

山东省街道/乡镇级格网化人口空间分布数据集 (2000、2010 年)

ISSN 2096-2223

CN 11-6035/N

王明明^{1,2}, 王卷乐^{2,3*}



文献 DOI:

10.11922/csdata.2019.0005.zh

数据 DOI:

10.11922/sciencedb.736

文献分类: 地球科学

收稿日期: 2019-02-28

开放同评: 2019-03-11

录用日期: 2019-05-06

发表日期: 2019-09-20

1. 山东理工大学建筑工程学院, 山东淄博 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中心, 南京 210023

摘要: 格网化人口空间分布数据是更精细尺度的人口空间分布数据, 对于支持地理、资源、环境、生态、灾害, 以及可持续发展等方面的研究具有重要意义。本文以山东省街道/乡镇级人口统计数据为控制单元, 在城乡二级分区的基础上, 将基于增强型植被指数修正的夜间灯光数据和土地利用数据相结合的空间化方法, 产出了 SDpop2000 和 SDpop2010 两期山东省 100 m 格网人口空间分布数据集。模型精度验证结果显示 2000 年与 2010 年分别有 78.0%和 78.4%的行政单元的相对误差绝对值小于 20%。空间化结果与精度较高的全球 WorldPop 人口数据产品对比分析, 结果显示: 2000 年的相关系数和相对均方根误差分别为 0.932 和 38.42%, 2010 年的相关系数和相对均方根误差分别为 0.948 和 40.46%。

关键词: 人口; 空间化; 夜间灯光数据; 土地利用数据; 街道/乡镇级; 山东省

数据库(集)基本信息简介

数据库(集)名称	山东省街道/乡镇级格网化人口空间分布数据集(2000、2010 年)
数据作者	王明明、王卷乐
数据通信作者	王卷乐 (wangjl@igsrr.ac.cn)
数据时间范围	2000、2010年
地理区域	中国山东省 (34°22'54"-38°24'0.6"N, 114°47'30"-122°42'18"E)
空间分辨率	100 m
数据量	36.07 MB
数据格式	*.tif, *.jpg
数据服务系统网址	http://www.sciencedb.cn/dataSet/handle/736
基金项目	中国科学院战略性先导科技专项(A类)资助(XDA19040501); 中国科学院“十三五”信息化专项科学大数据工程项目(XXH13505-07)。

* 论文通信作者

王卷乐: wangjl@igsrr.ac.cn

数据库（集）组成	本数据集由2部分组成：（1）“Dataset Result.Zip”包含2000年和2010年山东省格网人口空间分布数据。数据命名格式为SDpopYYYY.tif，YYYY代表年份。数据量为13.42 MB。（2）“Dataset Map.Zip”包含2000年和2010年山东省格网人口空间分布图。数据命名格式为SDpopYYYY.jpg。数据量为22.65 MB。
----------	--

引言

街道/乡镇级人口统计数据是我国目前公开的最精细的统计数据，街道/乡镇级人口空间分布数据成为地学研究的重要需求之一，但单纯地将统计数据以行政区划为单元分布到空间，往往存在行政单元尺度不够精细^[1]、行政区划系统存在变动^[2]、行政边界造成地学研究中的“可变元问题”^[3]等不足。格网化人口空间分布数据是传统人口统计数据的拓展和补充，其实现了行政单元尺度人口的重新展布，有效克服了人口统计数据存在的不足，对地理、资源、环境、生态、灾害以及可持续发展等跨学科研究具有重要意义。

目前涉及中国区域的格网化人口数据已有中国人口空间分布公里网格数据集（Chpop）、全球栅格人口数据集（GPW）、全球农村城镇制图计划数据集（GRUMP）和世界人口计划数据集（Worldpop）等，这些数据集中除了 GPW 外，其他均是以县级人口统计数据为基础，借助人口统计数据与土地利用、夜间灯光和土地覆被等数据的关系建立空间分布模型。尽管 GPW 使用了 2010 年的街道/乡镇级人口统计数据，但其街道/乡镇级行政边界数据不是实际边界数据，而是采用特殊方法的模拟数据^[4]。可见，真实的街道/乡镇级人口格网化数据产品目前是极为贫乏的。究其原因在于三点，一是街道/乡镇级行政区划矢量数据获取困难，且存在大量的行政区划和地名变更问题，难以直接利用。二是，街道/乡镇级人口统计数据的空间关联需要进行大量的涉及乡镇（街道）地名和行政区划的空间数据匹配处理工作，非专业的团队难以完成。三是，空间化的方法单一，缺少多源数据和中国人口地理学特点的综合运用，例如中国人口分布的城乡二级分布特征。以上这些原因使得当前人口数据空间化研究多是基于县级人口统计数据^[5-8]，即使有个别街道/乡镇级的研究^[9-10]，也只停留在小范围内。针对这一需求，本文完成了 2000 年和 2010 年山东省街道/乡镇级人口匹配处理，并利用土地利用数据和夜间灯光数据相结合的空间化方法，实现了两期人口数据的空间化工作。其中土地利用数据和街道/乡镇级行政单元相结合构建城乡二级分区方法，增强型植被指数（EVI）对 DMSP/OLS 夜间灯光数据进行修正，保证了空间化模型的精度。

