引用格式：王晓洁,王卷乐,薛润生.基于普查和手机定位数据的乡镇尺度人口空间化方法研究[J].地球信息科学学报,2020,22(5):1095-1105. [Wang X J, Wang J L, Xue R S. Research on population spatialization method in township scale based on census and mobile location data[J]. Journal of Geo-information Science, 2020,22(5):1095-1105.] **DOI:**10.12082/dqxxkx.2020.190806

基于普查和手机定位数据的乡镇尺度人口空间化方法研究

王晓洁[1,2],王卷乐[2,3*],薛润生[4]

1. 山东理工大学建筑工程学院,淄博 255049; 2. 中国科学院地理科学与资源研究所资源与环境信息系统国家重点实验室,北京 100101; 3. 江苏省地理信息资源开发与利用协同创新平台,南京 210023; 4. 山东科技大学,青岛 266590

Research on Population Spatialization Method in Township Scale based on Census and Mobile Location Data

WANG Xiaojie[1,2], WANG Juanle[2,3*], XUE Runsheng[4]

1. School of Civil and Architectural Engineering, Shandong University of Technology, Zibo 255049, China; 2. State Key Laboratory of Resources and Environment Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China; 3. Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing 210023, China; 4. Shandong University of Science and Technology, Qingdao 266590, China

Abstract: Quantifying the spatial distribution of population is a basis and hot issue in population geography researches. At present, there are large differences between different scales of spatialized population data in the world, because of various production methods, data sources, etc. This leads to the inconsistency of population spatialization, especially the 1 km-scale data which is widely needed. This paper takes Beijing-Tianjin- Hebei region as study area to build a population spatialized model at 1 km spatial resolution, based on multi-source data such as the township scale census data in 2000 and available mobile location data. The statistic population distribution weight (p) is calculated using the light projection method. Preliminary population spatialization is calculated using the area-weighted method, and the preliminary data is further modified by the exponential smoothing algorithm. Finally, the population spatialization dataset (PJ2000) with 1 km resolution in Beijing-Tianjin-Hebei region is obtained. This dataset integrates the small-scale characteristics of the township street demographic data and the advantages of mobile phone location data. The PJ2000 dataset reflects the actual location and the detailed characteristics of the population distribution in Beijing-Tianjin-Hebei region. Combined with the population density dataset (i.e., WorldPop) and China's kilometer gridded population spatial distribution dataset, the accuracy assessment of PJ2000 is carried out from three aspects: method difference, quantitative error, and regional comparison. The PJ2000 dataset

收稿日期：2019-12-26;**修回日期：**2020-03-17.

基金项目：中国科学院战略性先导科技专项(A类)(XDA19040501);中国工程科技知识中心建设项目(CKCEST-2019-3-6);中国科学院"十三五"信息化专项科学大数据工程项目(XXH13505-07)。 [**Foundation items:** the Strategic Priority Research Program of the Chinese Academy of Sciences, No.XDA19040501; Construction Project of China Knowledge Center for Engineering Sciences and Technology, No.CKCEST-2019-3-6; the Specific Informatization Scientific Research Science Program of the Chinese Academy of Sciences, No.XXH13505-07.]

作者简介：王晓洁(1995—),女,山东烟台人,硕士生,主要从事人口空间化研究。E-mail: wxj@lreis.ac.cn

***通讯作者：**王卷乐(1976—),男,河南洛阳人,博士,研究员,主要从事科学数据共享、地理信息系统与遥感应用研究。E-mail: wangjl@igsnrr.ac.cn

solves the problem of the different distribution of population density over the same land cover type but different towns, and addresses the large difference in the gridded data of population spatialization. The overall accuracy of PJ2000 dataset is 90%, with 87% townships (streets) showing relative error less than 0.5. The correlation coefficient (r) between PJ2000 and the pop2000 township demographic data in the year of 2000 is 0.95. In addition, the population density distribution of this dataset is relatively uniform at the local to large scale. Our results prove that the accuracy of the population density dataset with 1km scale is significantly improved. The population spatialization model is constructed by integrating multi- source data such as township- level demographic data and mobile location data. In the future, it is expected that this method could be applied to obtain the population spatialization distribution for other city agglomerations. Our model could provide high-quality population density dataset for collaborative development of urban agglomeration and risk assessment of natural and man-made disasters in cities, such as earthquake, flood, fire, and public infectious diseases.

Key words: Population density; Spatialization; Demography; Township level; Mobile phone location; Light projection; Exponential smoothing; Beijing, Tianjin and Hebei

***Corresponding author:** WANG Juanle, E-mail: wangjl@igsnrr.ac.cn

摘要:人口在空间上的实际分布是人口地理学研究的基础和热点问题。目前全球不同尺度的人口空间化数据产品因生产方法、数据源等有较大差异,空间化产品的一致性存在较大差异,尤其是共性需求集中的1 km数据产品。本文以京津冀地区为研究区,基于2000年乡镇尺度的人口普查数据和可开放获得的手机定位数据,利用光影投射法计算人口分布权重,结合面积权重法和指数平滑法得到京津冀地区1 km分辨率的人口空间化结果PJ2000。该产品较好地反映了京津冀人口实际分布细节特征。经精度评定,PJ2000人口空间化的总体精度为90%,人口空间化相对误差小于0.5的乡镇(街道)数约占87%,PJ2000与2000年乡镇街道人口统计数据pop2000的相关系数r高达0.95。结果证明,结合乡镇尺度人口统计数据和手机定位数据等多源数据所构建的人口空间化模型,所获1 km分辨率人口密度数据集精度得到显著提高。

关键词:人口密度;空间化;人口学;乡镇尺度;手机定位;光影投射;指数平滑;京津冀

1 引言

人口在空间上的实际分布是人口地理学一个长期和热点研究问题。我国人口普查十年一个周期,由于时间跨度大很难准确反映实际人口流动引致的人口再分布特征;且传统人口普查数据多保存在数值统计报表中,缺乏空间可视化表达。鉴于以上不足,部分学者相继开展全球、国家、州(省)级、县级等尺度的人口数据空间化研究[1-2],根据其发展历程大致分为以下3个阶段:依赖Clark模型[3]、空间自相关分析和地理信息系统分析[4]等区域插值[5]阶段,结合土地利用和夜间灯光等多源遥感数据[6]构建统计回归模型阶段,引入手机定位等社会感知数据构建精细尺度[7-8]人口空间化模型阶段。前2个阶段中形成了UNEP/GRID、GPW及GRUMP、Land-Scan、WorldPop、中国人口空间分布公里网格数据集等具有较大影响力的多尺度人口密度数据集[1-2]。其中1 km尺度的人口数据易于和其他多种类型的地表要素数据融合、叠加和统计分析,应用面最广。例如,田永中等[9]基于2000年《全国分县人口统计资料》结合分县控制、分城乡、分区建模的思路,得到基于土地利用类型的中国1 km尺度的栅格人口模型;卓莉等[10]基于1998年1:400万中国县级人口统计数据得到1 km尺度的中国人口密度图。然而上述数据空间化进展的最小空间粒度多是基于县级尺度人口统计数据进行。自2000年以来,国家统计局开放了乡镇尺度的人口统计数据,为开展乡镇尺度的人口数据空间化模拟提供了条件,但归因于其乡镇界线难于获取、空间化映射困难、行政变迁频繁等问题[6],基于乡镇尺度上的人口空间化模拟鲜有人开展研究。

