

引文格式: 成方龙,赵冠伟,杨木壮,等.集成地理探测器与随机森林模型的城市人口分布格网模拟[J].测绘通报,2020(1):98-103.DOI:10.13474/j.cnki.11-2246.2020.0021.

集成地理探测器与随机森林模型的城市人口分布格网模拟

成方龙¹,赵冠伟^{1,2},杨木壮^{1,2},刘月亮¹,李芳¹

(1. 广州大学地理科学学院,广东 广州 510006; 2. 广州大学国土资源与海岸带研究所,广东 广州 510006)

摘要: 精细尺度的城市人口分布信息是城市资源配置和综合管理的重要依据。本文以广州市越秀区、荔湾区、天河区、海珠区、白云区及黄埔区作为研究区域,基于人口统计、夜间灯光、兴趣点及土地利用等多源数据,利用地理探测器识别人口分布的影响因子,运用随机森林模型开展人口分布空间格网模拟研究。研究结果表明,与传统的相关分析相比,地理探测器能够更为准确地识别人口空间分布的重要影响因子。基于随机森林模型的人口分布格网模拟结果与街道(镇)实际人口的相关系数为0.774,平均相对误差约为30%。相比基于线性回归模型的模拟结果,随机森林模型的精度有明显提高。

关键词: 人口分布; 格网; 模拟; 随机森林; 地理探测器

中图分类号: P208

文献标识码: A

文章编号: 0494-0911(2020)01-0098-06

Simulation of urban population distribution grid by integrating the geodetector and random forest model

CHENG Fanglong¹, ZHAO Guanwei^{1,2}, YANG Muzhuang^{1,2}, LIU Yueliang¹, LI Fang¹

(1. School of Geographical Science, Guangzhou University, Guangzhou 510006, China; 2. Institute of Land Resources and Coastal Zone, Guangzhou University, Guangzhou 510006, China)

Abstract: Fine population distribution is important to urban resource allocation and management. In this paper, we take the Yuexiu District, Liwan District, Tianhe District, Haizhu District, Baiyun District and Huangpu District of Guangzhou city as the research areas, and base on multi-source data such as demography, night lighting, interest points and land use, using the geodetector to identify the influencing factors of population distribution, and simulate the population distribution grid by using random forest model. The results show that compared with the traditional correlation analysis, the geodetector can identify the important factors of spatial distribution of population more accurately. The correlation coefficient between the results of population distribution grid simulation based on random forest model and the actual population of streets (towns) is 0.774, with an average relative error of about 30%. Compared with the simulation results based on linear regression model, the accuracy of the stochastic forest model is significantly improved.

Key words: population distribution; grid; simulation; random forest; geodetector

人口空间分布是人口学、社会学、统计学的主要研究课题之一,更是人口地理学最主要的核心内容^[1]。如何有效、精确地获取人口在空间上的分布数量及特征,对协调人地关系、自然灾害的预防、应急响应、资源合理利用和环境保护都有巨大的现实意义。然而,传统的人口普查数据存在时空分辨率低、统计周期长、更新速度慢等缺点,导致不能准确、真实地反映人口在空间上的分布规律,更不能揭示小尺度范围内人口数量及密度的差异。尤其是当人口统计数据与多源数据进行融合或叠加分析时,往往还会出现行政区划边界与自然单元边界不吻合、

数据空间精度降低、“可变元”等一系列问题^[2]。人口统计数据空间化可以用格网的形式准确直观地表达人口在空间上的分布状况^[3],是解决上述问题的有效手段,越来越受到学者的重视。

人口空间化的研究起源可以追溯到1930年,美国社会地理学家Wright J.K.提出的Dasymetric方法将人口分布与土地利用类型相结合生成了马萨诸塞州科德角的人口空间化地图^[4]。目前,人口空间化模拟研究的方法较多,主要有空间插值法^[5-13]、遥感像元特征反演法^[14-16]、夜间灯光建模法^[17-26]、土地利用类型法^[27-31]和多源地理信息融合法^[32-36]等。近年来,随着计算机科学、遥感和地理信息技术的发展

收稿日期: 2019-04-08; 修回日期: 2019-10-19

基金项目: 国家自然科学基金(41671175; 41671430); 广东省自然科学基金(2017A030313240); 广东省高等学校优秀青年教师培养项目(YQ2015127)