1 数据采集和处理方法

1.1 数据来源

生产本数据集所采用的数据包括 DMSP/OLS 夜间非辐射定标的平均稳态数据、EVI、土地利用数据、世界人口计划数据集 WorldPop、街道/乡镇级人口统计数据和行政边界数据。数据的相关参数和来源如表 1 中所示。

表 1 数据源详细信息

序号	名称	分辨率	时间（年份）	来源	网址
1	DMSP/OLS 夜间非辐射定标的平均稳态数据	30"	2000/2010	美国国家环境数据中心	https://www.ncei.noaa.gov/
2	EVI	250 m	2000/2010	美国 NASA 一级大气存档和分配系统数据中心	https://ladsweb.modaps.eosdis.nasa.gov/
3	土地利用数据	100 m	2000/2010	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/
4	WorldPop	3"	2000/2010	WorldPop project 官方网站	https://www.worldpop.org/
5	街道/乡镇级行政边界数据	1:25 万	2000	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/
6	街道/乡镇级人口统计数据		2000/2010	中国知网	http://www.cnki.net/

1.2 数据处理

（1）街道/乡镇级人口空间数据

街道/乡镇级人口空间数据制作需要将街道/乡镇级人口统计数据与街道/乡镇级行政区划矢量数据的属性进行关联，实现街道/乡镇级人口统计值在空间上的分布。本文采用决策树的思想实现两者的属性关联，过程如图 1 所示。