随着大数据时代的到来,移动电话通讯数据、出租车轨迹数据、社交媒体网络数据等社会感知数据被广泛应用于人口分布模拟[11]。杨皓斐等[12]提出一种利用手机信号数据感知城市人口分布的方法;洪东升等[13]基于定位数据进行人口分布特征的研究,验证了定位数据在地理研究中的巨大潜力。这些位置大数据包含用户大量的实时定位信息,直接反映了人口的实际分布情况,在一定程度上也助力了精细化的人口空间化研究[14]。基于以上两类最

新数据源，如何综合利用乡镇尺度的统计数据和手机定位数据，结合土地利用等传统多源数据，研究精度更高的通用1 km尺度人口空间化数据产品是当前的紧迫需求。本文以人口集聚特征显著的京津冀地区为研究对象，开展基于乡镇尺度人口数据和手机定位数据的人口空间化方法研究。

2　研究区概况与数据源

2.1　研究区概况

京津冀城市群是中国重大城市群之一，实现京津冀地区的协同发展是国家重大发展战略。京津冀地区包括北京、天津、河北三省（直辖市），下辖北京、天津和石家庄、唐山、沧州、保定、秦皇岛、廊坊、承德、张家口、邢台、邯郸等主要城市，总面积约218 000 km^2（图1）。截至2000年底，京津冀辖区范围内共有2323个乡镇街道级行政单元。2000年公布的第五次人口普查结果显示，京津冀地区总人口为9010.2344万人，约占全国总人口的7%。

2.2　数据源及处理

2.2.1　数据源

本文选择的相关数据源包括2000年DMSP/OLS夜间灯光数据[15]，中国土地利用数据[16]，腾讯位置大数据[17]，2000年中国1:25万乡镇界线数据[18]、中国27省乡镇（街道）级人口密度数据集[19]，以及用来进行精度验证的中国人口空间分布公里网格数据集[20]和WorldPop数据集[21]。人口空间化数据源信息见表1。

随着移动通讯设备的发展，腾讯位置服务在微信、QQ、京东和滴滴出行等软件中得到深度应用，本文主要利用腾讯位置大数据来例证手机定位数据在人口空间化方向的应用。基于Python软件，采用网络爬虫的手段获取京津冀地区的定位数据。由于历史同期腾讯定位点数据无法获取，本文选用2019年的定位数据开展实验。腾讯位置数据时间为2019年8月19日至8月25日夜间19:00—22:00，共7 d的数据，该时段人口大多位于常住地，在一定程度上可以避免人口大量流动对腾讯定位数据的影响。由于手机定位数据与人口统计数据的年份相隔较大，为了验证该实验数据能否反应对应的人

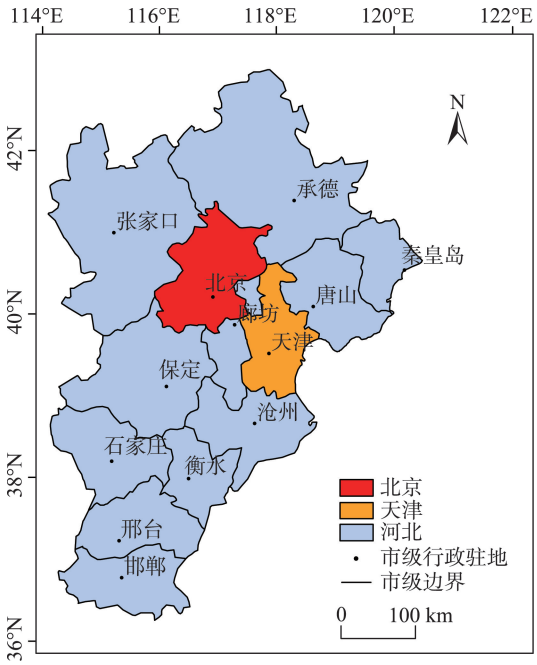

图1　京津冀研究区位置

Fig. 1　Map of study area of Beijing, Tianjin and Hebei

表1　人口空间化数据源

Tab.1 Population spatialization data source

数据名称	年份	精度	来源	作用
DMSP/OLS夜间灯光	2000	—	资源环境数据云平台（http://www.resdc.cn/data.aspx?DATAID=213）	实验数据可用性验证
中国土地利用数据	2000	100 m	资源环境数据云平台（http://www.resdc.cn/data.aspx?DATAID=97）	提供各土地利用类型，以确定不同类型的人口分配权重p
腾讯位置大数据	2019	—	腾讯位置大数据（https://heat.qq.com/index.php）	计算统计人口分配权重p
乡镇界线数据	2000	1:25万	国家科技基础条件平台—国家地球系统科学数据共享平台（http://www.geodata.cn）	提供乡镇街道的边界
27省乡镇（街道）级人口密度数据集	2000	1 km	Science Data Bank（http://www.csdata.org/paperView?id=2）	提供2000年乡镇街道统计人口
中国人口空间分布公里网格数据集	2000	1 km	资源环境数据云平台（http://www.resdc.cn/DOI/DOI.aspx?DOIid=32）	精度对比验证
WorldPop数据集	2000	100 m	WorldPop（https://www.worldpop.org/）	精度对比验证

口分布趋势，选用客观的夜间灯光数据作为参考，按照乡镇界限分别统计每个乡镇街道的腾讯定位次数总值和夜间灯光总值，将2000年乡镇尺度的人口统计数据分别与2019年腾讯定位次数总值以及2000年夜间灯光总值进行线性拟合，计算得到拟合优度R^2分别为0.683、0.462。考虑到夜间灯光分辨率与乡镇边界尺度差异明显、灯光溢出效应的影响和未进行城乡分区建模等相关处理，原始腾讯定位数据显示了其在乡镇尺度上拟合精度较高的优势。因此，从数据可用性角度认为可以采用腾讯定位数据参与人口空间化模拟。

