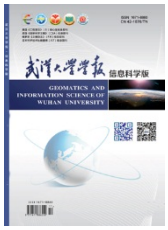


武汉大学学报(信息科学版)
Geomatics and Information Science of Wuhan University
ISSN 1671-8860,CN 42-1676/TN

《武汉大学学报(信息科学版)》网络首发论文

题目：顾及格网属性分级与空间关联的人口空间化方法
作者：吴京航，桂志鹏，申力，吴华意，刘洪波，李锐，梅宇翱，彭德华
DOI：10.13203/j.whugis20200379
收稿日期：2021-04-13
网络首发日期：2021-09-15
引用格式：吴京航，桂志鹏，申力，吴华意，刘洪波，李锐，梅宇翱，彭德华．顾及格网属性分级与空间关联的人口空间化方法．武汉大学学报(信息科学版).
https://doi.org/10.13203/j.whugis20200379

引用格式：

吴京航, 桂志鹏, 申力, 等. 顾及格网属性分级与空间关联的人口空间化方法[J]. 武汉大学学报·信息科学版. DOI: 10.13203/j.whugis20200379（WU Jinghang，GUI Zhipeng，SHEN Li，et al. Population Spatialization by Considering Pixel-level Attribute Grading and Spatial Association [J].Geomatics and Information Science of Wuhan University. DOI: 10.13203/j.whugis20200379）

顾及格网属性分级与空间关联的人口空间化方法

吴京航[1]，桂志鹏[1,2]，申力[1]，吴华意[2,3]，刘洪波[4]，李锐[3]，梅宇翱[1]，彭德华[3]

1 武汉大学遥感信息工程学院，湖北 武汉，430079

2 地球空间信息技术协同创新中心，湖北 武汉，430079

3 武汉大学测绘遥感信息工程国家重点实验室，湖北 武汉，430079

4 重庆市地理信息和遥感应用中心，重庆市，401147

摘要：现有人口空间化方法多基于行政单元构建回归模型并分配格网单元人口，但分析单元的尺度差异引发模型迁移问题。同时，格网特征建模仅考虑格网自身属性，导致格网间空间关联被人为割裂。为此，本文基于随机森林模型提出一种顾及格网属性分级与空间关联的人口空间化方法（Population Spatialization by Considering Pixel-level Attribute Grading and Spatial Association, PAG-SA）。该方法在格网特征建模中，1）基于自然断点法构造建筑区类别约束的夜间灯光分级特征，并在行政单元尺度统计各等级网格占比作为训练输入，以减小模型跨尺度误差；2）利用核密度估计刻画邻域兴趣点（Point of Interest, POI）对当前格网人口分布的影响及距离衰减效应；3）基于叠置分析统计不同类型建筑区轮廓包含的各类 POI 数量，提升特征建模精细度。论文选取武汉市作为实验区域，在街道尺度与 WorldPop、GPW 及中国公里网格人口数据集对比验证方法的有效性。结果表明该方法的平均绝对值误差仅为对比数据集的 1/6-1/3。此外，本文还探讨了特征构成、格网大小及核密度带宽对精度的影响。

关键词：人口空间化；随机森林；多源数据融合；跨尺度问题；核密度估计；叠置分析

收稿日期：2021-04-13
项目资助：国家重点研发计划（2018YFC0809806, 2017YFB0503704）；国家自然科学基金（41971349, U20A2091, 42090010）。
第一作者：吴京航，硕士，研究方向为人口空间化。wyw1294@whu.edu.cn
通讯作者：桂志鹏，博士，副教授。zhipeng.gui@whu.edu.cn

人口空间化是人口学及地理学的研究热点，旨在通过建立数学模型，将行政单元人口数据分配到细粒度格网中[1]，从而精细刻画人口分布。其在商业决策、区域规划及灾害救援等领域具有广泛的应用[2]，众多学者基于遥感数据和地理信息技术开展了深入的研究[3]。基于建模方法的异同，现有方法可分为区域插值法和回归建模法两大类。

区域插值法基于特定准则和插值方法将行政单元的人口数据转换到格网单元中，主要包括面积权重模型[4-5]、核密度估计模型[6-7]及分区密度模型[8-9]等。其中，面积权重模型假定行政区内人口密度均等，根据格网内各行政区面积实现人口分配。该模型虽简单易行，但未考虑影响人口分布的自然、经济和社会因素[10]，无法体现行政区内的人口密度差异。核密度估计模型假定人口密度从区域中心向外围递减，基于人口加权质心将人口密度内插到格网面。该模型能够模拟人口连续分布情况，但未考虑人口分布影响因素，带宽值确定较主观[11]。分区密度模型假定面元内同一类别分区上人口分布一致，通过面插值技术实现人口空间化[11]。该模型能够体现不同分区间的人口分布差异，但各分区内人口分布仍然具有均质性且权重分配较为困难。

回归建模法通过建立建模因子和人口数据间的回归模型估算人口分布，主要包括多元线性回归、随机森林及深度学习模型等。多元线性回归易建模、便于推广且结果较为可控[11]，但也存在容易过拟合且精细度不足的缺点，因此常用于粗粒度、大范围人口估算[12]。随机森林模型能够较好地避免模型过拟合，对异常值和噪声具有较高的容忍度[13]，适合处理高维数据建模问题。随着遥感及社会感知技术的发展，人口空间化建模的数据愈加多源化和精细化。基于随机森林模型融合多源数据进行人口估算是目前人口空间化研究的重要方向[14-17]。近年来，深度学习已用于建模卫星影像像素值和人口格网数据集间的回归关系[18-19]，但由于难以获取真实的格网人口样本，这类方法尚未得到广泛应用。因此，有研究基于全国区县人口融合社交媒体、夜光及数字高程模型等数据[20-21]构建深度学习模型，但由于训练样本的限制，此类方法无法针对小范围研究区域实现精细建模[21]。

相对于区域插值，回归建模能够通过特征提取考虑复杂因素对人口分布的影响，并通过模型再训练迁移到其他区域，但基于回归建模的人口空间化研究目前仍然存在一些不足。首先，由于缺乏真实格网人口数据，回归建模法通常使用行政单元数据建模，再将模型迁移到格网上，二者间的地理尺度差异导致训练与估算之间的跨尺度问题。同时，现有方法大多仅考虑格网本身属性，而未顾及邻近格网中不同类型空间要素对当前格网人口分布的影响，导致格网间的空间关联被人为割裂，影响空间化的合理性与准确性。为此，本文针对中小范围研究区域，基于随机森林模型提出一种顾及格网属性分级与空间关联的人口空间化方法。该方法在行政单元尺度引入格网属性分级统计信息，将特征提取统一在格网级别以减小跨尺度误差，并结合核密度估计构建邻域 POI 特征[22-23]，为不同类型 POI 选择合适的带宽。同时，该方法基于叠置分析统计不同类型建筑区轮廓包含的各类 POI 数量，提升特征建模精细度。本文以武汉市作为实验区域，将 PAG-SA 与 WorldPop、GPW 及中国公里网格人口数据在街道尺度进行对比。实验结果表明 PAG-SA 能够有效提升估算精度，其平均绝对值误差 7618 仅为对比数据集的 1/6-1/3，同时在高中低人口密度区域均具有更好的拟合优度。此外，本文还讨论了特征构成、格网尺度及核密度带宽对精度的影响。