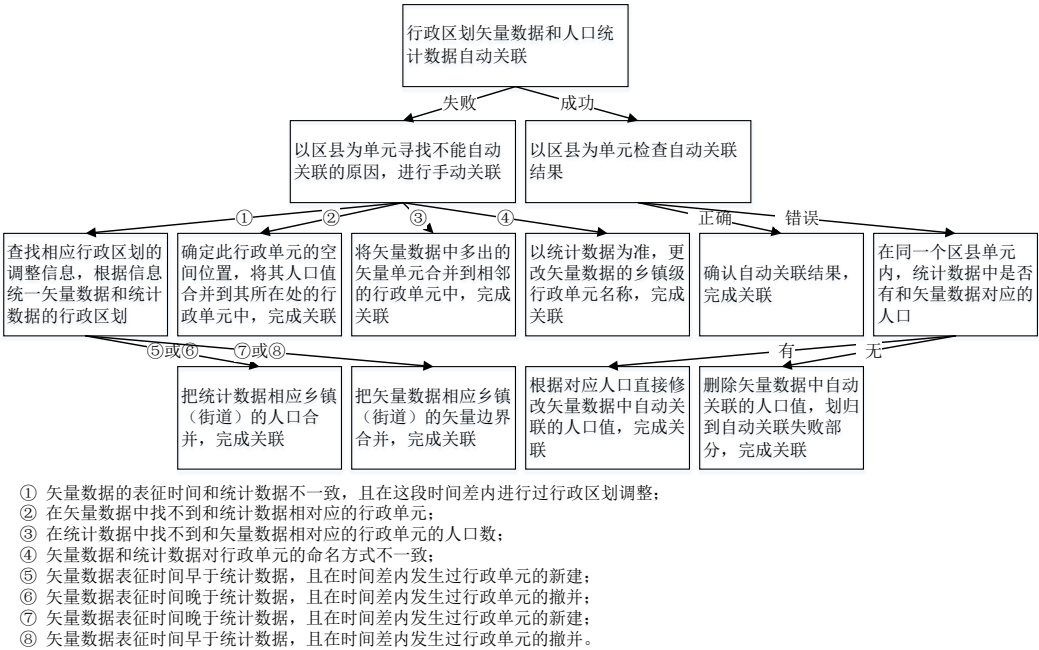


图 1 街道/乡镇级人口空间数据制作流程

（2）夜间灯光数据

DMSP/OLS 夜间灯光数据由于 OLS 传感器本身的缺陷存在饱和溢出现象，会影响到城市中心人口数据空间化的精度。有研究表明可以通过 EVI 对 DMSP/OLS 夜间灯光数据进行修正，生成增强型植被指数调整的夜间灯光指数（EANTLI），修正公式如式（1）^[11]。

$$EANTLI = \frac{1 + (NTL_n - EVI_a)}{1 - (NTL_n - EVI_a)} \times NTL \quad (1)$$

式中， NTL_n 为归一化后的 DMSP/OLS 夜间灯光指数， EVI_a 为处理后的 EVI 数据， NTL 为原始 DMSP/OLS 夜间灯光指数。

（3）城乡二级分区

建模之前将样本进行二级分区处理，技术流程如图 2。一级分区：利用土地利用数据将所有行政单元分为城镇地区（ A_1 ）、农村地区和其他地区。其中城镇地区中的居民地类型只包含城镇用地，农村地区的居民地类型只包含农村居民点，其他地区则包含这两种居民地类型。二级分区：以 1:4:1 的比例将农村地区分为低人口密度区（ A_2 ）、中人口密度区（ A_3 ）和高人口密度区（ A_4 ）；根据模型检验结果将其余地区分为 A_5 （相对误差绝对值 < 20%）、 A_6 （相对误差 < -20%）和 A_7 （相对误差 > 20%）3 个分区。

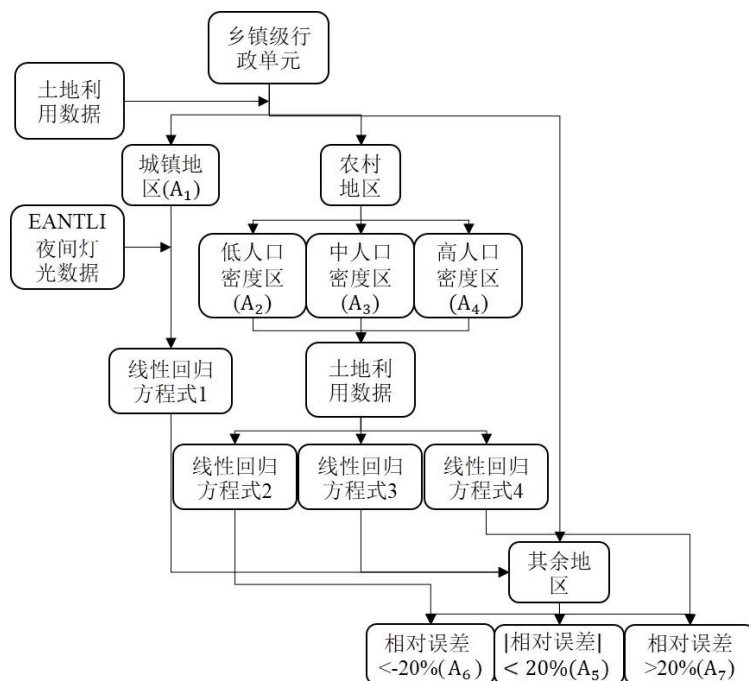


图 2 城乡二级分区和建模技术流程图

（4）空间化模型

分别利用线性回归模型和多层感知器（MLP）人工神经网络模型建立人口数据空间化模型。 A_1 、 A_2 、 A_3 和 A_4 分区的建模结果依次对应图 2 中的方程式 1—4。

①线性回归模型。 A_1 分区的人口数据空间化模型如式（2）所示， A_2 、 A_3 和 A_4 分区的人口数据空间化模型如式（3）所示。

$$pop = a \times EANTLI + b \quad (2)$$

$$pop_i = a_i \times area_i + b_i \quad (3)$$

式中, pop 和 pop_i 为街道/乡镇级行政单元内的人口值; $EANTLI$ 为行政单元中城镇用地范围内的 $EANTLI$ 灯光强度值; a 和 a_i 为回归系数; b 和 b_i 为常数项; $area_i$ 为行政单元中农村居民点的面积; i 表示第 i 个分区, 取值分别为 2、3、4。