2.2.2 数据处理

由于不同来源的数据格式、范围和投影坐标等不一致，因此需要在进行人口空间化计算之前进行处理。主要包括统一投影坐标系、统一空间分辨率重采样、数据范围裁剪以及相关的数据格式转换等。具体操作如下：

（1）统一投影坐标系：将1:25万乡镇界限数据（京津冀）以及2000年土地利用数据转换至Krasovsky_1940_Albers投影坐标系。

（2）1 km尺度重采样：将2000年土地利用数据重采样至1 km分辨率。

（3）掩模裁剪：根据1:25万乡镇界限数据（京津冀）对土地利用数据（2000年）、中国27省乡镇（街道）级人口密度数据集（2000年）进行掩模裁剪，得到2000年土地利用数据（京津冀）以及2000年乡镇（街道）级人口密度数据集（京津冀）。文件格式为TIFF。

（4）乡镇（街道）单元人口统计数据计算：利用1:25万乡镇界限数据（京津冀）、2000年乡镇（街道）级人口密度数据集（京津冀），基于ArcGis统计乡镇界限范围内的人口密度。根据人口密度公式：人口值等于人口密度乘以街道面积，计算得到乡镇（街道）单元人口统计数据。

（5）腾讯定位数据格式转换：获取的腾讯定位数据包括定位点的经纬度、定位次数、定位时间等属性信息的.txt格式文件。将其导入ArcGis，根据经纬度信息生成786 456个Shapefile点数据，并对点数据进行投影转换至Krasovsky_1940_Albers投影。

3 研究方法

3.1 人口数据空间化方法流程

本文基于乡镇尺度人口统计数据以及腾讯定位数据等多源数据开展京津冀地区1 km尺度的人口空间化。整个方法流程包括对乡镇（街道）人口密度数据集、腾讯位置大数据、土地利用数据等多源数据的预处理操作。结合光线投射法，利用腾讯定位点做射线，判断腾讯定位点数据所在位置，累计得到腾讯定位点在不同乡镇不同地类上的分布比例，即统计人口分配权重p，从而将2000年统计人口按照分配权重p分配到不同地类上。基于ArcGIS生成1 km×1 km尺度的渔网，统计每个格网不同地类的面积，并根据面积权重法（式（4））计算其人口模拟值，进行初步人口空间化。结合（二次）指数平滑法，利用9×9格网尺度窗口对初步空间化结果进行结果修正，窗口的中心栅格值利用平均值代替。最后，结合相对误差和显著相关性对空间化结果进行精度评估。具体方法流程如图2所示。

点包容性是指检查一个点是否在多边形里面，这是一个基本的地理空间操作[22]。通常情况下，判断点是否在简单多边形内部最有效的算法就是“光线投射法”，该算法也被称为“定向射线法”[22]。早在1962年，M. Shimart[23]就发现了上述算法。它的基本原理是：首先，从一个点引出一条射线，该射线与多边形若干条边相交，累计交点个数，若交点为奇数，判断点在多边形内部；若交点为偶数，则判断该点在多边形外部。而判断交点的位置结果分为以下3类情况：① 定位点位于多边形内部（交点个数为奇数）；② 定位点在多边形外部（交点个数为偶数）；③ 定位点位于临界边缘（点在多边形边界线上以及点在顶点处）。

面积权重法在社会经济数据空间化研究中应用最早[24]。这是一种基于变量值保持一致的方法，也是基于多边形对多边形的方法，即多边形叠加分析[25]。它的基本原理是：假设源区A范围内人口为均匀分布，目标区a的人口密度等于源区A的人口密度，按照式（1），计算目标区a的人口p_a。

$$p_a = \Sigma \frac{p_A}{s_A} \times s_a \tag{1}$$

式中：p_a代表目标a的预测人口值；s_a代表目标区a的面积；p_A代表源区A的人口统计值；s_A代表源区A的面积。

指数平滑法是时间序列模型中重要方法之一，分为一次指数平滑、二次指数平滑以及三次指数平滑[26]，本文采用二次指数平滑法，其中t代表距离。它的理论算法如式（2）-（3）。

$$p^{(2)} = a \times p^{(1)} + (1-a) \times p_t^{(2)} \tag{2}$$

$$\sum p \times a = N \tag{3}$$

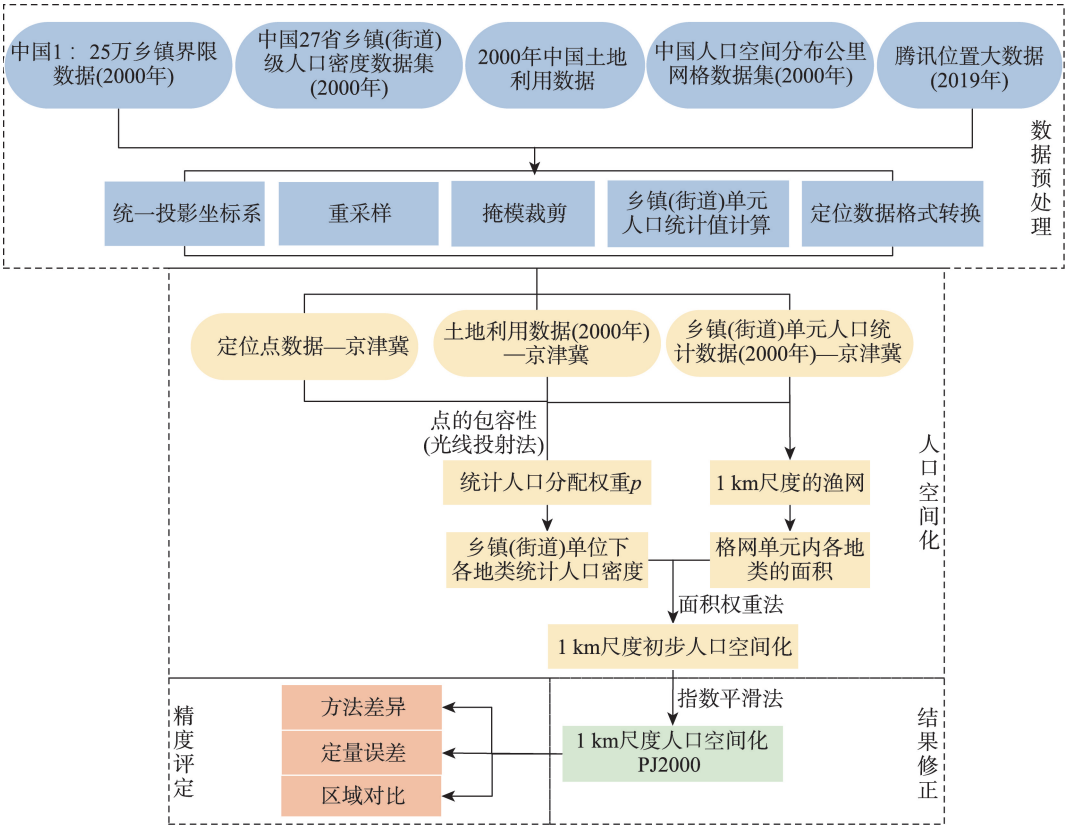

图2　人口数据空间化技术路线

Fig. 2　Technical route of spatial population data

式中：$p^{(1)}$ 代表一次指数平滑后的值；$p^{(2)}$ 代表二次指数平滑之后的值；$p_t^{(2)}$ 代表与初始值相邻 t 个单元的数据平均值；p 代表数据初始值；a 代表平滑系数，值域为(0,1)；N 代表源数据总值。