1 研究区概况和数据来源

1.1 研究区概况

本文研究区域为湖北省武汉市，其街道级行政区划及人口密度等级如图 1 所示。武汉市下辖 13 个区，185 个街道，总面积 8569.15 平方千米。2015 年武汉市户籍人口达 829.26 万人，常住人口达 1060.77 万人。13 个下辖区中包含 7 个主城区，即洪山区、青山区、武昌区、汉阳区、硚口区、江汉区和江岸区，占武汉市总人口的 61.67%；

6 个远城区分别为新洲区、江夏区、蔡甸区、黄陂区、东西湖区和汉南区。武汉市不仅具有人口分布众多的主城区，也包含地理范围广阔、人口密度较小的远城区，人口分布情况非常复杂，选择武汉市作为研究区域对于人口空间化研究具有借鉴意义。

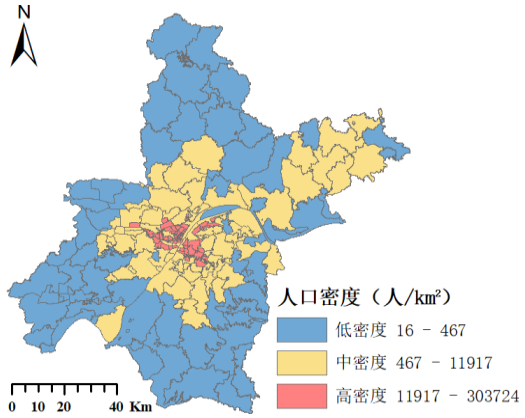


图 1. 武汉市街道行政区划及人口高中低密度区域

Fig.1 High Middle and Low Population Density Streets in Wuhan

1.2 数据来源

本文使用 NPP/VIIRS 夜光数据、高德 POI 及武汉市地理国情普查建筑区数据作为研究数据，详情如表 1 所示。夜间灯光数据能反映人类活动，是人口空间化建模的理想数据源[24-25]。POI 数据具有语义丰富且与人口分布高度相关的优点，常被用于人口建模[14-16]。地理国情普查建筑区数据提供的高精度建筑区轮廓及类别，有助于修正夜间灯光溢出的影响，区分不同用地类型，进而辅助人口估算。由于缺失武汉市 2015 年的 POI 数据，本文选取 2017 年数据代替，其他数据源采集时间均为 2015 年。

表 1. 所选用的研究数据

Tab.1 Dataset Used in This Study

数据类型	数据来源	年份	格式	描述
夜间灯光	美国国家环境中心	2015	栅格	NPP/VIIRS 全年月份数据合成夜间灯光影像，分辨率约为 500m
地理国情普查建筑区	武汉市测绘局	2015	矢量	基于分辨率低于 1m 的多源航空航天遥感影像数据生成。使用的建筑区类型包括高密度多层及以上房屋、低密度多层及以上房屋、高密度低矮房屋、低密度低矮房屋
POI	高德软件有限公司	2017	矢量	使用的 8 类 POI 包括休闲娱乐、住宿、医院、居民小区、科研教育、购物、金融服务及餐饮
武汉市行政区划	武汉市测绘局	2015	矢量	包括武汉市区县、街道级别的轮廓数据及对应的常住人口信息

2 PAG-SA 的计算与验证流程

PAG-SA 的计算与验证流程如图 2 所示，共由四个部分组成，包括数据预处理、特征提取、模型训练与估算及格网人口分配。数据预处理阶段对多源数据进行坐标转换、栅格数据重采样、格网信息统计及街道信息汇总。特征提取阶段融合建筑区轮廓数据、夜光数据及 POI 数据，生成训练及估算时的特征向量。模型训练与估算阶段使用随机森林模型，输入构建的特征向量，输出格网人口权重。格网人口分配阶段首先基于无房屋无人口原则[26]约束格网人口权重，然后在区县级别进行权重归一化并按权重分配格网人口。

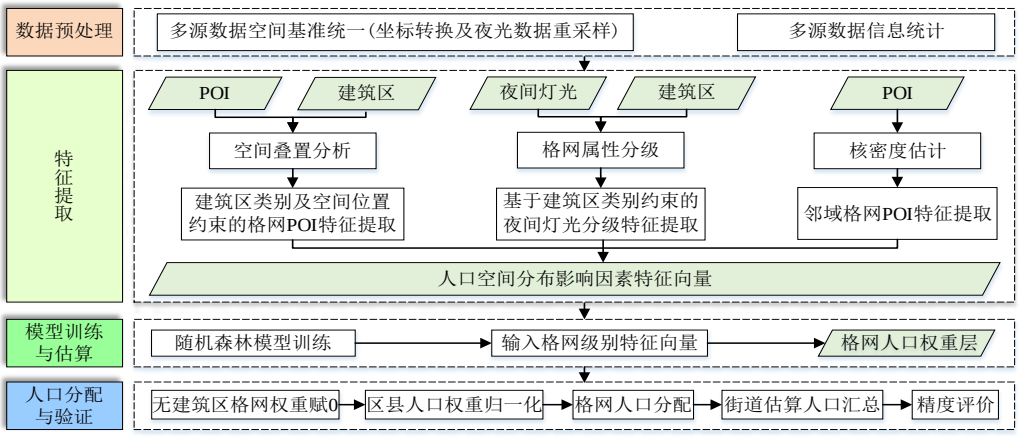


图 2 顾及格网属性分级与空间关联的人口空间化方法 PAG-SA 的计算与验证流程

Fig.2 Workflow of Calculation and Validation Process for PAG-SA

2.1 数据预处理

数据预处理主要包括多源数据空间基准统一及信息统计。使用 ArcGIS 等软件，将前述数据进行坐标转换，然后基于不同的格网尺度分别统计格网信息，具体包括，1）对夜光数据进行坐标转换及重采样，统计各个格网的夜光值。2）对建筑区轮廓数据进行坐标转换，基于 Java Topology Suite（JTS）统计各个格网的建筑区面积占比。3）对 POI 进行坐标转换。4）计算街道人口密度。

2.2 特征提取

PAG-SA 综合使用三种特征提取方法构建训练及估算向量。1）以建筑区轮廓为约束统计格网内各类 POI 数量特征，以建模不同类型建筑区与 POI 组合方式对人口密度的影响。2）使用格网属性分级方法提取基于建筑区类别约束的夜间灯光分级特征，利用格网属性分级方法减小模型跨尺度误差，结合建筑区类别约束缓解夜光值溢出问题。3）统计邻近格网 POI 在当前格网中心的核密度估计值，从而建模邻域 POI 与人口密度间的关系。