②MLP 人工神经网络模型。在 A_1 、 A_2 、 A_3 和 A_4 分区, 将土地利用数据和 $EANTLI$ 夜间灯光数据作为输入层, 街道/乡镇级人口统计数据作为输出层, 包含一个隐藏层, 模型如式 (4) 和式 (5) 所示。

$$f(x) = G(b_2 + W_2(S(b_1 + W_1x))) \quad (4)$$

$$S(a) = 1 / (1 + e^{-a}) \quad (5)$$

式中, G 为隐藏层到输出层的逻辑回归函数, S 为输入层到隐藏层的函数, W_1 和 W_2 分别为输入层到隐藏层和隐藏层到输出层的权重, b_1 和 b_2 分别为输入层到隐藏层和隐藏层到输出层的偏置, x 和 a 为变量。

(5) 格网单元人口模拟

选取模型精度最优的空间化模型模拟格网单元的人口值。经后文模型精度验证可知, 线性回归模型的精度优于 MLP 人工神经网络模型的精度, 因此本文利用线性回归模型的建模结果模拟格网单元的人口值。其中 A_1 、 A_2 、 A_3 和 A_4 分区依次利用方程式 1–4 模拟, A_5 分区利用方程式 1 和方程式 3 模拟, A_6 分区利用方程式 1 和方程式 2 模拟, A_7 分区利用方程式 1 和方程式 4 模拟。

(6) 格网单元人口修正

由于模型公式存在一定的误差, 行政单元内的人口模拟值和人口统计值不一致。建立人口模拟修正公式, 见式 (6), 对格网内的人口模拟值进行调整。

$$p_{ij}^{mo} = \frac{p_{ij}^0}{pop_i^0} \times pop_i^{sta} \quad (6)$$

式中, p_{ij}^{mo} 表示第 i 个行政单元内第 j 个格网单元人口模拟的修正值; p_{ij}^0 表示第 i 个行政单元内第 j 个格网单元初始人口模拟值; pop_i^0 表示第 i 个行政单元的初始人口模拟值; pop_i^{sta} 表示第 i 个行政单元的人口统计值。