3.2　人口数据空间化技术实现

3.2.1　计算统计人口分配权重

具体流程(图3)：① 将土地利用类型由TIFF转换格式写入数据表DataFrame，读取腾讯定位数据并进行投影转换。② 输入第一个腾讯定位点，根据其经纬度坐标判断该点位置。③ 判断其是否在行政界线 $poly$ 临界位置(顶点处、边框处)。若在临界位置，则点在行政界限内；若不在，从该腾讯定位点出发引一条射线，判断点与多边形交点个数，交点个数若为奇数，点在行政界线内；若为偶数，点在行政界线外。④ 遍历剩余786 455个腾讯定位点，判断点位于第 i 个乡镇街道单元内并累计定位次数。⑤ 累计定位次数总值写入DataFrame表，得到腾讯定位数据统计表 pos_{ij}(表2)，然后将第 i 个乡镇街道作为单位1，计算第 i 个乡镇街道单元第 j 类土地利用类型的统计人口分配权重 p_{ij}(表3)。

3.2.2　初步人口空间化

根据面积权重法的基本原理(式(1))，假设第 i 个乡镇单元第 j 类土地利用类型上的人口为均匀分布，第 i 个乡镇单元第 j 类土地利用类型的平均人口密度即为第 k 个格网单元内第 j 类土地利用类型的人口密度，根据式(4)计算第 k 个格网单元的初步人口预测值 p_pre_k。

$$p\_pre_k=\sum_{j=1}^{m}\frac{p_{ij}\times pop_i}{s_{ij}}\times s_{kj} \tag{4}$$

式中：p_pre_k 代表第 k 个格网单元的初步人口预测值；p_{ij} 代表第 i 个乡镇街道第 j 类土地利用类型的统计人口分配权重；pop_i 代表第 i 个乡镇街道2000年的统计人口值；s_{ij} 代表第 i 个乡镇街道第 j 类土地利用类型的面积；s_{kj} 代表第 k 个格网单元第 j 类土地利用类型的面积；m 代表土地利用类型的总数。

3.2.3　人口空间化结果修正

由面积权重法得到初步人口空间化结果 p_pre_k，人口分布趋势不明显，分布结果数据边缘跳跃性太大，不符合实际人口的分布情况。本文结合(二次)指数平滑法对初步空间化结果进行修正处理(式(5)—(6))。以单位栅格为中心，设置周围

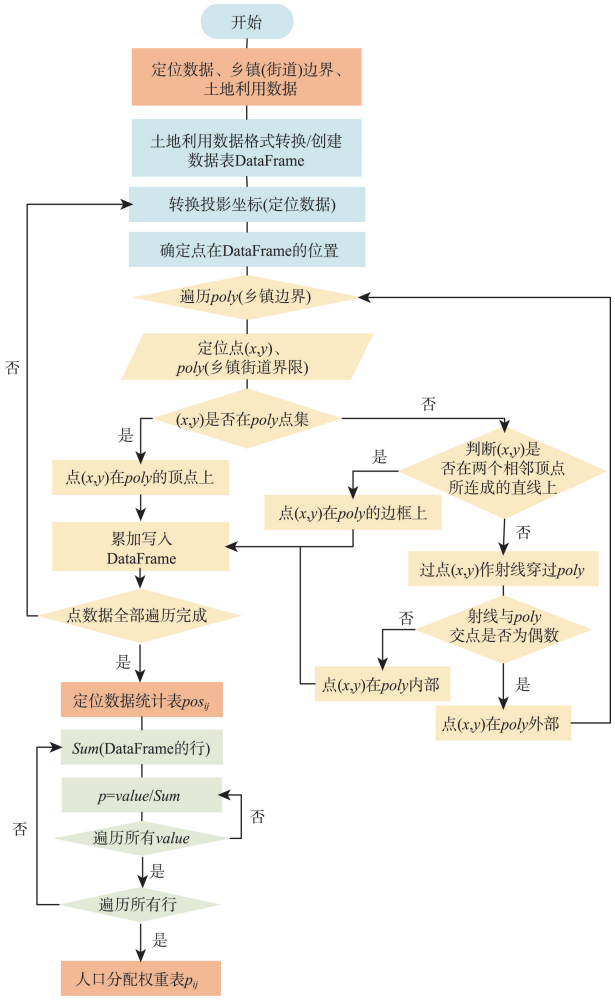

图3 人口分配权重的计算流程

Fig. 3 Calculation process of population distribution weight

9×9格网单元为平滑区域,利用区域平均值代替栅格中心值进行人口修正。使得人口分布趋势符合实际人口分布,满足数据的连续性特点。

$$PJ2000_{st}=a\times p_{st}^{o}+(1-a)\times \overline{p_{st}} \tag{5}$$

式中:$PJ2000_{st}$代表指数平滑之后第s行t列的格网人口值;p_{st}^{o}代表指数平滑之前第s行t列的格网人口值;$\overline{p_{st}}$代表与该格网相邻的9×9范围内修正后的格网人口平均值。

$$\sum p_{st}\times a=N \tag{6}$$

式中:p_{st}代表初步人口空间化的格网单元人口值;a代表平滑系数,值域为(0,1);N代表京津冀地区的统计人口总值。

4 结果与分析

4.1 分配权重p

由3.2.1节得到腾讯定位点在不同乡镇(街道)不同地类上的数据统计表(表2)以及统计人口分配权重p(表3),可以看出人口主要分布在旱地、有林地、灌木林、疏林地以及其他林地、建筑用地等区域内,其他地类上人口分布相对较少。