2.2.1 建筑区类别及空间位置约束的格网 POI 特征提取

作为一种易获取的地理空间数据，POI 具有语义丰富且与人口分布高度相关的特点[14]。基于 POI 数据进行人口空间化，相比于土地利用类型数据能够更好地保留人口空间分布的细节信息。目前 POI 数据在人口空间化中应用广泛[14-16]，但大多只考虑了格网内 POI 的绝对数量，忽略了建筑区对 POI 的潜在空间位置约束。不同建筑类型具有不同的人口密度，分布于不同建筑区类型内的 POI 对人口的吸引力也存在差异。例如，分布于高密度多层建筑区内的 POI 对人口的吸引力可能比位于低密度低矮建筑区内的 POI 更高。因此，PAG-SA 基于建筑区类别及空间位置约束，统计不同建筑区类别内的 POI 数量，以便模型拟合其与人口密度间的相关关系。具体步骤如下：1）将建筑区数据和 POI 数据进行空间叠置分析，统计各个格网中分布于各类建筑区类别内的各类 POI 数量。假设建筑区种类数为$Type_{building}$，POI 种类数为$Type_{poi}$，则空间叠置后的特征维数为$Type_{building} \cdot Type_{poi}$。2）统计街道内所有格网中各类特征的平均值作为模型训练的输入。

2.2.2 基于建筑区类别约束的夜间灯光分级特征提取

夜间灯光数据能反映居民点、交通道路及产业结构等多种信息，但存在夜光值溢出的现象[24]，从而影响人口空间化的精度。针对上述问题，有学者提出使用土地利用数据进行约束[24]，统计格网内城镇用地和农村居民用地的总面积，如果总面积大于 0 则表示该夜光值有效。该方法能够缓解夜光值溢出问题，但无法应对灯光来源的复杂性。为此，PAG-SA 在格网尺度对建筑区数据和夜光数据进行属性分级，将数量信息转换成类别信息，然后使用类别合并的方法实现数据融合，具体步骤如下。

1）格网属性分级

本文采用自然断点法对夜光值和建筑区类型的面积占比进行分级，通过戴维森堡丁系数（Davies-Bouldin Index, DBI）确定最佳分级数量，流程如图 3 所示。

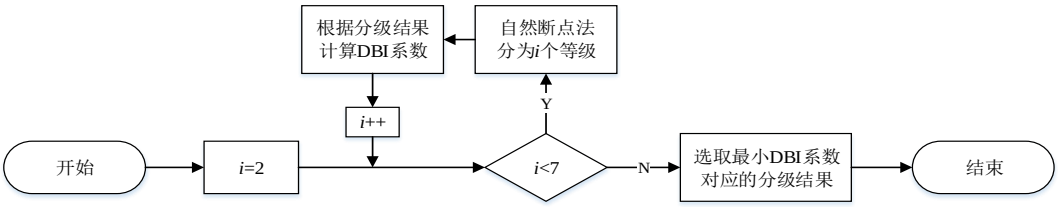


图 3 自然断点法分级流程图

Fig.3 Workflow of Attribute Grading based on Natural Breaks

自然断点法是一种数据分级算法，算法原理是对分类间隔加以识别，实现类间方差最大，类内方差最小。DBI 系数是一种评估聚类算法优劣的指标，取值范围为[0,+∞)，DBI 系数越小表明等级内距离越小，等级间距离越大，计算方法如式 1 所示。

$$DB^k = \frac{1}{k}\sum_{x=1}^{k}\max_{y\neq x}\left(\frac{\overline{\alpha_x}+\overline{\alpha_y}}{|\sigma_x-\sigma_y|}\right) \quad (1)$$

其中DB^k表示自然断点法分类数为k时对应的 DBI 系数，$\overline{\alpha_x}$和$\overline{\alpha_y}$分别是分级结果中第x和第y个等级的类内平均距离，σ_x和σ_y分别是第x和第y两个等级中心间的距离。

2）特征向量构建

对于一个格网单元，根据分级结果确定各类数据的等级。若格网属于第t类数据第k个等级，就在第t类数据第k个等级对应的特征向量编码处将特征值赋为 1，在第t类数据其他等级编码处赋为 0。例如，若经过格网分级，夜光亮度值介于 0-100 之间、100-200 之间和 200-255 之间分别为第 1、2、3 类，则对应特征向量编码分别为[1,0,0]、[0,1,0]和[0,0,1]。根据以上方法获取特征向量后，将夜光数据与建筑区数据的特征向量按照与运算进行融合以构建组合向量，如式 2 所示：

$$\boldsymbol{y} = merge([l_1, l_2, \dots, l_m], [b_1, \dots, b_n]) = [l_1\&b_1, \dots l_1\&b_n, \dots, l_m\&b_1, \dots, l_m\&b_n] \quad (2)$$

其中，$\boldsymbol{y}$表示组合向量，$[l_1, l_2, \dots, l_m]$和$[b_1, \dots, b_n]$分别表示该格网夜光和建筑区数据的特征向量，&表示与运算。例如，当l_2和b_3取值均为 1，则$l_2\&b_3$等于 1，表示该格网夜光亮度为第二等级且建筑密集程度为第三等级。

在街道单元尺度，若第i个街道的总格网数为N_i，属于第t类建模数据第k个等级的格网数量为$N_i^{t,k}$，则其特征向量可表示为式 3：

$$\boldsymbol{\beta_i} = \frac{(N_i^{1,1},\dots,N_i^{1,c1},N_i^{2,1},\dots,N_i^{2,c2},\dots,N_i^{t,1},\dots,N_i^{t,ct})}{N_i} \quad (3)$$

2.2.3 邻域格网 POI 特征提取

2.2.1 节中提取的特征只包含格网本身的 POI 语义信息，导致格网间的空间关联被人为割裂。为此本文使用核密度估计提取邻域 POI 特征，以刻画邻近格网 POI 对当前格网人口分布的影响及距离衰减效应[27]。核密度估计是分析点事件分布和识别热点[28-29]的一种常用方法。相关研究及本文实验表明，核函数的选择对结果影响不大[14]，带宽（搜索半径）是核密度估计的主要参数[22]。为此，本文选用密度函数较为平滑且使用场景广泛的高斯核作为核函数（式 4）。点对象核密度估计方法如式 5 所示[29]。本文针对每一类 POI，通过比较多种带宽取值获取相对最优带宽。

$$k\left(\frac{d_{is}}{r}\right) = \begin{cases} \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{{d_{is}}^2}{2r^2}\right) & \text{if } 0 < d_{is} \le r \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

$$\mu(s) = \sum_{i=1}^{i=n}\frac{1}{\pi r^2}k\left(\frac{d_{is}}{r}\right) \quad (5)$$

其中k为核函数，$\mu(s)$表示位置s处的核密度估计值，r为带宽，d_{is}为i点到当前位置s的距离。

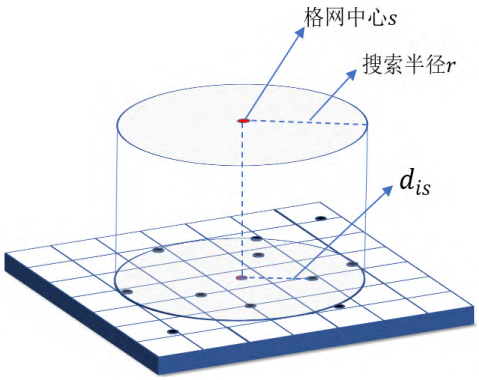


图 4. 基于格网中心的核密度估计示意

Fig.4. Illustration of Kernel Density Estimation upon Grid Cell Centre

本文对 POI 进行核密度估计时，位置s为当前格网中心点，如图 4 所示。图中红点表示当前格网中心，蓝色点表示 POI，圆柱半径表示搜索半径。各 POI 相对于当前格网中心的核密度估计值记作$k\left(\frac{d_{is}}{r}\right)$，格网中心最终核密度估计值为$\mu(s)$。最后，统计各个街道内所有格网的平均值作为模型训练的输入。