作者简介: 成方龙(1992—),男,硕士生,主要研究方向为地理时空数据分析。E-mail: 784628444@qq.com

通信作者: 赵冠伟。E-mail: zgwl1124@163.com

展,社交媒体、兴趣点、夜间灯光等多源地理信息数据的不断涌现,数理统计、空间分析、机器学习等方法不断改进,为人口空间化模拟研究提供了丰富的数据来源和先进的技术手段。

地理探测器由王劲峰等研发,是探测空间分异性及揭示其背后驱动因子的一种新的统计学方法,该方法无线性假设,具有优雅的形式和明确的物理含义^[67]。地理探测器模型在应用时没有过多的假设条件,可以克服统计方法处理变量的局限性,因此被广泛应用于社会经济因素和自然环境因素的影响机理研究。随机森林模型是一种基于分类树的机器学习算法,最早于 2001 年由 Leo Breiman 和 Cutlery Adele 提出,由多棵分类和回归树构成的^[68]。与多元线性回归、空间自相关、地理加权回归、贝叶斯回归和人工神经网络等模型方法相比,随机森林算法具有以下优点:①在进行回归分析过程中,不需要考虑一般线性回归分析所面临的多元共线性问题,也不需要做变量选择;②与其他机器学习方法相比,随机森林算法在降低了运算量的同时也提高了模型预测精度;③随机森林算法在每棵树的生长过程中使用的是 bootstrap 采样方法,这使得每棵树只选择部分样本及部分特征,在一定程度上避免过拟合;④由于随机森林每棵树在随机选择样本的同时随机选择特征,这使得模型对异常值和噪声具有很高的容忍度^[69];⑤随机森林算法运算速度快,可以较好地处理高维度数据,并且不需要做特征选择。由于随机森林算法优点明显,在人口模拟^[65-66]、生态学、经济学、医学等领域得到广泛应用。综上所述,本文以广州市中心六区为研究区,综合兴趣点、夜间灯光、土地利用、人口统计等多源地理信息数据,利用地理探测器准确识别人口分布的影响因子,运用随机森林模型开展城市人口分布格网模拟研究,以期城市管理提供精细尺度的人口数据。

1 研究区概况与数据来源

1.1 研究区概况

广州市是广东省省会城市,粤港澳大湾区的核心城市之一,有中国“南大门”之称。广州市行政面积为 7 434.4 km²,截至 2013 年底常住人口为 1 292.68 万人。为确保与 2013 年人口统计数据口径的一致性,选择行政区划调整之前的广州市天河区、海珠区、荔湾区、越秀区、白云区以及黄埔区(不含原萝岗区)等 6 个区,共计 106 个街道作为研究区域。截至 2013 年底,研究区总面积为 1 198.14 km²,常住人口总数为 783.02 万人。

1.2 数据来源与预处理

研究需要的数据包括:行政区划数据、人口统计数据、土地利用数据、数字高程模型、路网数据、POI 数据、房屋建筑区、夜间灯光数据等。行政区划和路网数据根据 2014 年高德电子地图进行数字化得到,其中路网包括高速公路、省道、市区道路和县道。人口数据由广州市公安局提供,包含 2013 年研究区各个街道(镇)的人口数量及各个区的人口数量总和。数字高程模型数据(GDEM V2)下载自地理空间数据云平台,空间分辨率为 30 m,空间参考为 GCS_WGS_1984。土地利用数据来自中国科学院资源与环境数据中心,土地利用类型包括:耕地,林地,草地,水域,城乡、工矿、居民用地和未利用土地共 6 个大类和 25 个小类,空间分辨率为 30 m。POI 数据主要通过 Python 语言调用百度地图 API 接口进行爬取,包括餐饮设施、公共设施、公司企业、购物设施、交通设施、金融保险、科教文化设施、商务住宅、生活服务设施、体育休闲设施、医疗服务设施、政府机关及社会团体和住宿设施共 13 类,采集时间为 2018 年 7 月。夜间灯光数据从美国国家海洋局气象中心网站下载,空间参考为 GCS_WGS_1984。房屋建筑区数据是参考全国地理国情普查成果进行绘制。房价数据是利用 Python 语言编程从链家网进行爬取,数据属性包括小区名称、房价均价、所在区域、地址、建筑年代、建筑类型、经纬度等。为方便后文对数据进行空间分析,将所涉及的空间数据统一为 Albers 投影坐标系,参考椭球体为 Krasovsky_1940 椭球体。