4.2 人口空间化最终结果

京津冀地区人口空间化结果数据集PJ2000(图4),数据格式是TIFF格式,空间分辨率为1 km×1 km。从PJ2000空间分布看出,人口密度大于等于3000人/km^{2}的人口稠密地区主要集中分布在北京市、唐山市、石家庄市、天津市、邢台市以及邯郸市等京津冀的东南部区域,京津冀西北部地区人口则相对稀疏,人口密度多集中在500人/km^2

表2 腾讯定位数据统计表pos_{ij}

Tab. 2 Tencent positioning data statistics table pos_{ij}

乡镇街道(序号)/地类	11	12	21	22	23	24	31	32	33	41	42	…
0	0	2063	403	0	0	0	0	0	0	0	0	…
1	0	14	1002	0	0	55	0	0	0	0	0	…
2	0	588	49	0	0	0	0	0	0	0	0	…
3	0	227	66	0	737	0	0	0	0	0	0	…
4	0	13 063	0	0	0	0	0	0	0	0	0	…
5	0	16 711	0	0	0	0	0	0	0	0	0	…
6	0	0	30	80	34	49	0	0	0	0	0	…
7	0	2320	0	0	0	0	0	0	0	0	0	…
8	0	0	91	234	0	0	0	0	0	0	0	…
9	0	0	0	0	0	0	0	0	0	0	0	…
10	0	0	0	0	1197	0	0	0	0	0	0	…
…	…	…	…	…	…	…	…	…	…	…	…	…
2322	0	1003	0	0	0	0	0	0	0	0	0	…

表3　统计人口分配权重p_{ij}

Tab. 3　Statistical population distribution weight p_{ij}

乡镇街道（序号）/地类	11	12	21	22	23	24	31	32	33	41	42	…
0	0.00	0.84	0.16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	…
1	0.00	0.01	0.94	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.00	…
2	0.00	0.92	0.08	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	…
3	0.00	0.22	0.06	0.00	0.72	0.00	0.00	0.00	0.00	0.00	0.00	…
4	0.00	0.89	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	…
5	0.00	0.57	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	…
6	0.00	0.00	0.16	0.41	0.18	0.25	0.00	0.00	0.00	0.00	0.00	…
7	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	…
8	0.00	0.00	0.28	0.72	0.00	0.00	0.00	0.00	0.00	0.00	0.00	…
9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	…
10	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	…
…	…	…	…	…	…	…	…	…	…	…	…	…
2322	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	…

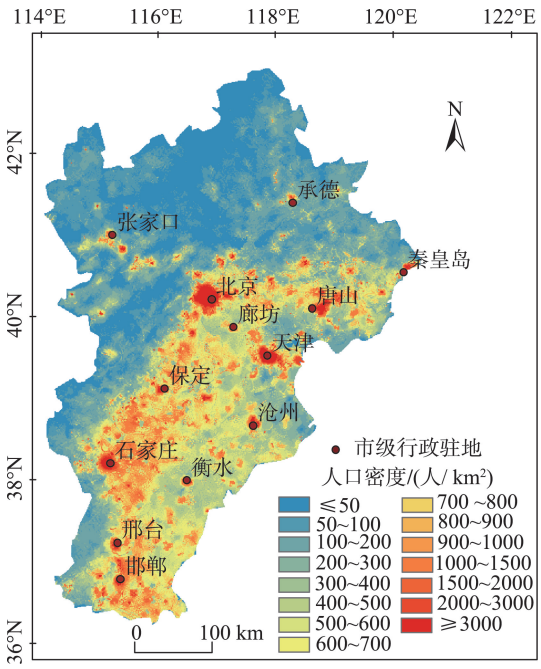

图4　京津冀2000年人口密度数据集PJ2000

Fig. 4　Population density dataset PJ2000 of Beijing, Tianjin and Hebei in 2000

以内。人口密度大多呈现从城市中心向外辐射递减的趋势。北京市、天津市、张家口市、承德市以及秦皇岛市这几个地区的人口中心不位于城市几何中心，尤其是承德市和秦皇岛市，人口中心主要集中在偏东南部地区，这与其地形特点有关。承德市人口主要集中分布在滦河周边，秦皇岛北依燕山山脉，南邻渤海，地势呈现北高南低的趋势，东部靠近渤海港口，所以人口相对集中东部地势平缓地区以及靠近港湾地区。

4.3　精度验证

结合中国公里格网人口密度数据集、WorldPop2000年人口密度数据集以及人口空间化结果PJ20003种数据集进行精度验证，包括方法差异、定量误差以及区域对比这3方面的验证。（注：下文称本文的空间化结果为PJ2000；称中国公里格网人口密度数据集为tpop2000；称乡镇街道人口统计值为pop2000；称WorldPop2000年人口密度数据集为WorldPop。）

（1）方法差异。传统面积权重法的目标区为整个地类，实际上不同街道内部不同地类的人口分布具有差异。本文引入了手机定位数据，利用该数据表征人口实时分布等优势，结合普通面积权重法的集本原理，将目标区缩小到乡镇街道内部不同土地利用类型上，在一定程度上克服了传统方法目标区范围较大造成乡镇街道人口空间化结果不准确的缺陷。对比结果如图5所示。

（2）定量误差。结合相对误差（式（7））以及显著相关性检验（表4）对空间化结果*PJ*2000进行定量精度验证。

$$Error=\frac{PJ2000-pop2000}{pop2000} \tag{7}$$

式中：*Error*代表统计误差；*PJ*2000代表空间化模拟人口根据乡镇（街道）行政区的统计值；*pop*2000代表2000年的人口统计值。

计算得到*PJ*2000与*pop*2000的相对误差值（图6）

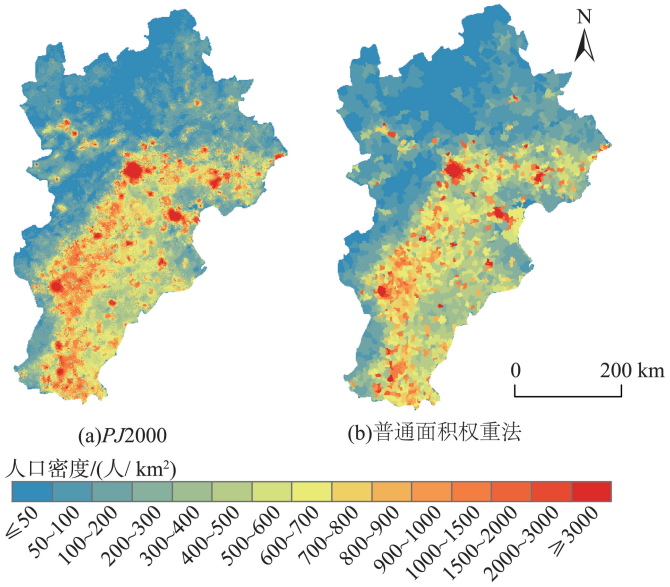

图5 *PJ*2000与普通面积权重法结果对比

Fig. 5 Comparison between PJ2000 and common area weight method

小于等于0.5的个数为2037个，正确率约占乡镇（街道）总数的87%。*worldpop*与*pop*2000之间相对误差值（图7）小于等于0.5的个数有1920个，正确率约占83%。*tpop*2000与*pop*2000之间相对误差值（图8）