2.3 模型训练与估算

本文选用随机森林构建回归模型，原因有三：1）该模型对异常值和噪声具有较高的容忍度[13]。在人口空间化中，由于数据源的多样性及人口分布的复杂性，在特征向量中往往存在异常值却难以发现。例如以街道数据进行训练时，面积较小的街道可能存在数据分布极端的训练样本。2）融合多源数据导致特征维数增多，造成筛选及降维的困难，而随机森林模型能够处理高维数据，避免人为特征选择。3）随机森林模型中决策树相互独立，利于并行实现，训练速度快。

PAG-SA 的训练与估算流程如图 5 所示，训练阶段输入街道级别特征拟合街道人口密度，估算阶段输入格网级别特征生成格网人口权重。

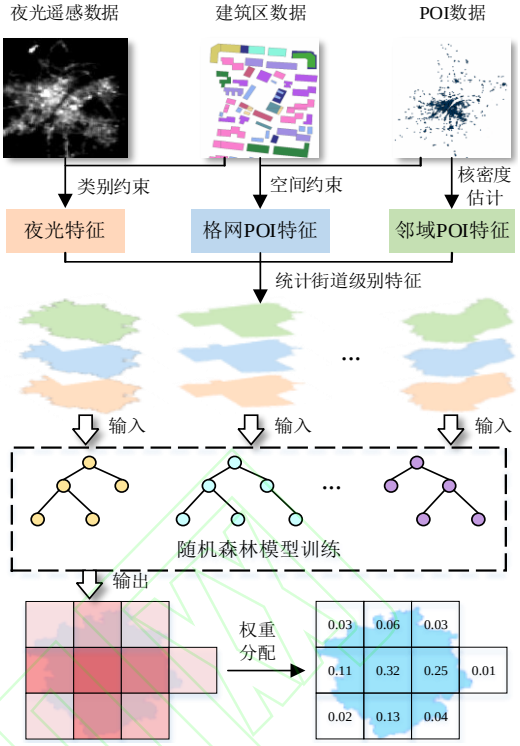


图 5. PAG-SA 训练与估算流程

Fig.5 Workflow of Training and Estimation for PAG-SA

2.4 人口权重修正与人口分配

针对人口权重可能存在无建筑区但权重非零的问题，本文根据无房屋无人口原则[26]，将无建筑区的格网人口权重赋 0。

经过建筑区数据修正后，对各个区县进行权重归一化并将区县人口按照权重分配到各个格网中，最后根据格网与街道的映射关系计算街道人口。格网人口计算方法如式 6 所示：

$$SI_{ij} = SI_i * \frac{W_{ij}}{\sum_{u=1}^{N_i} W_{iu}} \quad (6)$$

其中，i表示第i个区县，j表示第j个格网，SI_{ij}表示第i个区县第j个格网的最终人口值，SI_i表示第i个区县的人口总值，W_{ij}、W_{iu}分别表示第i个区县第j个和第u个格网权重值，N_i表示第i个区县的格网总数。

3 实验结果与分析

3.1 精度评价指标

本文选取平均绝对值误差（Mean Absolute Error，MAE）、均方根误差（Root Mean Square Error，RMSE）、决定系数（Coefficient

of Determination，R^2）三种指标进行精度评价（式 7），其中 MAE 反映人口估算误差的绝对值，RMSE 刻画人口估算值与真实值之间的偏差程度，决定系数度量人口估算值与真实人口的拟合程度。

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|Predict_i - Real_i|$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Predict_i - Real_i)^2} \quad (7)$$

$$R^2 = 1 - \frac{\sum_i(Real_i - Predict_i)^2}{\sum_i\left(Real_i - \overline{Real}\right)^2}$$

其中N表示街道总数，$Predict_i$表示街道i的估算值，$Real_i$表示街道i的真实值，$\overline{Real}$表示所有街道真实人口的平均值。

3.2 实验分析

3.2.1 总体精度验证实验

为了验证特征提取方法融合多源数据的有效性，将仅使用 POI 密度的方法、综合使用 POI、夜间灯光及建筑区数据进行特征向量直接拼接的方法与 PAG-SA 在 200m、500m 和 1000m 三种格网尺度下进行精度对比，结果如图 6 所示。其中，特征向量直接拼接的方法不考虑特征之间的关联，POI 统计值为 POI 在街道或格网内的密度，夜光统计值为街道或格网内的平均夜光亮度，建筑区统计值为建筑区在街道或格网内的面积占比。3.2.1、3.2.2 及 3.2.4 节实验中，各类 POI 的核密度带宽取值见表 2“所选带宽”列。

图 6 表明 PAG-SA 能够有效提升精度，且不同格网尺度的效果存在显著差异。直接拼接方法在三种格网尺度下均出现 R^2 下降、MAE/RMSE 上升的情况，说明使用建筑区及夜光数据在街道尺度上训练构建的模型对格网尺度不适用，引发模型跨尺度问题。同时，不恰当的数据融合方式可能导致精度降低。而相对于仅使用 POI 的方法，PAG-SA 的拟合优度及准确度在三个尺度下均有所提升。MAE 下降 4%~16%，且 200m 格网尺度精度最优，随着格网尺度增大精度逐渐降低，说明不同尺度下特征的精细程度和表达能力不同。

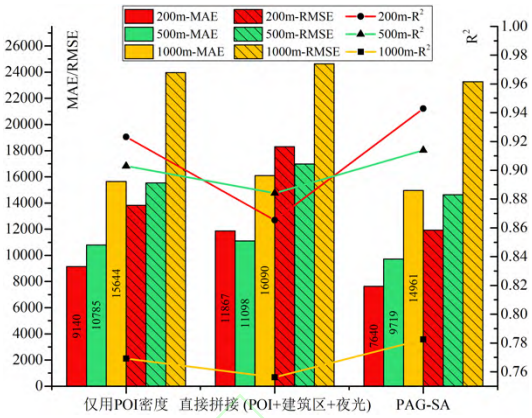


图 6. PAG-SA 与仅使用 POI 密度、三种数据特征直接拼接法的精度对比

Fig.6 Accuracy Comparison of Three Feature Extraction Methods

3.2.2 特征提取各步骤精度对比

为了进一步验证特征提取各步骤的有效性，本实验对比了四种特征提取方法，实验结果如图 7 所示。第一种方法仅使用 POI 密度作为参照；第二种提取建筑区类别及空间位置约束的 POI 特征，记为 P1；第三种在第二种基础上提取夜间灯光分级特征，记为 P2；第四种在第三种的基础上引入 POI 核密度特征，记为 P3。

