基于随机森林的人口密度模型优化试验研究

刘 艺[1]，杨歆佳[1]，刘劲松[1, 2, 3, 4*]

1. 河北师范大学资源与环境科学学院，石家庄 050024；2. 河北省环境变化遥感识别技术创新中心，石家庄 050024；3. 河北省环境演变与生态建设实验室，石家庄 020024；4. 河北省地理科学实验教学示范中心，石家庄 050024

摘 要：人口密度是表征人口分布特征的定量指标。近 30 年来，栅格人口密度模型成为人口密度研究的重要领域，并形成了“自上而下的人口统计数据分解算法”和“自下而上的人口调查数据估计算法”。其中，基于随机森林的人口密度模型备受关注。笔者以石家庄为例，通过对比多套人口密度数据集，发现基于随机森林的人口密度模型仍存在区群谬误、MAUP、混淆人口分布规律、遴选影响因子欠缜密等问题，围绕这些问题，探讨了基于随机森林的人口密度模型的优化方案，即：以禀赋分区为建模单位，分区构建随机森林模型，以避免在人口密度成果数据集中混淆人口分布法则；以公顷栅格为采样单元，分区开展随机采样，避免模型产生区群谬误问题，避免样本数据质量受到 MAUP 影响；分区开展遴选影响因子制图试验，避免将错误因子引入模型。优化方案为改善人口密度模型的信度和效度提供了系统思路。

关键词：人口密度；随机森林模型；禀赋分区；随机采样；影响因子；石家庄
DOI: https://doi.org/10.3974/geodp.2020.04.15

1 前言

人口密度是表征区域人口分布特征的定量指标，人口密度图是揭示人口分布规律的基本依据，人口密度变化对区域政治、经济、文化、资源、环境等会产生重要影响[1]，也是全球土地覆被变化的关键驱动因子[2–3]。因此，开展人口密度研究不仅可以揭示人口分布规律，也是评估全球变化效应和灾害风险等级、制定国土空间规划、优化资源配置等重大决策的关键基础性工作[4]。

20 世纪 90 年代以前，世界各国均使用矢量格式人口密度图表征人口分布特征。由于人口普查数据获取频率低（许多国家每 10 年进行一次人口普查）、时效性差、步调不一致（受战争、瘟疫等影响，有的国家不能按时发布人口普查数据，有的国家会延迟发布）[5–6]，故难以编制人口数据同步的矢量格式全球人口密度图。

矢量格式人口密度图的人口统计单元一般采用行政区划单元或经纬网格单元[7–8]。无论是行政区划单元，还是经纬网格单元，均与自然地理单元（流域、样带等）在形状、大小、位置等方面不相匹配，人口密度的转换运算总会受到可塑性面积单元问题（The Modifiable

收稿日期：2020-10-30；**修订日期：**2020-12-16；**出版日期：**2020-12-24
基金项目：国家自然科学基金（41671138, 42071167, 40871073）；河北省自然科学基金（D2007000272）
***通讯作者：**刘劲松，河北师范大学资源与环境科学学院，liujinsong@hebtu.edu.cn
引用格式：[1] 刘艺, 杨歆佳, 刘劲松. 基于随机森林的人口密度模型优化试验研究[J]. 全球变化数据学报, 2020, 4(4): 402–416. https://doi.org/10.3974/geodp.2020.04.15.

Areal Unit Problem，MAUP）的困扰[9–10]；矢量格式人口密度图还假设制图单元内的人口密度相同，掩盖了制图单元内人口分布的异质性特征。由于上述两个原因，真实人口分布规律常被矢量格式人口密度图扭曲[11]，导致人口、自然、资源、环境数据的叠加分析大打折扣[5]。由于全球变化研究须借助基于栅格数据的分析模型，故全球环境变化人文因素计划（The Human Dimensions of Global Environmental Change Programme，HDP）第 3 工作组在 1990 年代初，倡议研制全球栅格人口密度图[12]。

2　国内外出版的典型人口密度数据集

覆盖全球的典型人口密度数据集主要包括：GPW（Gridded Population of the World）、GRUMP（Global Rural Urban Mapping Project）、LandScan（LandScan Global Population Database）、GHS-POP（Global Human Settlement Layer-Population）、WPE（World Population Estimate）、WorldPop、HYDE（History Database of the Global Environment Population Grid）。覆盖我国的典型人口密度数据集主要是 CnPOP（表 1）。

隶属于哥伦比亚大学（Columbia University）的美国国际地球科学信息网络中心（Center for International Earth Science Information Network，CIESIN）是编制全球栅格人口密度数据集最早的世界数据中心和美国国家航空航天局社会经济数据中心，其不仅独立研发了每 5 年间隔的世界 1 km 人口密度栅格数据集，而且还和其他科研机构合作，联合编制了 GRUMP、GHS-POP、WPE 等人口密度数据集[13]。

CIESIN 运用面积加权模型，生产了 GPW 人口密度数据集，目前已演进到第 4 个版本，该版本提供了 2000、2005、2010、2015 和 2020 年 5 个年份的全球人口密度数据集，空间分辨率为 30 arcsec（ca.1km）。

CIESIN 与国际粮食政策研究所（International Food Policy Research Institute，IFPRI）、世界银行（The World Bank）、国际热带农业中心（International Center for Tropical Agriculture，CIAT）合作，采用线性回归模型，生产了 GRUMP 人口密度数据集，目前仅有第 1 版本，该版本提供了 1990、1995 和 2000 年 3 个年份的全球人口密度数据集，空间分辨率为 30 arcsec（ca.1km）[14]。