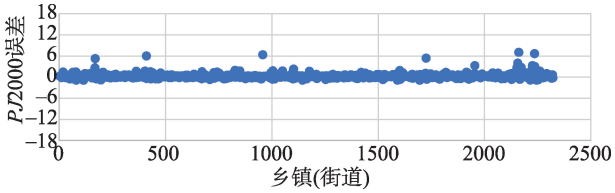

图6 *PJ*2000误差

Fig. 6 *PJ*2000 error

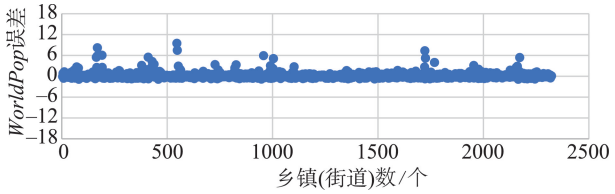

图7 *WorldPop*误差

Fig. 7 *WorldPop* error

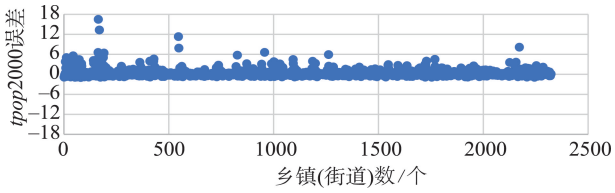

图8 *tpop*2000误差

Fig. 8 *tpop*2000 error

小于等于0.5的个数有1721个，正确率仅有74%。

按照乡镇边界分别统计*PJ*2000、*tpop*2000以及*WorldPop*，分别与2000年乡镇人口普查数据*pop*2000进行显著相关性检验，结果如表4所示。

表4 显著相关性检验

Tab. 4 Significant correlation test

	*PJ*2000	*WorldPop*	*tpop*2000
*pop*2000	0.950**	0.935**	0.905**

注：**表示在1%水平（双侧）上显著相关。

（3）区域对比。结合*tpop*2000数据集以及*PJ*2000进行区域对比，结果证明*PJ*2000在张家口市、承德市、唐山等局部地区的人口空间分布在细节表达上更准确，整体的人口空间化分布边缘跳跃性较小。结果如图9—图10所示。

5 讨论

在方法方面，本研究利用光影投射法准确判断腾讯定位点的位置，累计腾讯定位次数得到统计人口在乡镇街道内部不同土地利用类型上的实际分配权重，按照该权重来分配乡镇统计人口，有效解决了乡镇街道模拟人口和统计人口差异较大的问题。不同乡镇街道同地类人口分布具有差异性，基于传统面积权重法的基本原理，将目标区从整个土地利用类型缩小到乡镇街道内部不同土地利用类型上，结合二次指数平滑法构建京津冀2000年人口空间化模型，得到人口空间化初步模拟数据更加符合人口的实际分布。未来可以利用主成分分析[27]计算统计人口分配权重，结合K-means聚类以及距离衰减[28]等方法构建更高精度的人口空间化模型。随着手机定位数据等大数据的开放性和可获取性越来越强，需要结合相同时间尺度的乡镇街道人口统计数据、POI数据[29]以及其他类型的手机信号数据，探究多源数据与人口分布的关系，构建完整的人口空间化基础数据体系。

在结果方面，本文得到京津冀地区1 km尺度人口空间化产品*PJ*2000，可以更好地刻画京津冀地区的人口空间分布格局。结合国际以及国内认可度较高的WorldPop2000年人口密度数据集、中国人口空间分布公里网格数据集*tpop*2000以及2000年乡镇街道人口普查数据*pop*2000进行精度验证，具体包括方法差异、定量误差以及区域对比3方面。通过对比得到*PJ*2000的人口分布的细节更加显著，证明