2 基于多元线性回归模型的人口分布格网模拟

2.1 格网尺度确定

已有研究表明,人口数据库的精度不仅与建模方法、源数据的空间分辨率、人口数据精度有关,也与所选取的格网大小有关。因此,选择适宜的格网大小,对于更好地展现人口分布空间分异特征,提高人口数据空间数据集质量具有重要的影响。通过文献梳理发现,确定适宜格网大小的方法主要有两类:基于源数据的适宜格网确定方法和基于人口数据格网结果表达度的方法。由于在开展模拟之前需要确定适宜的格网尺度,因此本文主要通过分析数据源来确定适宜格网的大小。在确定格网尺度时,以各街道(镇)的大小作为参考标准,综合文献[39—40]的研究成果,为避免面积最小的街道(镇)全部落入同一格网中,将格网大小确定为面积最小街道(镇)的 10%。在本文研究中面积最小的街道为沙面街

道,其面积约为 308 348 km²,当格网大小为 150 m 时,格网面积与最小街道(镇)面积的 10% 最为接近。因此本文将格网大小设置为 150 m。

2.2 基于地理探测器的人口空间分布影响因素识别

作为探测空间分异性及揭示其背后驱动因子的一种新的统计学方法,地理探测器在度量空间分异性、探测解释因子、分析变量之间的交互关系方面有着重要的突出作用,已经广泛的被应用于社会、经济及自然环境因素的影响机理研究中^[8],并取得了丰硕的研究成果。在前人的研究成果上,结合自身研究的需求,在自然因素方面选取了土地利用及海拔高度两个指标;在社会经济因素方面选取了兴趣点数据(包括 13 个类别)、路网分布、夜间灯光强度、小区房价、房屋建筑面积等 5 个指标,共计 7 个指标作为研究区影响人口空间分布的重要因素。对选取的数据进行离散化处理后,将因变量人口密度值 Y 和自变量人口空间分布影响因子 X 输入到 GeoDetector2015 软件中,计算影响因子 X 对因变量 Y 的解释力。由计算结果可知,所选取的人口空间分布影响因子除了水域指数和草地指数外,其他影响因子均通过 0.05 的显著性检验。

2.3 回归模型构建

基于土地利用数据和夜间灯光数据,运用多元线性回归模型实现人口数据空间格网模拟是经典的人口分布格网化研究方法。具体模型构建为

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

式中, P_i 为研究区第 i 个街道(镇)的人口数据; a_j 为第 j 类土地利用类型的人口分布初始系数; S_{ij} 为第 i 个街道(镇) j 类土地利用指数; n 为所选取的土地利用类型种类; k 为夜间灯光强度系数; M_{nl} 为各街道平均夜间灯光强度值; b 为常数项。

以研究区各街道(镇)人口密度作为因变量,分别选取了城镇用地指数(X_1)、耕地指数(X_2)、林地指数(X_3)、农村用地指数(X_4)、其他建设用地指数(X_5)及平均夜间灯光强度(X_6)作为自变量,将数据导入 SPSS 22.0 软件中进行多元线性回归分析,得到回归模型,其线性相关系数 R 为 0.764, R^2 为 0.584。具体模型为

$$Y = 70.098X_1 - 3.678.228X_2 - 176.539X_3 - 2.052X_4 - 121.901X_5 + 702.048X_6 \quad (2)$$