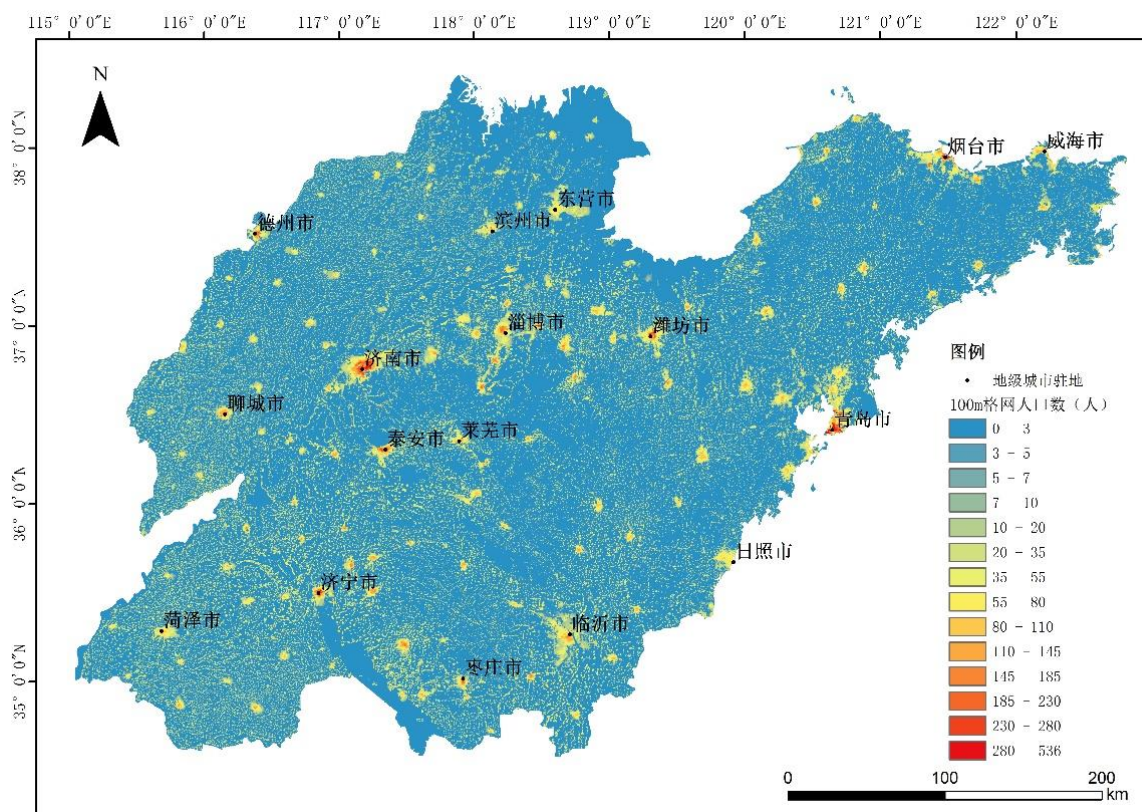
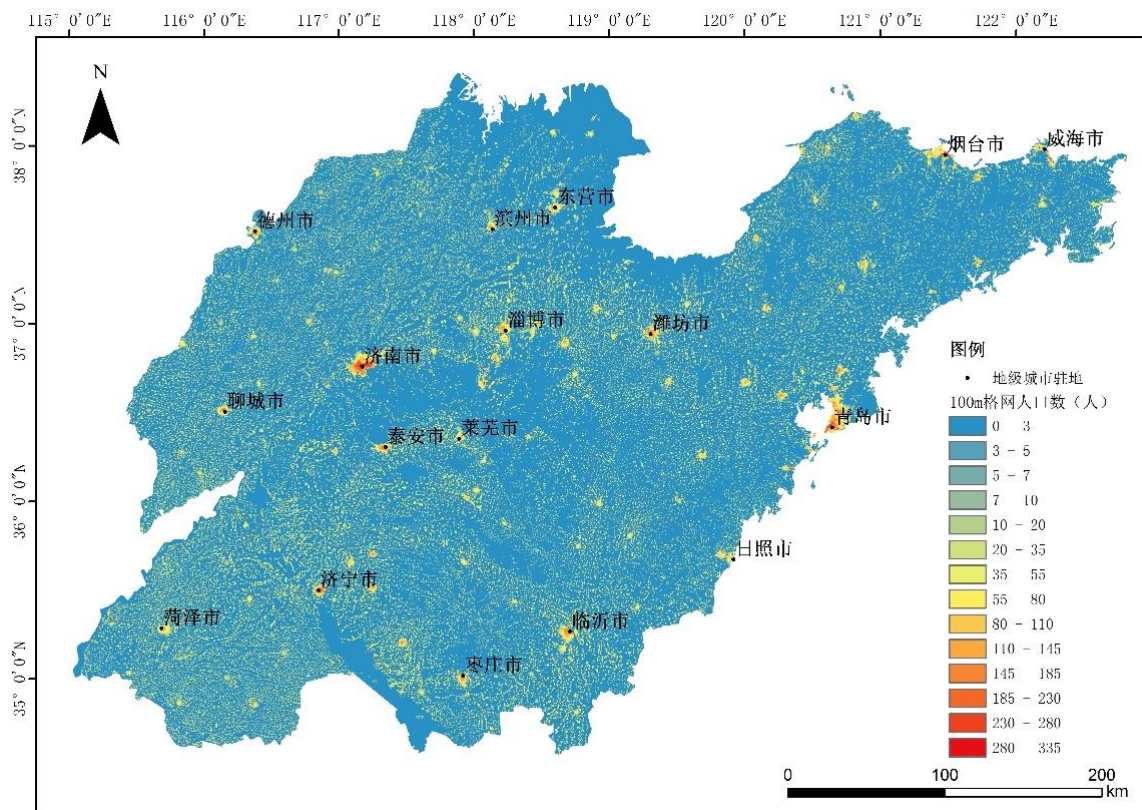
2 数据样本描述

经过数据处理得到 2000 年和 2010 年山东省 100 m 格网人口空间分布数据集 (图 3、图 4)。本数据集比较清晰地反映了山东省的人口分布特点, 高人口密度主要集中在济南、青岛、潍坊、淄博、烟台、德州、威海和济宁等地级城市驻地, 济南和青岛两地的人口密度和数量最高; 鲁北沿海、鲁中山地丘陵区 and 济宁南部的人口分布较为稀少。从 2000 年到 2010 年的人口分布变化来看, 地级城市高人口密度区有明显的扩大, 2010 年的最大人口密度值较 2000 年有所提高。

