在中国县域人口分析中,地理加权回归(GWR)方法明显优于普通最小二乘法(OLS)。这表明考虑地理空间异质性对人口分布模型的建立是至关重要的;

(2)在本研究的数据对比中,腾讯定位大数据的表现最为优异,其次是兴趣点POI数据,相较于土地覆盖/利用数据和夜间灯光数据,前两者的模拟精度更高;

(3)本研究采用GWR模型和综合腾讯位置大数据、土地利用/覆盖数据和夜间灯光数据等多源数据进行县域人口的估计,其拟合优度(R^2)为0.972,并且实现了85.4%的县域人口空间模拟精度。

这些结论不仅对于人口空间分布研究具有重要意义,而且还为相关领域的进一步研究提供了有益的指导和参考。但是在本研究中可以看出,各数据对于胡焕庸线以西,我国西部部分地区的人口估算精度尚待提高,且不能够体现县域内部的小区域差异,未来的研究尺度可以进一步精细。

参考文献:

[1] Yong C. An assessment of China's fertility level using the variable-r method[J]. Demography, 2008,45(2):271-281.

[2] Wu J W,Yu Z,Wei Y D, et al. Changing distribution of migrant population and its influencing factors in urban China: Economic transition, public policy, and amenities[J]. Habitat International,2019,94:102063.

[3] Chen J D,Fan W,Li K, et al. Fitting Chinese cities' population distributions using remote sensing satellite data[J]. Ecological Indicators,2019,98:327-333.

[4] Wang L T,Wang S X,Zhou Y, et al. Mapping population density in China between 1990 and 2010 using remote sensing[J]. Remote Sensing of Environment,2018,210:269-281.

[5] Tan M H,Li X B,Li S J, et al. Modeling population density based on nighttime light images and land use data in China[J]. Applied Geography,2018,90:239-247.

[6] Yu S S,Zhang Z X,Liu F. Monitoring population evolution in China using time-series DMSP/OLS nightlight imagery[J]. Remote Sensing,2018,10(2):194.

[7] Zhang G,Guo X Y, Li D R, et al. Evaluating the potential of LJ1-01 nighttime light data for modeling socio-economic parameters[J]. Sensors,2019,19(6):1465.

[8] 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 of Environment, 1999, 68(1): 77-88.

[9] Sutton P, Roberts D, Elvidge C, et al. Census from heaven: An estimate of the global human population using night-time satellite imagery[J]. International Journal of Remote Sensing, 2010, 22(16): 3061-3076.

[10] Chen X, Nordhaus W. A test of the new VIIRS lights data set: Population and economic output in Africa[J]. Remote Sensing, 2015, 7(4): 4937-4947.

[11] Zeng C Q, Zhou Y, Wang S X, et al. Population spatialization in China based on night-time imagery and land use data[J]. International Journal of Remote Sensing, 2011, 32(24): 9599-9620.

[12] Li X M, Zhou W Q. Dasymetric mapping of urban population in China based on radiance corrected DMSP-OLS nighttime light and land cover data[J]. Science of the Total Environment, 2018, 643: 1248-1256.

[13] Guo W, Liu J K, Zhao X S, et al. Spatiotemporal dynamics of population density in China using nighttime light and geographic weighted regression method[J]. International Journal of Digital Earth, 2023, 16(1): 2704-2723.

[14] Lung T, Lübker T, Ngochoch J K, et al. Human population distribution modelling at regional level using very high resolution satellite imagery[J]. Applied Geography, 2013, 41: 36-45.

[15] Ye T T, Zhao N Z, Yang X C, et al. Improved population mapping for China using remotely sensed and points-of-interest data within a random forests model[J]. Science of the Total Environment, 2019, 658: 936-946.

[16] Wu T J, Luo J C, Dong W, et al. Disaggregating county-level census data for population mapping using residential geo-objects with multisource geo-spatial data[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote