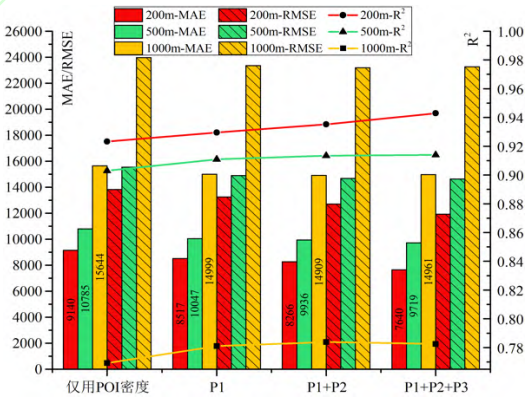


图 7 PAG-SA 各特征提取步骤的精度提升

Fig.7 Accuracy Improvement at Each Feature Fusion Step of PAG-SA

由图 7 可知，除了 1000m 格网外，各特征提取步骤均有助于精度提升，但是不同格网尺度的效果不同。1）从格网尺度上看，200m 的 R^2 呈现接近线性的上升趋势，而 500m 和 1000m 的 R^2 先上升后趋于平缓，说明 PAG-SA 相对适合较小格网尺度下的数

据融合。2）使用建筑区类型及空间位置约束的格网 POI 特征在三个格网尺度下均取得显著的精度提升，原因是该方法有助于提升 POI 特征的精细度。3）融合基于建筑区类别约束的夜间灯光分级特征在 200m 下有一定的精度提升，而其他两个尺度下提升较小，说明格网尺度较小时夜光分级特征能更真实地反映人口分布规律。4）邻域格网 POI 特征提取在 200m 尺度下精度提升较明显，而 500m 提升较小，在 1000m 甚至出现精度下降。原因在于尺度较大时格网自身已包含相对丰富的信息，引入邻域格网特征反而增大误差。

3.2.3 核密度带宽的选取实验

考虑到各类 POI 的辐射范围[22]不同，为了获取核密度相对最优带宽，实验对比 13 种带宽下（200m~1000m 间距为 100m，2000m~5000m 间距为 1000m），单独使用每一类 POI 构建邻域格网特征时的精度。由于带宽选择受到格网大小的影响，本文实验了 200m、500m 及 1000m 三种尺度下带宽选择对精度的影响。实验表明，200m 格网下三种精度评价指标对带宽选择最敏感，且 200m 下所获取的相对最优带宽区间包含 500m 及 1000m 对应的最优区间，因此本文采用 200m 尺度格网开展带宽选取实验。实验结果如图 8 所示，图中虚线和实线分别表示引入 POI 核密度特征前后对应的精度评价指标。由图 8 可知，不同类型 POI 的相对最优带宽不同，且各类 POI 在各自相对最优带宽处均能提高精度。1）与引入核密度前对比，在某些带宽区间使用核密度后 MAE/RMSE 降低，同时 R^2 提高，说明合适带宽下各类 POI 的核密度特征对提高精度均有效。2）从 MAE/RMSE 的变化趋势上看，科研教育、住宿及金融服务这三类 POI 的整体变化幅度不大，对带宽的选择不敏感。而医院、休闲娱乐、购物及餐饮四类 POI 形成了明显的波峰。因此，选择核密度带宽时需结合具体的 POI 类型，不同 POI 类型设置同一个带宽可能会引入误差。3）从相对最优带宽上看，医院和休闲娱乐的相对最优带宽约为 3km，可能原因是医院和休闲娱乐场所的辐射距离较大，空间服务范围较广；而餐饮、购物及居民小区的相对最优带宽较小，说明这三类 POI 总体上辐射距离较小，空间服务范围较为有限。科研教育、住宿及金融服务的带宽取值对精度的影响不敏感，反映出这几类设施的辐射能力较强且空间服务范围广泛的特点。各类 POI 的相对最优带宽取值范围如表 2 所示。

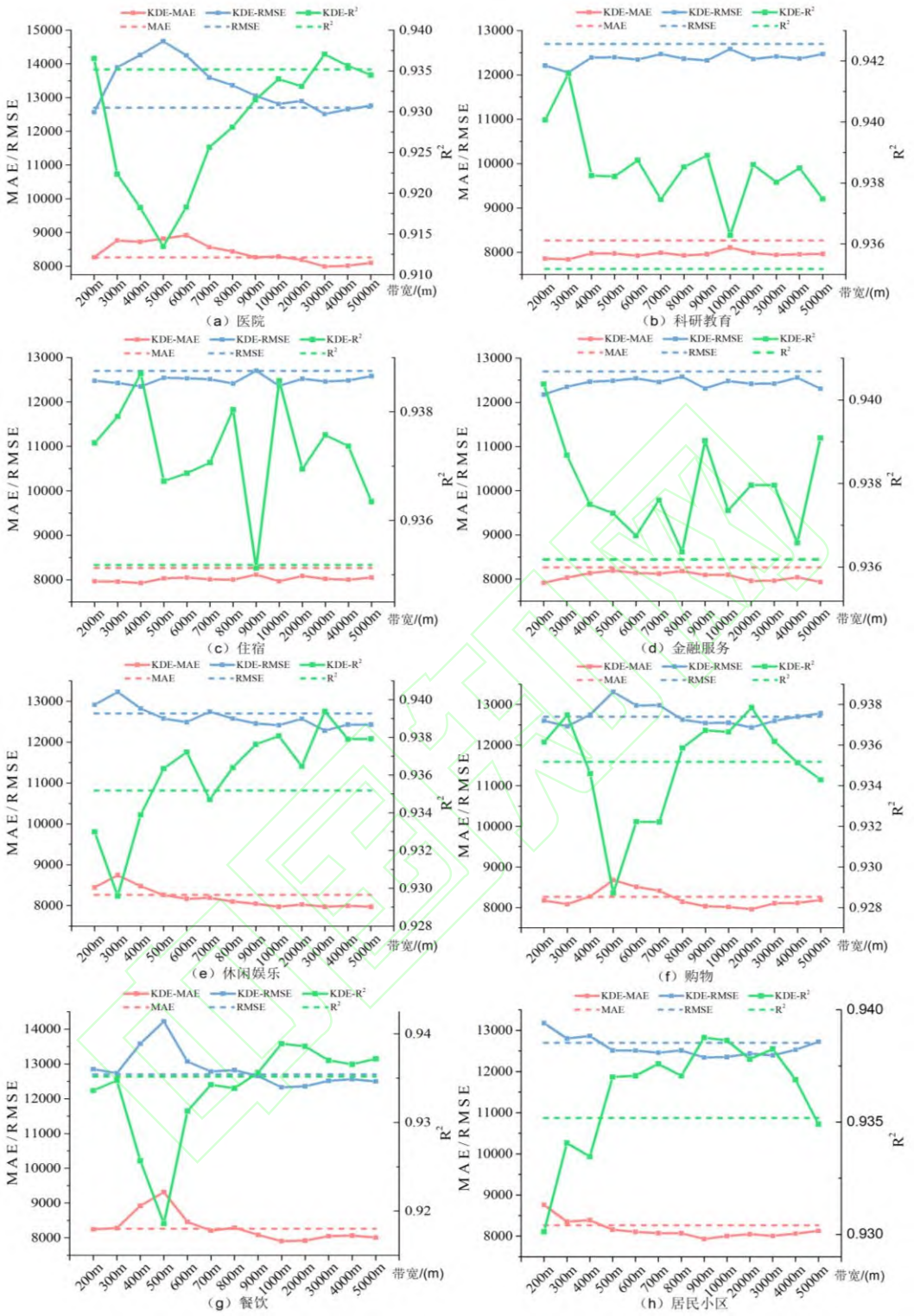


图 8. POI 带宽对估算精度的影响分析

Fig. 8. Impact of Bandwidth on Estimation Accuracy for Different POI Types

表 2. 各类 POI 相对最优核密度带宽区间及本文选用带宽

Tab.2 Relative Optimal Bandwidth Ranges and the Selected Bandwidths for Different POI Types