3 数据质量控制和评估

3.1 质量控制

本文利用街道/乡镇级人口统计数据作为基础数据、运用城乡二级分区建模方法、EVI 对 DMSP/OLS 夜间灯光数据进行修正等做法保证空间化产品的精度。为对空间化结果做出准确的评价, 本文对空间化模型进行验证并对空间化结果进行同类产品对比分析。验证和对比结果如下文所述。



3.2 质量评估

(1) 模型精度验证

在模拟格网单元内的人口之前, 检验空间化模型的精度。利用方程式 1-4 模拟其余地区乡镇(街道)单元内的人口, 按式(7) 计算其余地区街道/乡镇级人口统计值和模拟值之间的相对误差。结果显示, 线性回归模型中 2000 年和 2010 年的其余地区分别有 78.0%和 78.4%的行政单元的相对误差绝对值小于 20%, MLP 人工神经网络模型中 2000 年和 2010 年的其余地区分别有 71.0%和 72.7%的行政单元的相对误差绝对值小于 20%。可见在本文人口数据空间化模型构建中, 线性回归模型比 MLP 人工神经网络模型更有优势。

将线性回归模型的相对误差分布到空间(图 5), 可以看出误差较大(相对误差绝对值大于 20%, 图中蓝色和红色区域)的行政单元主要分布于地级城市周边和北部沿海区域, 且 2000 年的相对误差在中部山地丘陵区也分布较多。主要原因有以下两点: ①在城市周边和山地丘陵区容易出现城镇用地和农村居民点用地的错分现象; ②在沿海地区存在港口码头, 导致夜间灯光和人口分布的相关性降低。

$$\delta = \frac{pop_i^{sta} - pop_i^{mo}}{pop_i^{sta}} \times 100\% \quad (7)$$

式中: pop_i^{sta} 为其余地区第 i 个行政单元的人口统计值, pop_i^{mo} 为其余地区第 i 个行政单元的人口模拟值。

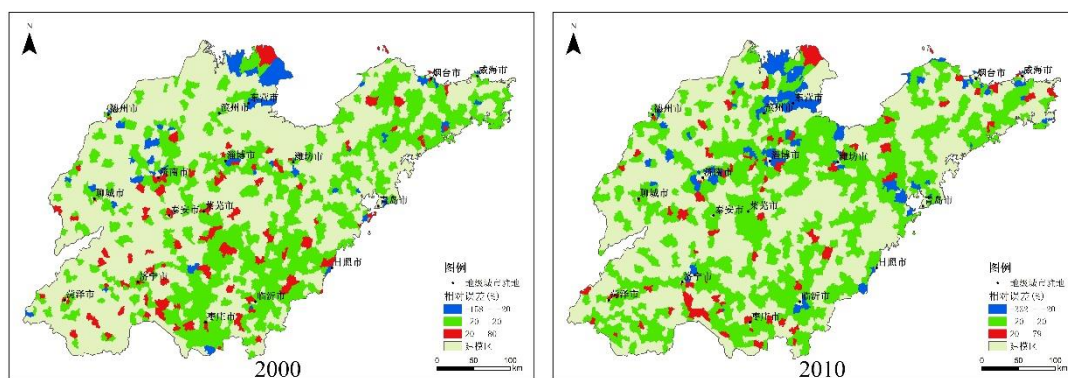


图 5 模型精度检验结果空间分布图

(2) 空间化结果对比

利用山东省街道/乡镇级人口统计数据对 Chpop、GPW 和 Worldpop 进行精度评价, 计算 Chpop2000、GPW2000、WorldPop2000、Chpop2010、GPW2010 和 Worldpop2010 六个数据集街道/乡镇级行政单元尺度内的人口模拟值和统计值之间的皮尔逊相关系数(R)、平均绝对误差(MAE)、均方根误差(RMSE)和相对均方根误差(%RMSE), 其中%RMSE 是将均方根误差除以人口普查数据的平均值得到, 可以反映模型模拟的精度高低^[12]。MAE、RMSE 和%RMSE 的计算公式分别如式(8)、式(9)和式(10)所示, 计算结果如表 2 所示。结果显示 WorldPop 的人口模拟值最精确。

$$MAE = \frac{1}{N} \sum_{i=1}^N |PE_i - PS_i| \quad (8)$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (PE_i - PS_i)^2} \quad (9)$$