2.4 人口分布格网模拟及精度验证

基于式(2)的多元线性回归方程,调用 ARCGIS 中的栅格计算器,生成 150 m 分辨率的研究区人口

分布格网数据,结果如图 1 所示。

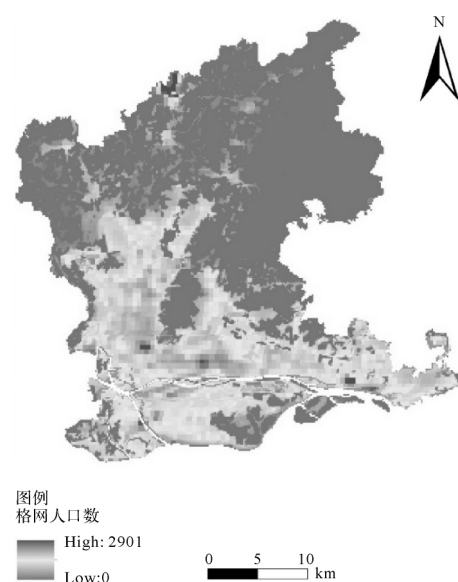


图 1 基于线性回归模型的人口分布格网模拟结果

由图 1 可知,基于线性回归模型的人口分布格网模拟结果并不理想。整体上,人口数量由城市中心向四周逐渐下降的扩散式特征。同时,由于模拟结果受到建设用地空间格局及夜间灯光强度的影响过大,容易导致建设用地区域上的人口数量过度高估,人口分布的空间分异特征不明显。另外,人和镇北部由于靠近广州白云国际机场、工业区分布集中,导致人口估算出现明显的异常值,与实际情况不符合。为了进一步验证模拟结果精度,将街道(镇)模拟人口与统计人口进行相关系数计算及线性拟合发现:相关系数 R 为 0.06、拟合优度 R^2 仅为 0.003 9,绝大部分街道(镇)人口模拟结果与实际人口数量偏差较大,模拟效果不佳。

3 基于随机森林模型的人口数据空间化

为了减少模拟误差及提高模拟精度,对人口空间分布影响因子的选取进行优化,并采用随机森林模型,实现研究区人口数据空间化。

3.1 构建模型特征库

人口空间分布是自然因素和社会经济因素综合作用的结果。考虑线性回归模型的模拟结果及误差影响因素,重新选择土地类型中的耕地指数和林地指数、夜间灯光强度、海拔高度、13 类 POI 影响强度、房屋建筑面积、路网密度、房价均价共 20 个指标,基于街道(镇)尺度,构建随机森林模型特征数据库。由于影响因子量纲不一致,不能用于直接建模,还需对数据进行标准化处理,消除量纲不统一和

极值的影响。

3.2 随机森林模型构建及训练

以研究区范围内各街道人口密度值作为因变量,耕地指数、林地指数、夜间灯光强度、海拔高度、13 类 POI 影响强度、房屋建筑面积、路网密度及房价均价共 20 个特征数为自变量,利用 Python 语言调用 scikit-learn 库,对随机森林模型进行训练。在训练的过程中,采用 bootstrap 有放回的采样方法,即训练过程中约 63.2% 原始样本被作为训练数据,剩下的约 36.8% 样本数据组成袋外数据(out-of-bag, OOB)。同时,使用 OOB 样本来评估模型的泛化性能,即模型的精度。

随机森林模型构建过程中需要确定 2 个重要参数,包括子树最大深度(max_depth)和子树数量(n_estimators)。首先,在子树数量较大的前提下(n_estimators = 500),测试子树深度不同取值时随机森林模型的泛化性能。由图 2 可知,随着子树深度的增加,模型泛化性能逐渐增强,当子树深度等于 6 时,模型泛化性能达到最佳。其次,测试当子树深度为 6 的时候,子树数量与模型泛化性能的关系。由图 3 可知,随着子树数量的增加,模型泛化性能波动较大;当子树数量等于 320 时,模型泛化性能达到最优;随后,当子树数量大于 600 之后模型泛化性能较为稳定,且模型效果较好。综上所述,将子树最大深度设置为 6,子树数量设置为 600,搭建随机森林模型。