POI 类型	带宽区间/(km)	MAE	RMSE	R^2	选用带宽/(km)
医院	3-4	7994	12515	0.93705	4
科研教育	0.2-5	7840	12055	0.94159	5
住宿	0.2-5	7928	12349	0.93870	5

金融服务	0.2-5	7915	12179	0.94038	5
休闲娱乐	3	7973	12280	0.93939	3
餐饮	1-2	7905	12330	0.93889	2
居民小区	0.9-1	7933	12343	0.93876	1
购物	2	7961	12434	0.93784	2

3.2.4 与其他人口数据集的精度对比

PAG-SA 在 200m 格网划分下的人口空间化结果如图 9 所示，其人口分级采用自然断点法。从空间分布模式上看，武汉市人口呈现中心城区集聚且周边多核的空间结构，人口值大于 275 的格网主要分布于中心城区。

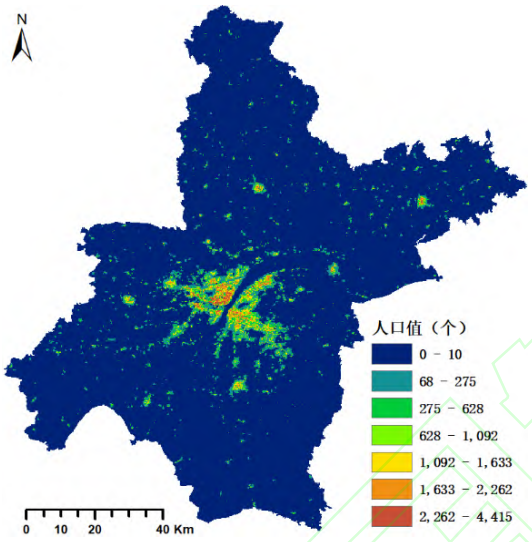
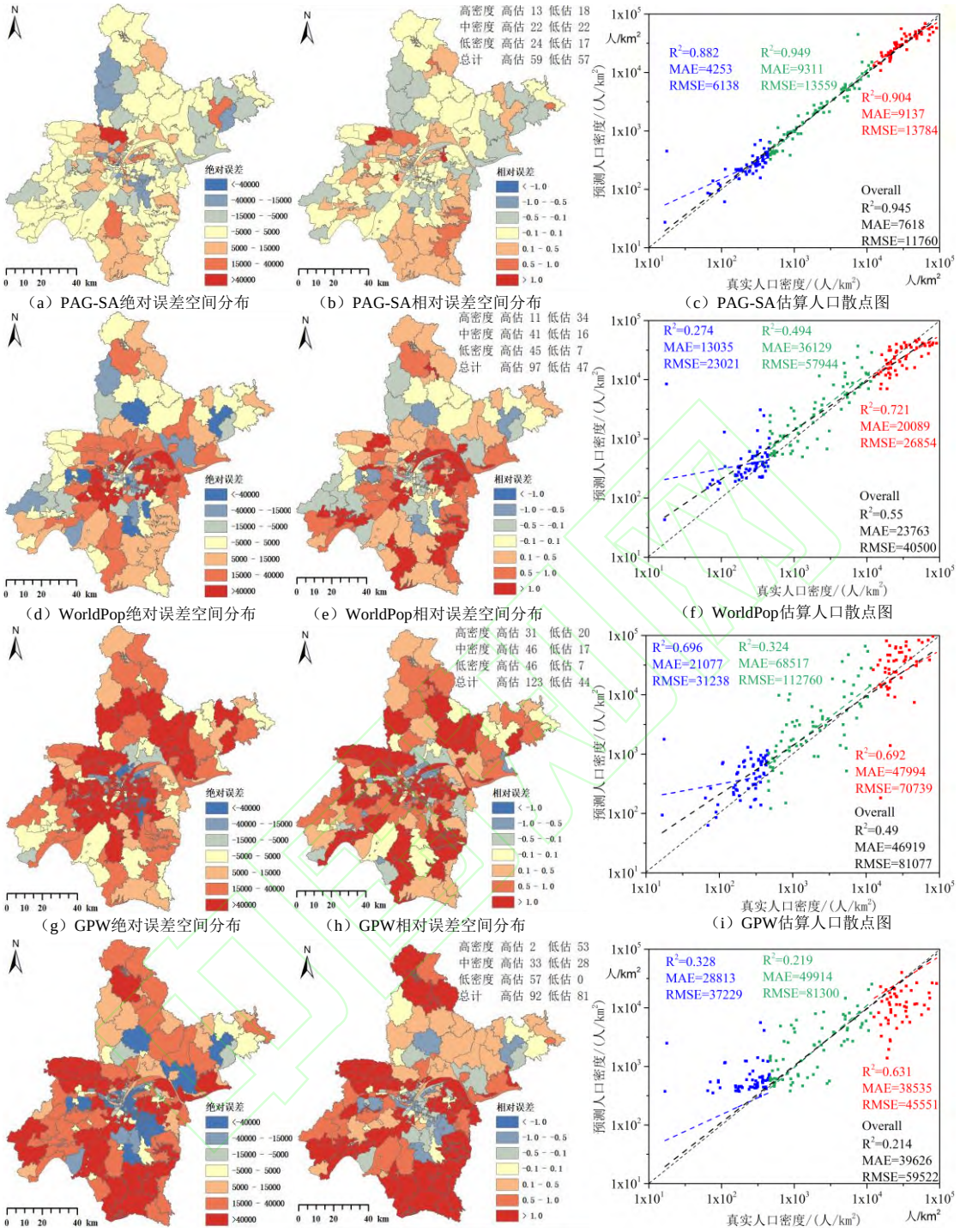


图 9 200m 格网 PAG-SA 人口空间化结果

Fig.9 Results of PAG-SA in 200m Grid Size

对于不同人口密度区域，其人口建模特征的空间分布存在差异，空间化结果亦呈现不同模式。为此，本文使用自然断点法将武汉市 185 个街道按人口密度值划分为高中低三个密度等级区域，并将 PAG-SA 与 WorldPop、GPW 及中国公里网格人口（PopulationGrid_China）数据集进行对比。街道人口密度分级结果如图 1 所示，其中低密度区 60 个街道，中密度区 69 个街道，高密度区 56 个街道。各街道误差如图 10 所示，其绝对误差为人口估算值与人口普查值之差，相对误差为绝对误差与对应街道人口普查值的比值，高低估街道定义为相对误差绝对值大于 0.1 的街道。散点图中蓝、绿、红三种颜色分别表示低密度区、中密度区及高密度区。