$$\%RMSE = \frac{RMSE}{\frac{1}{N} \sum_{i=1}^N PS_i} \times 100\% \quad (10)$$

式中， PE_i 为第 i 个乡镇（街道）内的人口模拟值， PS_i 为第 i 个乡镇（街道）内的人口统计值， N 山东省境内街道/乡镇级行政单元个数。

表 2 山东省格网化人口分布数据集的人口模拟值与统计值的评价指标

数据集	R	MAE (人)	RMSE (人)	%RMSE (%)
Chpop2000	0.703**	16940	38096	77.58
GPW2000	0.920**	13283	22579	45.95
WorldPop2000	0.934**	13042	22428	45.64
Chpop2010	0.599**	19243	30748	57.89
GPW2010	0.796**	13832	24233	45.63
WorldPop2010	0.858**	11756	21225	39.96

注：**表明在0.01水平（双侧）上显著相关。

利用目前可开放获取的精度相对最高的人口空间分布数据集 WorldPop 对本文产出的 SDpop2000 和 SDpop2010 进行对比验证。分别统计 WorldPop 和 SDpop 在 10 km 格网内的人口值，采用散点图、R、MAE、RMSE 和 %RMSE 来衡量两者的差异。MAE、RMSE 和 %RMSE 分别采用式（8）、式（9）和式（10）计算，其中 PE_i 和 PS_i 分别带入 10 km 格网内 SDpop 和 WorldPop 的人口模拟值， N 代表 10 km 格网的个数。检验结果如表 3 所示，SDpop 与 WorldPop 的散点图如图 6 所示。

表 3 山东省格网化人口空间分布数据验证结果

数据	R	MAE (人)	RMSE (人)	%RMSE (%)
SDpop2000	0.932**	11598	20028	38.42
SDpop2010	0.948**	13028	22529	40.46

注：**表明在0.01水平（双侧）上显著相关。

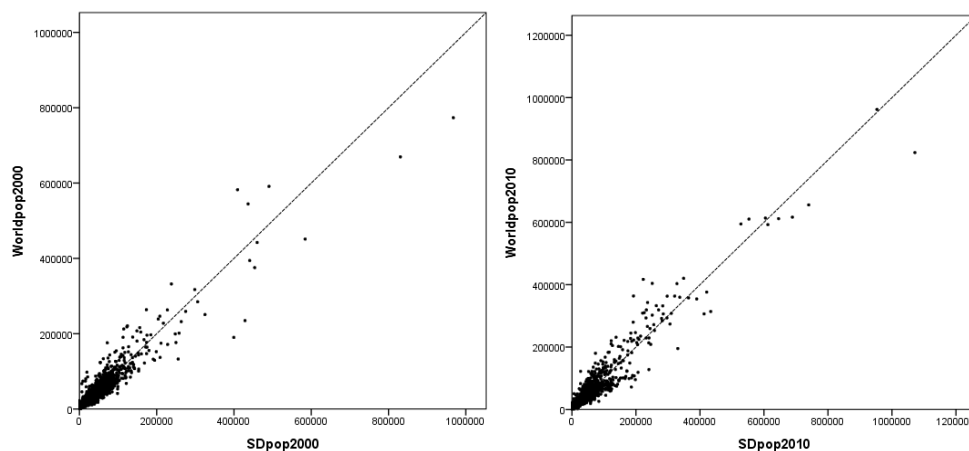


图 6 SDpop 与 WorldPop 的 10 km 格网人口值散点图

4 数据价值

格网化人口空间分布数据打破了行政单元的禁锢,实现了更精细的人口空间分布,这不仅对地理、资源、环境、生态、灾害等学科中和人口相关的跨学科研究具有重要意义,而且对政府部门进行土地使用、公共设施和开发计划等规划具有重要参考意义。本文通过对街道/乡镇级人口统计数据、独特分区方法和基于 EVI 修正的夜间灯光数据的使用,有效地提高了人口数据的空间化精度,产出的 2000 年和 2010 年山东省格网化人口空间分布数据集,丰富了人口格网化产品。