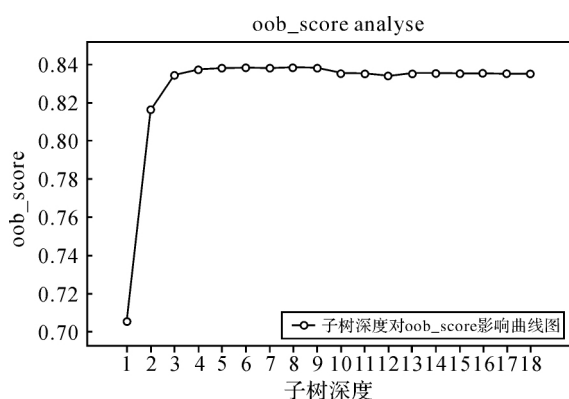


图 2 模型泛化性能与子树深度之间的关系

3.3 模型结果及精度评价

基于上文设置的参数,以研究区街道(镇)为基本单位进行随机森林模型训练,并将生成的随机森林模型应用到每个格网中,预测每个格网的人口密度。将得到的格网人口密度值与格网面积相乘,最后生成研究区人口分布格网(如图 4 所示)。可以看出,研究区人口空间分布呈明显的“核心—边缘”

格局,即人口高度集中在研究区中心地带,而边缘地区人口分布数量较少。其中,越秀区、荔湾区东北部等地区的格网人口数量最多,人口分布集聚效应明显。

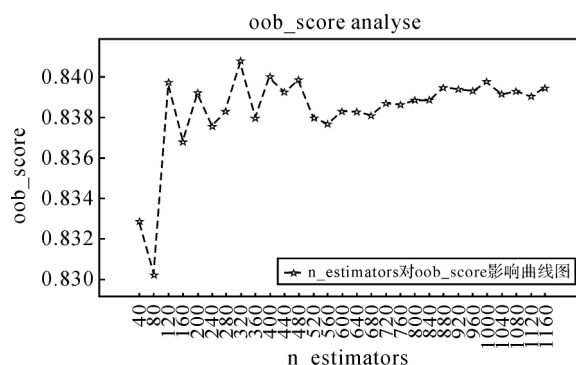


图 3 模型泛化性能与子树数量之间的关系

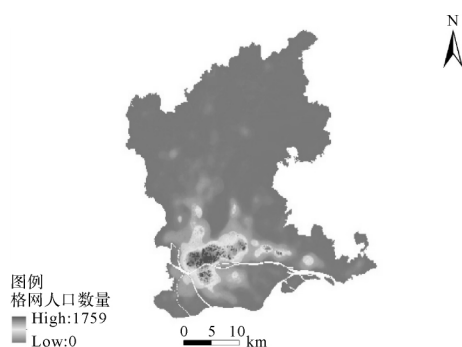


图 4 基于随机森林模型的人口分布格网模拟结果

将随机森林模型生成的街道(镇)模拟人口与实际统计人口进行相关分析和线性拟合可知:模型的相关系数 R 为 0.774,相关性较强;拟合优度为 0.599。研究区内所有街道(镇)的平均相对误差约为 30%,与回归模型相比模拟精度有了显著提升。

4 结论与展望

4.1 主要结论

(1) 利用地理探测器模型来识别人口空间分布的主要影响因子,结果表明除草地指数和水域指数外,其他因子均通过 0.05 的显著性检验,其中政府机关和社会团体对人口空间分布的解释力 q 值最大,公共设施次之,其他建设用地指数解释力最小。

(2) 基于地理探测器模型识别人口空间分布影响因子,并利用随机森林模型生成的研究区人口分布格网模拟结果与街道(镇)实际人口的相关系数为 0.8,平均相对误差为 21.98%。相比多元线性回归模拟结果,模拟精度有明显的提高,与实际人口分

布特征更为符合。

4.2 不足与展望

由于受个人能力和精力的限制,本文还存在以下不足:①研究过程中涉及的数据量和类型较多,在数据转换过程中不可避免产生误差,进而影响模拟结果。②数据精度有待提高。对研究区而言,土地利用数据的空间分辨率为30 m,精度不高,尤其在研究区的核心地带,建设用地集中连片的分类结果导致模拟结果无法反映其原本存在的特征;此外,NPP/VIIRS 夜间灯光数据的空间分辨率约为500 m,精度虽比 DMSP/OLS 夜间灯光数据高,但仍有改进的空间。今后可考虑利用珞珈一号夜间灯光数据源以提升分辨率。③数据时序不一致。研究所选用的人口统计数据(2013年)与土地利用(2015年)、夜间灯光(2016年)、POI(2018年)等其他数据存在时间上的不匹配,从而影响模拟结果。④人口空间分布的影响因子选取依旧存在改进的余地。当前所选取的影响因子中,POI数据成为了人口数据空间化建模的主导因子。同时,本文在研究过程中,只考虑了每类POI数据对人口空间分布的影响,并没有从综合的角度考虑多种POI共同作用下,对人口分布的影响。上述不足都有待在未来的研究中进一步改进。