由图 10 可知，1）从三种评价指标的数值上看，PAG-SA 在高中低三种密度区域相对于对比人口数据集均有更小的误差。2）从拟合效果上看，PAG-SA 的散点大体集中分布于对角线两侧，而对比数据集的散点较分散且距离对角线较远，说明 PAG-SA 能够更好地拟合真实人口分布。3）从绝对误差空间分布上看，总体而言，GPW 与中国公里网格人口数据集在高中低三种密度区域均存在大量绝对误差大于 4 万的街道，WorldPop 的绝对误差主要分布于中密度区域。而 PAG-SA 除武汉市东南区域及其他零星分布街道外，均有较为明显的精度提升，显著降低了误差等级。4）从相对误差上看，PAG-SA 的高估和低估街道数量相对均衡，而 WorldPop 与中国公里网格人口数据集在低密度及中密度区域易高估，而在高密度区域易低估。原因是 WorldPop 及中国公里网格人口数据集的估算范围较广，提取的特征不够精细，因此估算结果较为平均化。GPW 数据集在高中低三个密度区域均易高估，原因是 GPW 数据集基于格网内的行政单元面积进行人口分配，未顾及其他组合因素的影响。5）PAG-SA 在武汉东南区域没有明显的精度提升，原因是该区域 POI 等设施较齐全但实际人口较少，使用武汉全部街道训练构建的模型不适用于该区域，可通过分区域训练提高精度。



（a）PAG-SA绝对误差空间分布　（b）PAG-SA相对误差空间分布　（c）PAG-SA估算人口散点图

（d）WorldPop绝对误差空间分布　（e）WorldPop相对误差空间分布　（f）WorldPop估算人口散点图

（g）GPW绝对误差空间分布　（h）GPW相对误差空间分布　（i）GPW估算人口散点图

（j）PopulationGrid_China绝对误差空间分布（k）PopulationGrid_China相对误差空间分布　（l）PopulationGrid_China估算人口散点图

图 10 PAG-SA 与 WorldPop、GPW 及中国公里网格人口数据集的精度对比

Fig.10 Comparison of Accuracy between PAG-SA and WorldPop, GPW and PopulationGrid_China

4 总结与展望

本文提出一种顾及格网属性分级与空间关联的人口空间化方法。该方法，1）基于自然断点法分别对建筑区密度和夜间灯光值分级并融合二者构建组合向量，在行政尺度使用各等级格网占比信息作为训练输入，以减小模型跨尺度误差；2）通过实验为不同类型 POI 选择合适的核密度估计带宽构造邻域格网 POI 特征；3）基于叠置分析统计建筑区类别及空间位置约束的格网

POI 特征，以便刻画多种属性不同空间聚合方式与人口密度间的关联关系，提高特征建模精细度。本文以武汉市为实验区域，通过与 WorldPop、GPW 及中国公里网格人口数据集的对比验证了方法的有效性。实验结果表明本文方法街道尺度 MAE 远小于对比数据集，在高中低人口密度区域均取得较好的拟合优度，并有效提升空间化精度。同时，本文特征提取方法中各步骤的有效性均得到验证，且 200m 格网尺度精度提升最为明显。不同类型 POI 的辐射作用范围不同，合理的核密度带宽阈值与各类 POI 的社会职能相关，需通过实验选取。

本文方法存在以下不足有待进一步研究。本文利用建筑区轮廓和 POI 间的空间关系进行数据融合，提取的特征虽然保留 POI 在不同建筑区类型内的数量信息，但未考虑空间分布信息。POI 的空间分布模式，如均匀、随机或聚集，可能对人口分布产生影响进而有助于刻画人口分布[30]，今后将尝试引入空间分布特征。同时，本文使用枚举方式选择 POI 核密度带宽，今后可研究最优带宽的自适应提取方法[31]以提升带宽选择的效率及可解释性。此外，由于依赖于建筑区轮廓及 POI 等细粒度数据，因此本文模型无法直接迁移到相关数据缺失的区域。但其特征建模方法依然具有一定适用性及参考价值，如格网属性分级、核密度估计及空间叠置分析等，在后续工作中尝试将上述建模方法迁移到不同类型的区域进行验证与分析。

参考文献

[1]Hu Yunfeng, Wang Qianqian, Liu Yue, et al. Index System and Transferring Methods to Build the National Society and Economy Grid Database[J]. *Journal of Geo-Information Science*, 2011, 13(5): 573-578(胡云锋, 王倩倩, 刘越, 等. 国家尺度社会经济数据格网化原理和方法[J]. 地球信息科学学报, 2011, 13(5): 573-578).

[2]Bai Zhongqiang, Wang Juanle, Yang Fei. Research Progress in Spatialization of Population Data[J]. *Progress in Geography*, 2013, 32(11): 1692-1702(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).

[3]Wu Shuosheng, Qiu Xiaoming, Wang Le. Population Estimation Methods in GIS and Remote Sensing: A Review[J]. *GIScience and Remote Sensing*, 2005, 42(1): 80-96.

[4]Flowerdew R, Green M. Developments in Areal Interpolation Methods and GIS[J]. *The Annals of Regional Science*, 1992, 26(1): 67-78.

[5]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.

[6]Lu Anmin, Li Chengming, Lin Zongjian, et al. Spatial Distribution of Statistical Population Data[J]. *Geomatics and Information Science of Wuhan University*, 2002, 27(3): 301-305(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(3): 301-305).

[7]Yan Qingwu, Bian Zhengfu, Zhang Ping, et al. Census Spatialization Based on Settlements Density[J]. *Geography and Geo-Information Science*, 2011, 27(5): 95-98(闫庆武, 卞正富, 张萍, 等. 基于居民点密度的人口密度空间化[J]. 地理与地理信息科学, 2011, 27(5): 95-98).

[8]Mennis J. Generating Surface Models of Population Using Dasymetric Mapping[J]. *The Professional Geographer*, 2003, 55(1): 31-42.

[9]Su M D, Lin R C, Hsieh R I, et al. Multi-layer Multi-class Dasymetric Mapping to Estimate Population Distribution[J]. *Science of the Total Environment*, 2010, 408(20): 4807-4816.

[10]Fu Haiyue, Li Manchun, Zhao Jun, et al. Summary of Grid Transformation Models of Population Data[J]. *Human Geography*, 2006, 21(3): 115-119(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, 21(3): 115-119).

[11]Dong Nan, Yang Xiaohuan, Cai Hongyan. Research Progress and Perspective on the Spatialization of Population Data[J]. *Journal of Geo-Information Science*, 2016, 18(10): 5-14(董南, 杨小唤, 蔡红艳. 人口数据空间化研究进展[J]. 地球信息科学学报, 2016, 18(10): 5-14).

[12]Zeng Chuiqing, Zhou Yi, Wang Shixin, 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.

[13]Fang Kuangnan, Wu Jianbin, Zhu Jianping, et al. A Review of Technologies on Random Forests[J]. *Statistics and Information Forum*, 2011, 26(3): 32-38(方匡南, 吴见彬, 朱建平, 等. 随机森林方法研究综述[J]. 统计与信息论坛, 2011, 26(3): 32-38).

[14]Yang Xuchao, Ye Tingting, Zhao Naizhuo, et al. Population Mapping with Multisensor Remote Sensing Images and Point-Of-Interest Data[J]. *Remote Sensing*, 2019, 11(5): 574.