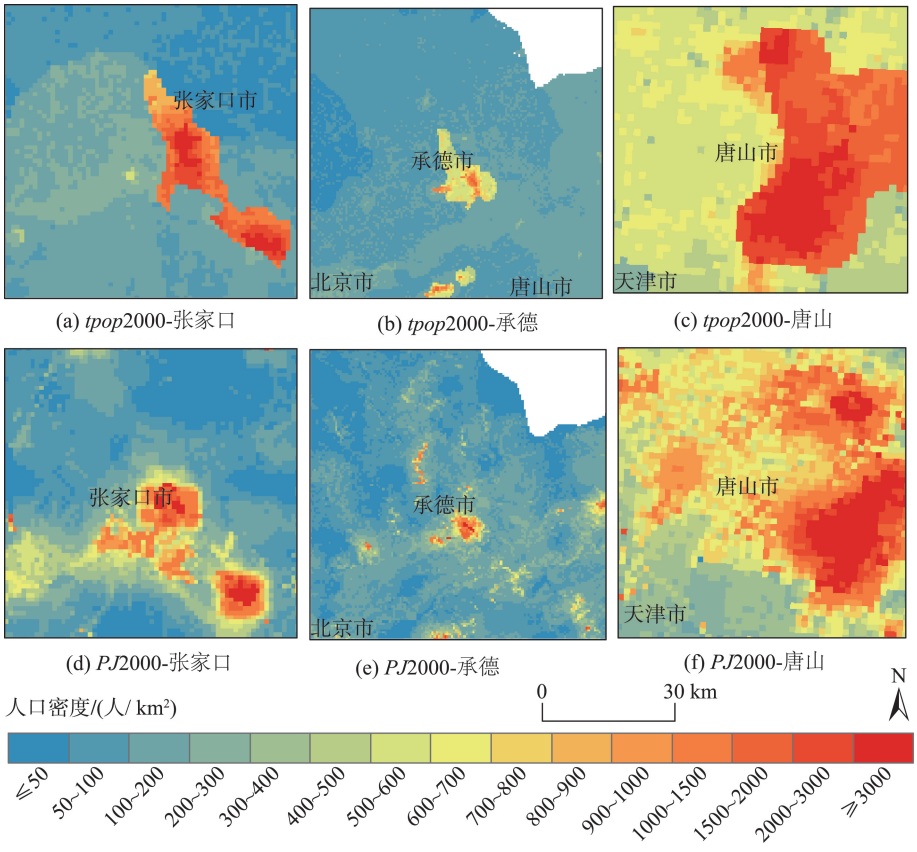

图9　*PJ*2000与*tpop*2000局部对比

Fig. 9　Local Contrast Graph of *PJ*2000 and *tpop*2000

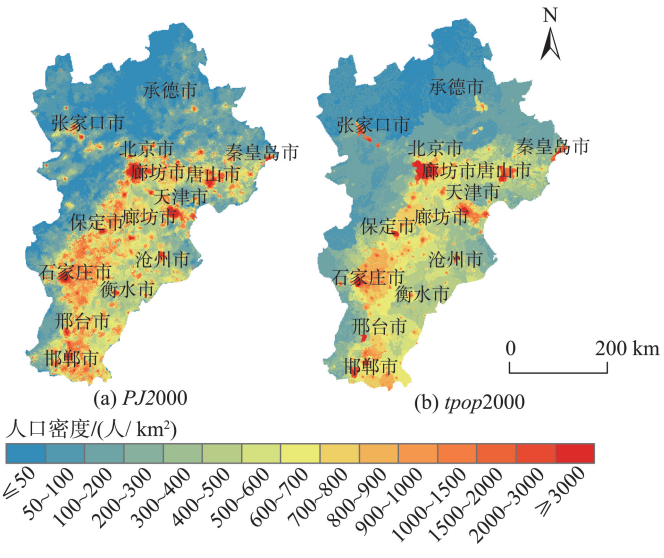

图10　*tpop*2000与*PJ*2000全局对比

Fig. 10　Global comparison between *tpop*2000 and *PJ*2000

本文构建的空间化方法在一定程度上提高了1 km尺度的人口空间化模拟精度。从*PJ*2000结果得到，人口密度≥3000人/km²集中分布在京津冀的东南部地区，西北地区人口密度主要在≤50人/km²范围内，人口分布总体格局呈现由东南地区向西北地区递减的趋势。另外，受地形、水资源、政治以及经济等因素的影响，人口分布中心往往偏离城市的几何中心，这在1 km尺度的人口空间化结果*PJ*2000中均得到体现。

在数据产品应用方面，结合乡镇尺度人口普查统计数据和手机定位数据等多源数据产出的高精度城市群地区人口空间化数据产品具有巨大的应用潜力。这不仅能够为我国京津冀等城市化地区的创新协同发展提供高精度基础人口数据支持，而且预期将其方法可应用于长江三角洲、珠江三角洲、成渝等更多城市群提供人口空间化产品，服务于国土空间规划以及人口资源环境配置等协同发展的辅助决策支持，为地震、洪水、火灾、传染病等城市区域的自然和人为灾害[30]的风险评估提供数据支持。

本研究在综合利用多源数据的同时也存在局限性。在利用腾讯位置大数据优势的同时，不可忽略腾讯位置大数据的有偏性。具体表现在3个方面：① 几十万的腾讯位置大数据量仍不能完全表征

该地区准确的实际人口情况；② 微信、QQ、滴滴打车等应用腾讯位置的软件得经过位置授权才可以采集到实时的定位点数据；③ 不同年龄段的用户对于手机腾讯定位使用普及率不是100%。目前也有很多学者针对社交媒体大数据的有偏性问题进行研究，本文通过拟合优度证明腾讯定位数据与乡镇（街道）统计人口数据具有较强相关性，显示出社交媒体大数据的全覆盖优势，但是没有对腾讯数据的有偏性进行定量深入探讨，这一部分工作在后续研究中会继续深入。

6 结论

针对人口数据空间化产品对技术方法的需求，本文基于2000年首次公布的乡镇（街道）级人口普查数据和越来越开放且可能获取的手机定位数据等多源数据进行人口空间化方法研究，获得以下结论。

（1）融合了乡镇街道人口统计数据尺度较小的特点和手机定位数据反映人口实际位置等优势，构建京津冀地区1 km尺度人口空间化模型。该模型能够充分利用多源数据优势，更客观地反映该地区人口真实的空间分布，为传统1 km尺度人口空间化产品反演提供了大数据的解决方法。

（2）在面积权重法的基础上结合光影投射法和二次指数平滑法进行统计人口空间化及其结果修正。避免了不同乡镇同地类之间人口分布的差异对空间化结果精度的影响，显著减少了空间化初步模拟人口分布跳跃性过大的问题，并刻画出区域内部人口的精细分布特征。

（3）结合精度较高的中国公里格网人口密度数据集和WorldPop数据集，从方法差异、定量误差以及区域对比三方面对本研究的人口空间化产品进行精度评估，其总体精度为90%，与2000年乡镇人口普查数据相关系数r高达95%，乡镇尺度统计人口的相对误差小于50%的个数约占乡镇总个数的87%，结果证明人口空间化数据产品的精度得到了显著提高。

参考文献(References)：

[1] 杨晓荣,陈楠.基于多源数据的福建省人口数据空间化研究[J].贵州大学学报(自然科学版),2019,36(2):79-84,95. [Yang X R, Chen N. Spatialization of population data for Fujian province based on multi-source data[J]. Journal of Guizhou University (Natural Science Edition), 2019,36(2):79-84,95.]

[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 population data[J]. Progress in Geography, 2013,32(11):1692-1702.]

[3] 陈彦光.城市人口空间分布函数的理论基础与修正形式——利用最大熵方法推导关于城市人口密度衰减的Clark模型[J].华中师范大学学报(自然科学版),2000,34(4):489-492. [Chen Y G. Derivation and generalization of Clark's model on urban population density using entropy--maximising methods and fractal ideas[J]. Journal of Central Normal University(Natural Science), 2000,34(4):489-492.]

[4] Zhang J, Zhu Y. The population spatial distribution model based on the spatial statistics in Shandong province, China[J]. International Conference on Geoinformatics, 2011:1-4. DOI:10.1109/GeoInformatics.2011.5981117。

[5] 邓顺强.基于随机森林算法和多源数据的人口空间分布模型研究[D].上海:华东师范大学,2018. [Deng S Q. Multi-source data based spatial model of population distribution using random forests[D]. Shanghai: East China Normal University, 2018.]

[6] 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-level population based on nighttime light and land use data in Shandong province[J]. Journal of Geo-Information Science, 2019,21(5):699-709.]

[7] 吴中元,许捍卫,胡钟敏.基于腾讯位置大数据的精细尺度人口空间化——以南京市江宁区秣陵街道为例[J].地理与地理信息科学,2019,35(6):61-65. [Wu Z Y, Xv H W, Hu Z M. Fine-scale population spatialization based on tencent location big data: A case study of moling subdistrict, Jiangning district, Nanjing[J]. Geography and Geo-Information Science, 2019,35(6):61-65.]