数据作者分工职责

王明明(1993—),男,河北沧州人,硕士研究生,研究方向为人口数据空间化。主要承担工作:基础数据处理、人口数据空间化及论文撰写。

王卷乐(1976—),男,河南洛阳人,博士,研究员,研究方向为资源环境数据集成与共享、GIS 和遥感应用。主要承担工作:数据集设计和技术指导。

参考文献

- [1] BRIGGS D J, GULLIVER J, FECHT D, et al. Dasymetric modelling of small-area population distribution using land cover and light emissions data[J]. Remote sensing of Environment, 2007, 108(4): 451-466.
- [2] ZENG C, ZHOU Y, WANG S, 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.
- [3] 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.
- [4] 柏中强. 基于乡镇尺度的中国人口分布格局及空间化问题研究[D]. 北京: 中国科学院大学, 2016.
- [5] 付晶莹, 江东, 黄耀欢. 中国公里网格人口分布数据集(PopulationGrid_China) [DB/OL]. 全球变化科学研究数据出版系统, 2014. DOI:10.3974/geodb.2014.01.06.V1.
- [6] GAUGHAN A E, STEVENS F R, HUANG Z, et al. Spatiotemporal patterns of population in mainland China, 1990 to 2010[J]. Scientific Data, 2016, 3: 160005.
- [7] 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.
- [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(10): 122-130.
- [9] 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.
- [10] 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(10): 42-49.
- [11] 卓莉, 张晓帆, 郑璟, 等. 基于 EVI 指数的 DMSP/OLS 夜间灯光数据去饱和方法[J]. 地理学报, 2015, 70 (8): 1339- 1350.

[12] 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. DOI:10.1371/journal.pone.0107042.

论文引用格式

王明明, 王卷乐. 山东省街道/乡镇级格网化人口空间分布数据集（2000、2010 年）[J/OL]. 中国科学数据, 2019, 4(3). (2019-04-18). DOI: 10.11922/csdata.2019.0005.zh.

数据引用格式

王明明, 王卷乐. 山东省街道/乡镇级格网化人口空间分布数据集（2000、2010 年）[DB/OL]. Science Data Bank, 2019. (2019-02-28). DOI: 10.11922/sciencedb.736.

Gridded data of township-level population distribution in Shandong Province (2000 & 2010)

Wang Mingming^{1,2}, Wang Juanle^{2,3*}

1. School of Civil and Architectural Engineering, Shandong University of Technology, Zibo 255049, P. R. 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, P. R. China

3. Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing 210023, P. R. China

*Email: wangjl@igsrr.ac.cn

Abstract: Capturing a fine-scale population spatial distribution, gridded data is of great significance for supporting research on geography, resources, environment, ecology, disasters, and sustainable development. On the basis of urban-rural partition and township-level demographic data, this study uses the spatialization method that combines enhanced vegetation index adjusted nighttime light index and land use data to produce the gridded population data of Shandong Province, namely, SDpop2000 and SDpop2010 (100 m resolution). The model verification results showed that 78.0% of the administrative units in 2000 and 78.4% in 2010 had an absolute relative error of less than 20%. When validated against global WorldPop population data products, our dataset demonstrated a correlation coefficient and a %RMSE of 0.932 and 38.42% in 2000 respectively, and 0.948 and 40.46% in 2010.

Keywords: population; spatialization; nighttime light data; land use data; township-level; Shandong Province

Dataset Profile

Title	Gridded data of township-level population distribution in Shandong Province (2000 & 2010)
Data corresponding author	Wang Juanle (wangjl@igsnrr.ac.cn)
Data authors	Wang Mingming, Wang Juanle
Time range	2000, 2010
Geographical scope	Shandong Province, China (34°22'54"N–38°24'0.6"N, 114°47'30"E–122°42'18"E)
Spatial resolution	100 m
Data volume	36.07 MB
Data format	*.tif, *.jpg
Data service system	< http://www.sciencedb.cn/dataSet/handle/736 >
Sources of funding	Strategic Priority Research Program of the Chinese Academy of Sciences (XDA19040501); Scientific Data Project of the 13 th Five-Year Informatization Program of the Chinese Academy of Sciences (XXH13505-07).
Dataset composition	The dataset is composed of two subsets: (1) Gridded data of population spatial distribution in Shandong Province (2000 & 2010), with a data volume of 13.42 MB. Each data document is recorded as SDpopYYYY.tif, where YYYY represents the year. (2) Gridded map of population spatial distribution in Shandong Province (2000 & 2010), with a data volume of 22.65 MB. Each data document is recorded as SDpopYYYY.jpg.