[15]Liu Zhenglian, Gui Zhipeng, Wu Huayi, et al. Fine-Scale Population Spatialization by Synthesizing Building Survey Data and Point of Interest Data. *Journal of Geomatics*, 2021, DOI：10.14188/j.2095-6045.2019182 (刘正廉, 桂志鹏, 吴华意, 等. 融合建筑物与兴趣点数据的精细人口空间化研究. 测绘地理信息, 2021, DOI：10.14188/j.2095-6045.2019182).

[16]Ye Tingting, Zhao Naizhuo, Yang Xuchao, 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.

[17]Sinha P, Gaughan A E, Stevens F R, et al. Assessing the Spatial Sensitivity of a Random Forest Model: Application in Gridded Population Modeling[J]. *Computers, Environment and Urban Systems*, 2019, 75: 132-145.

[18]Robinson C, Hohman F, Dilkina B. A Deep Learning Approach for Population Estimation From Satellite Imagery[C]. *Proceedings of the 1st ACM SIGSPATIAL Workshop on Geospatial Humanities*, Los Angeles, 2017.

[19]Chen Jie, Pei Tao, Shaw Shih-Lung, et al. Fine-Grained Prediction of Urban Population Using Mobile Phone Location Data[J]. *International Journal of Geographical Information Science*, 2018, 32(1): 1-17.

[20]Zhao Song, Liu Yanxu, Zhang Rui, et al. China's Population Spatialization Based on Three Machine Learning Models[J]. *Journal of Cleaner Production*, 2020, 256: 120644.

[21]Leyk S, Gaughan A E. The Spatial Allocation of Population: A Review of Large-Scale Gridded Population Data Products and Their Fitness for Use[J]. *Earth System Science Data*, 2019, 11(3): 1385-1409.

[22]YU Wenhao, AI Tinghua, YANG Min, et al. Detecting "Hot Spots" of Facility POIs Based on Kernel Density Estimation and Spatial Autocorrelation Technique[J]. Geomatics and Information Science of Wuhan University, 2016, 41(2): 221-227(禹文豪, 艾廷华, 杨敏, 等. 利用核密度与空间自相关进行城市设施兴趣点分布热点探测[J]. 武汉大学学报(信息科学版), 2016, 41(2): 221-227).

[23]YANG Xiping, FANG Zhixiang, ZHAO Zhiyuan, et al. Analyzing Space-Time Variation of Urban Human Stay Using Kernel Density Estimation by Considering Spatial Distribution of Mobile Phone Towers[J]. Geomatics and Information Science of Wuhan University, 2017, 42(1): 49-55(杨喜平, 方志祥, 赵志远, 等. 顾及手机基站分布的核密度估计城市人群时空停留分布[J]. 武汉大学学报(信息科学版), 2017, 42(1): 49-55).

[24]Chen Qing, Hou Xiyong. An Improved Population Spatialization Model by Combining Land Use Data and DMSP/OLS Data [J]. *Journal of Geo-Information Science*, 2015, 17(11): 1370-1377(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).

[25]Yu Bailang, Lian Ting, Huang Yixiu, et al. Integration of Nighttime Light Remote Sensing Images and Taxi GPS Tracking Data for Population Surface Enhancement[J]. *International Journal of Geographical Information Science*, 2019, 33(4): 687-706.

[26]Langford M. Obtaining Population Estimates in Non-census Reporting Zones: An Evaluation of the 3-class Dasymetric Method[J]. *Computers Environment and Urban Systems*, 2006, 30(2): 161-180.

[27]Guo Yuchen, Huang Jinchuan, Lin Haoxi. Spatialization of China's Population Data Based on Multisource Data[J]. Remote Sensing Technology and Application, 2020, 35(1): 219-232(郭雨臣, 黄金川, 林浩曦. 多源数据融合的中国人口数据空间化研究

[J]. 遥感技术与应用, 2020, 35(1): 219-232).
[28]Chainey S P. Examining the Influence of Cell Size and Bandwidth Size on Kernel Density Estimation Crime Hotspot Maps for Predicting Spatial Patterns of Crime[J]. *Bulletin of the Geographical Society of Liege*, 2013, 60(1): 7-19.
[29]Lin Yupin, Chu Hone-Jay, Wu Chenfa, et al. Hotspot Analysis of Spatial Environmental Pollutants Using Kernel Density Estimation and Geostatistical Techniques[J]. *International Journal of Environmental Research and Public Health*, 2011, 8(1): 75-88.
[30]Du Guoming, Zhang Shuwen, Zhang Youquan. Analyzing Spatial Auto-Correlation of Population Distribution: A case of Shenyang city[J]. *Geographical Research*, 2007, 26(2): 383-390.
[31]Yuan Kunxiaojia, Cheng Xiaoqiang, Gui Zhipeng, et al. A Quad-Tree-Based Fast and Adaptive Kernel Density Estimation Algorithm for Heat-Map Generation[J]. *International Journal of Geographical Information Science*, 2019, 33(12): 2455-2476.

Population Spatialization by Considering Pixel-level Attribute Grading and Spatial Association

WU Jinghang[1], GUI Zhipeng[1,2], SHEN Li[1], WU Huayi[2,3], LIU Hongbo[4], LI Rui[3], MEI Yu'ao[1], PENG Dehua[3]

1 School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China

2 Collaborative Innovation Center of Geospatial Technology, Wuhan 430079, China

3 State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China

4 Chongqing Geomatics and Remote Sensing Center, Chongqing, 401147

Abstract: Existing population spatialization methods mainly use administrative-unit-level data to train regression model, and transfer it to grid cell-level to achieve population allocation. However, the significant scale difference between the analytical units in training and estimation leads to the issues of cross-scale model transfer. Meanwhile, only the attributes of current cell are considered in cell-level feature modeling, which causes the innate spatial association between cells to be eliminated and cells to be isolated. Therefore, this paper proposes a novel population spatialization based on random forest by considering pixel-level attribute grading and spatial association (PAG-SA). In the cell-level feature modeling, we firstly constructs the night light grading features embedded with building category constraints based on natural breaks, and counts the grid proportion of each grading level at the administrative-unit-level as the training input to reduce the cross scale error; secondly, the influence and distance attenuation of neighborhood point of interests (POIs) upon the current cell is modelled by using kernel density estimation; thirdly, based on overlay analysis, the numbers of POIs in the contours of different building types are counted to improve the precision of feature modeling. To verify the effectiveness of the proposed method, we selected Wuhan city as the experimental area and compared its spatialization accuracy with the datasets of WorldPop, GPW and PopulationGrid_China at street scale. The results show that the mean absolute error of PAG-SA is only 1/6-1/3 of the comparison datasets. In addition, the influence of feature composition, grid size and kernel density bandwidth on the accuracy is also discussed.

Keywords: population spatialization; random forest; multi-source data fusion; cross-scale issues; kernel density estimation; overlay analysis

First author: WU Jinghang, master, specializes in population spatialization. E-mail: wyw1294@whu.edu.cn
Corresponding author: GUI Zhipeng, PhD, associate professor. E-mail: zhipeng.gui@whu.edu.cn
Foundation support: The National Key Research and Development Program of China (2018YFC0809806, 2017YFB0503704); The National Natural Science Foundation of China (41971349, U20A2091, 42090010).