[8] 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 population in the pearl river delta in 30 m grids using random forest model[J]. Progress in Geography, 2017,36(10):1304-1312.]

[9] 田永中,陈述彭,岳天祥,等.基于土地利用的中国人口密度模拟[J].地理学报,2004,59(2):283-292. [Tian Y Z, Chen S P, Yue T X, et al. Simulation of Chinese population density based on land use[J]. Acta Geographica Sinica, 2004,59(2):283-292.]

[10] 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. [Zhuo L, Chen J, Shi P J, et al. Modeling population density of China in 1998 based on DMSP/OLS nighttime light image [J]. Acta Geographica Sinica, 2005,60(2):266-276.]

[11] 黄益修.基于夜间灯光遥感影像和社会感知数据的人口空间化研究[D].上海:华东师范大学,2016. [Huang Y X. Spatialization of population using nighttime light remote sensing images and social sensing data[D]. Shanghai: East China Normal University, 2016.]

[12] 杨皓斐,曹仲,李付琛.基于手机大数据的动态人口感知[J].计算机系统应用,2018,27(5):73-79. [Yang H F, Cao Z, Li F C. Dynamic population perception based on mobile phone big data[J]. Computer Systems & Applications, 2018,27(5):73-79.]

[13] 洪东升.基于定位数据的人口分布特征研究[D].北京:中国地质大学(北京),2015. [Hong D S. Research on characteristics of population distribution based on positioning data[D].Beijing: China University of Geosciences (Beijing), 2015.]

[14] 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. [Xiao D S, Yang S. A review of population spatial distribution based on nighttime light data[J]. Remote Sensing for Land & Resources, 2019,31(3):10-19.]

[15] 资源环境数据云平台.DMSP/OLS夜间灯光数据(2000)[DB/OL].http://www.resdc.cn/data.aspx?DATAID=213. [Resource and Environment Data Cloud Platform. DMSP/OLS night light data(2000)[DB/OL]. http://www.resdc.cn/data.aspx?DATAID=213.]

[16] 资源环境数据云平台.中国土地利用现状遥感监测数据(2000)[DB/OL]. http://www.resdc.cn/data.aspx?DATAID=97. [Resource and Environment Data Cloud Platform. Remote sensing monitoring data of land use in China(2000)[DB/OL]. http://www.resdc.cn/data.aspx?DATAID=97.]

[17] 腾讯位置大数据.腾讯定位开放平台定位次数[DB/OL]. https://heat.qq.com/index.php.2019. [Tencent's Location Big Data. Positioning of tencent positioning open platform[DB/OL]. https://heat.qq.com/index.php. 2019.]

[18] 国家科技基础条件平台—国家地球系统科学数据共享服务平台.中国1:25万乡镇界线数据(2000)[DB/OL]. http://www.geodata.cn. 2017. [National Earth System Science Data Sharing Infrastructure, National Science & Technology Infrastructure of China.China's 1:25 million township boundary data (2000) [DB/OL]. http://www.geodata.cn. 2017.]

[19] 柏中强,王卷乐.中国27省乡镇(街道)级人口密度数据集[J/OL].中国科学数据,2016,1(1). http://www.csdata.org/paperView?id=2. DOI: 10.11922/csdata.170.2015.0002. [Bai Z Q, Wang J L. A dataset of population density at township level for 27 provinces of China[J/OL]. China Science Data, 2016,1(1).http://www.csdata.org/paperView?id=2. DOI: 10.11922/csdata.170.2015.0002.]

[20] 徐新良.中国人口空间分布公里网格数据集[DB/OL].资源环境数据云平台,http://www.resdc.cn/DOI/DOI.aspx?DOIid=32, 2017, DOI:10.12078/2017121101. [Xv X L. China's population spatial distribution kilometer grid dataset[DB/OL]. Resource and Environment Data Cloud Platform, http://www.resdc.cn/DOI/DOI.aspx?DOIid=32, 2017, DOI:10.12078/2017121101.]

[21] WorldPop. China population 2000[DB/OL]. https://www.worldpop.org/geodata/summary?id=1552, 2018.

[22] Lawhead, J. Python地理空间分析指南(邓世超译)[M].北京:人民邮电出版社,2017. [Lawhead, J. Learning geospatial analysis with python(Deng S C, trans)[M]. Beijing: Posts &Telecom Press, 2017.]

[23] Shimrat, M. Algorithm112: Position of point relative to polygon[J]. Communications of the Association for Computing Machinery, 1962,5(8):434.

[24] 吴吉东,王旭,王菜林,等.社会经济数据空间化现状与 发展趋势[J].地球信息科学学报,2018,20(9):1252-1262. [Wu J D, Wang X, Wang C L, et al. The status and development trend of disaggregation of socio-economic Data[J]. Journal of Geo-information Science, 2018,20(9):1252-1262.]

[25] 潘志强,刘高焕.面插值的研究进展[J].地理科学进展,2002,21(2):146-152. [Pan Z Q, Liu G H. The research progress of areal interpolation[J]. Progress in Geography, 2002,21(2):146-152.]

[26] 张秋悦.重庆市GDP值的趋势预测分析——基于指数平滑法[J].价值工程,2019,38(27):187-189. [Zhang Q Y. Trend prediction and analysis of GDP value in Chongqing: Based on exponential smoothing methord[J]. Value Engineering, 2019,38(27):187-189.]

[27] 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/OL].热带地理,2019(12):1-16. [Zhao X, Song Y Q, Liu Y L, et al. Population spatialization based on satellite remote sensing and POI data: Regarding Guangzhou as an example[J/OL]. Tropical Geography, 2019(12):1-16.]

[28] 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. [Chun J, Zhang X C, Huang J F, et al. A gridding method of redistributing population based on POIs[J]. Geography and Geo-Information Science, 2018,34(4):83-89,124.]

[29] 杨旭.人口统计数据空间化不同方案及其误差评价[D].开封:河南大学,2015. [Yang X. The difference schemes and error evaluation on spatialization of statistical population data[D]. Kaifeng: Henan University, 2015.]

[30] 吴安坤,田鹏举,黄天福,等.基于人口/GDP数据空间化的雷电灾害风险评价[J].气象科技,2018,46(5):1026-1031. [Wu A K, Tian P J, Huang T F, et al. Risk assessment of lightning disasters based on spatial population/GDP data [J]. Meteorological Science and Technology, 2018,46(5):1026-1031.]