Sensing, 2020, 13: 1189-1205.
[17] Baynes J, Neale A, Hultgren T. Improving intelligent dasymetric mapping population density estimates at 30 m resolution for the conterminous United States by excluding uninhabited areas[J]. Earth System Science Data, 2022, 14(6): 2833-2849.
[18] Stevens F R, Gaughan A E, Linard C, et al. Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data[J]. PLoS One, 2015, 10(2): e0107042.
[19] Doxsey-Whitfield E, MacManus K, Adamo S B, et al. Taking advantage of the improved availability of census data: A first look at the gridded population of the world, version 4[J]. Papers in Applied Geography, 2015, 1(3): 226-234.
[20] Freire S, Schiavina M, Florczyk A J, et al. Enhanced data and methods for improving open and free global population grids: Putting "leaving no one behind" into practice[J]. International Journal of Digital Earth, 2020, 13(1): 61-77.
[21] Bai Z Q, Wang J L, Wang M M, et al. Accuracy assessment of multi-source gridded population distribution datasets in China[J]. Sustainability, 2018, 10(5): 1363.
[22] McKenzie G, Janowicz K, Gao S, et al. POI pulse: A multi-granular, semantic signature-based information observatory for the interactive visualization of big geosocial data[J]. Cartographica: The International Journal for Geographic Information and Geovisualization, 2015, 50(2): 71-85.
[23] Yoshida D, Song X F, Raghavan V. Development of track log and point of interest management system using Free and Open Source Software[J]. Applied Geomatics, 2010, 2: 123-135.
[24] 张健. 基于 POI 大数据的城市零售商业空间布局与人口耦合关系研究——以上海市为例[J]. 复旦学报(自然科学版),2019,58(2):151-161.
[25] Bakillah M, Liang S, Mobasheri A, et al. Fine-resolution population mapping using OpenStreetMap points-of-interest[J]. International Journal of Geographical Information Science, 2014, 28(9): 1940-1963.
[26] Cai J X, Huang B, Song Y M. Using multi-source geospatial big data to identify the structure of polycentric cities[J]. Remote Sensing of Environment, 2017, 202: 210-221.
[27] 淳锦,张新长,黄健锋,等. 基于 POI 数据的人口分布格网化方法研究[J]. 地理与地理信息科学,2018,34(4):83-89, 124.
[28] 赵鑫,宋英强,刘轶伦,等. 基于卫星遥感和 POI 数据的人口空间化研究——以广州市为例[J]. 热带地理,2020,40(1):101-109.
[29] 刘正廉,桂志鹏,吴华意,等. 融合建筑物与 POI 数据的精细人口空间化研究[J]. 测绘地理信息,2021,46(5):102-106.
[30] 高航. 基于多源地理信息和随机森林模型的高精度人口空间化——以长沙市为例[D]. 长沙:湖南师范大学,2020.
[31] França U, Sayama H, McSwiggen C, et al. Visualizing the "heartbeat" of a city with tweets[J]. Complexity, 2016, 21(6): 280-287.
[32] Patel N N, Stevens F R, Huang Z J, et al. Improving large area population mapping using geotweet densities[J]. Transactions in GIS: TG, 2017, 21(2): 317-331.
[33] Deville P, Linard C, Martin S, et al. Dynamic population mapping using mobile phone data[J]. Proceedings of the National Academy of Sciences of the United States of America, 2014, 111(45): 15888-15893.
[34] Cheng Z F, Wang J H, Ge Y. Mapping monthly population distribution and variation at 1-km resolution across China[J]. International Journal of Geographical Information Science, 2022, 36(6): 1166-1184.
[35] Zhao S, Liu Y X, Zhang R, et al. China's population spatialization based on three machine learning models[J]. Journal of Cleaner Production, 2020, 256: 120644.
[36] Tu W N, Liu Z, Du Y Y, et al. An ensemble method to generate high-resolution gridded population data for China from digital footprint and ancillary geospatial data[J]. International Journal of Applied Earth Observation and Geoinformation, 2022, 107: 102709.
[37] Chen Y H, Xu C C, Ge Y, et al. A 100-m gridded population dataset of China's seventh census using ensemble learning and geospatial big data[J]. Earth System Science Data Discussions, 2024, 2024: 1-19.
[38] Xu Y, Song Y M, Cai J X, et al. Population mapping in China with Tencent social user and remote sensing data[J]. Applied Geography, 2021, 130: 102450.
[39] 李佳洺,陆大道,徐成东,等. 胡焕庸线两侧人口的空间分异性及其变化[J]. 地理学报,2017,72(1):148-160.
[40] Chen M X, Xian Y, Huang Y H, et al. Fine-scale population spatialization data of China in 2018 based on real location-based big data[J]. Scientific Data, 2022, 9(1): 624.
[41] 张海平,周星星,汤国安,等. 基于 GIS 场模型的城市餐饮服务热点探测及空间格局分析[J]. 地理研究,2020,39(2):354-369.

【责任编辑: 孙向荣】