参考文献:

- [1] 吴忠观.人口学[M].重庆:重庆大学出版社,1993.
- [2] 柏中强,王卷乐,姜浩,等.基于多源信息的人口分布格网化方法研究[J].地球信息科学学报,2015,17(6):653-660.
- [3] 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.
- [4] LI G, WENG Q. Using Landsat ETM+Imagery to measure population density in indianapolis, Indiana, USA[J]. Photogrammetric Engineering & Remote Sensing, 2005, 71(8):947-958.
- [5] 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.
- [6] 刘建军,李春来.基于遥感和GIS的巢湖流域人口信息提取[J].科学通报,2002,47(23):1835-1837.
- [7] 潘顺.长三角人口数据格网化及其人口空间分布特征分析[D].上海:上海师范大学,2016.
- [8] 毛莹莹.城市人口数据空间化研究[D].福州:福建师范大学,2016.
- [9] 刘焕金.基于多源数据的太湖流域人口空间化研究[D].南京:南京农业大学,2012.
- [10] WRIGHT J K. A method of mapping densities of population[J].The Geographical Review, 1936:103-110.
- [11] MARTIN D. Mapping population data from zone centroid locations[J]. Transactions of the Institute of British Geographers, 1989, 14(1):90-97.
- [12] FLOWERDEW R, GREEN M. Developments in areal interpolation methods and GIS[J]. The Annals of Regional Science, 1992, 26(1):67-78.
- [13] GOODCHILD M F, ANSELIN L, DEICHMANN U. A framework for the areal interpolation of socioeconomic data[J]. Environment and Planning A, 1993, 25:383-397.
- [14] LISAKA J, HEGEDUA E. Population estimates from Landsat imagery[J]. Remote Sensing of Environment, 1982, 12:259-272.
- [15] HARVEY J T. Population estimation models based on individual TM pixels[J]. Photogrammetric Engineering & Remote Sensing, 2002, 68(11):118-1192.
- [16] 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(4):36-45.
- [7] 杨眉,王世新,周艺,等.DMSP/OLS 夜间灯光数据应用研究综述[J].遥感技术与应用,2011,26(1):45-51.
- [8] 王鹤饶,郑新奇,袁涛.DMSP/OLS 数据应用研究综述[J].地理科学进展,2012,10(1):11-18.
- [9] ELVIDGE C D, BAUGH K E, KIG E A, et al. Relation between satellite observed visible-near infrared emissions, population, economic activity and electric power consumption[J]. International Journal of Remote Sensing, 1997, 18(6):1373-1379.
- [20] SUTTON K, ROBERTS D, ELVIDGE C. Census from heaven: an estimate of the global human population using night-time satellite imagery[J]. International Journal of Remote Sensing, 2001, 22(16):3061-3076.
- [21] BRIGGS D J, GULLIVER J, FECHT D, et al. Dasymeric modelling of small-area population distribution using land cover and light emissions data[J]. Remote Sensing of Environment, 2007, 108(4):451-466.
- [22] 卓莉,陈晋,史培军,等.基于夜间灯光数据的中国人口密度模拟[J].地理学报,2005,60(2):266-276.
- [23] 曹丽琴,李平湘,张良培.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,2009,2009(1):83-87.
- [24] 高义,王辉,王培涛,等.基于人口普查与多源夜间灯光数据的海岸带人口空间化分析[J].资源科学,2013,35(12):2517-2523.
- [25] 李翔,陈振杰,吴洁璇,等.基于夜间灯光数据和空间

- 回归模型的城市常住人口格网化方法研究 [J]. 地球信息科学学报, 2017, 19(10): 1298-1305.
- [26] 高倩, 阿里木江·卡斯木. 基于 DMSP/OLS 夜间灯光数据的天山北坡城市群人口分布空间模拟 [J]. 西北人口, 2017, 38(3): 113-120.
- [27] WU C, MURRAY A T. A Cokriging method for estimating population density in urban areas [J]. Computers, Environment and Urban Systems, 2005, 29: 558-579.
- [28] GAUGHAN A E, STEVENS F R, LINARD C, et al. High resolution population distribution maps for Southeast Asia in 2010 and 2015 [J]. Plos One, 2013, 8(2): e55882.
- [29] 颜清梅. 基于格网的人口空间分布特征分析 [D]. 成都: 四川师范大学, 2013.
- [30] 吴桂平, 曾永年, 邹滨, 等. 基于 GIS 的区域人口密度空间分布模拟——以张家界市永定区为例 [J]. 测绘科学, 2009, 34(2): 237-240.
- [31] 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.
- [32] 林文棋, 陈会宴, 谢盼, 等. 基于多源数据的北京市朝阳区人口时空格局评估与预测 [J]. 地球信息科学学报, 2018, 20(10): 1467-1477.
- [33] 许玲丽, 颜梅春. 多源信息多尺度视角的南京市街道级人口模拟研究 [J]. 测绘通报, 2019(2): 71-75.
- [34] 张晓瑞, 华茜, 程志刚. 基于空间句法和 LBS 大数据的合肥市人口分布空间格局研究 [J]. 地理科学, 2018, 38(11): 1809-1816.
- [35] 黄安, 许月卿, 孙丕苓, 等. 基于多源数据人口分布空间化研究——以河北省张家口市为例 [J]. 资源科学, 2017, 39(11): 2186-2196.
- [36] 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.
- [37] 王劲峰, 徐成动. 地理探测器: 原理与展望 [J]. 地理学报, 2017, 72(1): 116-134.
- [38] 李欣海. 随机森林模型在分类与回归分析中的应用 [J]. 应用昆虫学报, 2013, 50(4): 1190-1197.
- [39] 林飞娜, 赵文吉, 张萍. 基于 GIS 的城市人口空间分布模型与应用——以长春市区为例 [J]. 测绘科学, 2008, 33(4): 163-165.
- [40] 徐凤娇. 人口数据空间化格网尺度适宜性研究——以安徽省为例 [D]. 吉林: 延边大学, 2017.

(责任编辑: 杨瑞芳)

(上接第 71 页)

- [5] 王海起, 王劲峰. 空间数据挖掘技术研究进展 [J]. 地理与地理信息科学, 2005, 21(4): 6-10.
- [6] 李德仁, 王树良, 史文中, 等. 论空间数据挖掘和知识发现 [J]. 武汉大学学报(信息科学版), 2001, 26(6): 491-499.
- [7] 高凯. 实战 Elasticsearch, Logstash, Kibana: 分布式大数据搜索与日志挖掘及可视化解决方案 [M]. 北京: 清华大学出版社, 2015.
- [8] 李祥池. 基于 ELK 和 Spark Streaming 的日志分析系统设计与实现 [J]. 电子科学技术, 2015, 2(6): 674-678.
- [9] 王裕辰. 基于 ELK Stack 的实时日志分析系统的设计与实现 [D]. 北京: 北京邮电大学, 2018.
- [10] 姜康, 冯钧, 唐志贤, 等. 基于 Elasticsearch 的元数据搜索与共享平台 [J]. 计算机与现代化, 2015(2): 117-121, 126.
- [11] 白俊, 郭贺彬. 基于 Elasticsearch 的大日志实时搜索的软件集成方案研究 [J]. 吉林师范大学学报(自然科学版), 2014, 35(1): 85-87.
- [12] 朱建平, 章贵军, 刘晓藏. 大数据时代下数据分析理念的辨析 [J]. 统计研究, 2014, 31(2): 10-17.
- [13] 王元卓, 靳小龙, 程学旗. 网络大数据: 现状与展望 [J]. 计算机学报, 2013, 36(6): 1-5.
- [14] 王凯, 曹建成, 王乃生, 等. Hadoop 支持下的地理信息大数据处理技术初探 [J]. 测绘通报, 2015(10): 114-117.
- [15] 王净. 空间数据挖掘和知识发现与地理可视化的集成 [J]. 测绘通报, 2005(12): 20-23.
- [16] 薛明, 刘春, 肖学年. 基于 WebGIS 的复杂城市交通网络信息服务 [J]. 测绘通报, 2011(12): 64-68.

(责任编辑: 马保卫)