欧盟委员会联合研究中心（European Commission Joint Research Centre，JRC）与 CIESIN 合作，采用线性回归模型，生产了 GHS-POP 人口密度数据集，提供了 1975、1990、2000 和 2015 年 4 个年份的全球人口密度数据集，每个数据集又分别提供 4 种空间分辨率（250 m、1 km、9 arcsec、30 arcsec）的数据[15]。

美国环境系统研究所（Environment Systems Research Institute，ESRI）与 CIESIN 合作，采用线性回归模型，生产了 WPE 人口密度数据集，提供了 2013、2015 和 2016 年 3 个年份的全球人口密度数据集，其中，2013 和 2015 年数据集的空间分辨率为 250 m，2016 年数据集的空间分辨率为 150 m[16]。该数据集仅对 ESRI 正版软件用户开放。

荷兰环境评估署（Netherlands Environmental Assessment Agency，PBL）采用线性回归模型，生产了 HYDE 人口密度数据集，目前已演进到 3.2 版本，提供了自 10000aBC 到 2016AD 的全球人口密度数据集，其中，从 10000aBC 年到 1000aBC，每逢 1000 年提供 1 套全球人

口密度数据集；从 0AD 到 1700AD，每逢 100 年提供 1 套全球人口密度数据集；从 1700AD 到 2000AD，每逢 10 年提供 1 套全球人口密度数据集；从 2000 年到 2016 年，每年提供 1 套全球人口密度数据集。上述数据集的空间分辨率均为 5 arcmin（ca.10km）[17]。HYDE 数据集在全球变化研究中发挥了独特作用。

美国橡树岭国家实验室（Oak Ridge National Laboratory，ORNL）采用线性回归模型，生产了 LandScan 人口密度数据集，提供了 2000–2019 每年的全球人口密度数据集，其空间分辨率为 30 arcsec（ca.1km）[18]。

隶属于南安普顿大学（University of Southampton）的 WorldPop 采用随机森林模型，生产了 WorldPop 人口密度数据集，自 2000–2020 年，每年提供 1 套全球人口密度数据集，其空间分辨率为 30arcsec（ca.1km）[19]。

中国科学院地理科学与资源研究所利用线性回归模型，研发了中国公里格网人口密度数据集(CnPop)，提供了 2005 年和 2010 年中国人口密度数据集，其空间分辨率为 1 km[20–21]。

表 1　典型人口密度数据集（参考[4]，有改动）

数据集	研发机构	算法	空间分辨率	成图年份	数据共享网址
GPW v4.11	CIESIN; Columbia University	面积加权模型	30 arcsec (ca.1 km)	2000; 2005; 2010; 2015; 2020	https://sedac.ciesin.columbia.edu/data/collection/gpw-v4
GRUMP v1	CIESIN; IFPRI; The World Bank; CIAT	线性回归模型	30 arcsec (ca.1 km)	1990; 1995; 2000	https://sedac.ciesin.columbia.edu/data/collection/grump-v1
GHS-POP	JRC; CIESIN	线性回归模型	250 m	1975; 1990; 2000; 2015	https://ghsl.jrc.ec.europa.eu/ghs_pop.php
WPE	ESRI; CIESIN	线性回归模型	250 m 250 m 150 m	2013 2015 2016	https://sites.google.com/ciesin.columbia.edu/popgrid/find-data/esri
HYDE v3.2	PBL	线性回归模型	5 arcmin (ca.10 km)	10000BC–2016	https://themasites.pbl.nl/tridion/en/themasites/hyde/ download/index-2.html
LandScan	ORNL	线性回归模型	30 arcsec (ca.1 km)	2000–2019	https://landscan.ornl.gov/
WorldPop	WorldPop, University of Southampton	随机森林模型	30 arcsec (ca.1 km)	2000–2020	https://www.worldpop.org/
CnPOP	IGSNRR, CAS	线性回归模型	1 km	2005; 2010	https://doi.org/10.3974/geodb.2014.01.06.V1

目前，上述 7 个主流全球人口密度数据集（GPW、GRUMP、LandScan、GHS-POP、WPE、HYDE、WorldPop）已被广泛应用于灾害风险评估[22–25]、土地利用变化[26–28]、公共卫生管理[29–31]、环境变迁中的人为影响[2, 32–35]等工作中，在全球变化研究和环境治理工作中发挥了基础性支撑作用。

国内外学者通过对比评估认为，基于随机森林的人口密度模型生产的 WorldPop 人口密度数据集具有较高效度[4, 36–37]，基于随机森林的人口密度模型显示出显著优势。

30 多年来，我国学者也研发了栅格人口密度模型[8, 10, 37–51]，但这些模型或用于区域尺

度或用于地方尺度。在覆盖全球的人口密度数据集中，尚未看到属于我国学者研发的人口密度数据集。

3　栅格人口密度模型

近 30 年来，栅格人口密度模型层出不穷，并逐步形成了两类栅格人口密度模型[52]，即“自上而下的人口统计数据分解算法”和“自下而上的人口调查数据估计算法”。

3.1　自上而下的人口统计数据分解算法

“自上而下的人口统计数据分解算法”是在人口普查工作开展较好的国家或地区，按照某种演绎规则，将区域人口总数分解到人口普查（或人口登记）单元所覆盖的各个栅格内的算法。此类算法主要包括：面积加权模型[8, 43, 47, 53]、距离衰减模型[38, 40, 42, 45, 54]、线性回归模型[2, 39, 41, 43, 46, 48–49, 55–60]和随机森林模型[6, 37, 50–52, 61–62]。

3.1.1　面积加权模型

面积加权模型简单易行。CIESIN 运用此类模型，编制了 GPW 栅格人口密度数据集。该模型的优点是，利用面积加权模型，能够开展人口密度的尺度下推和尺度上推计算[43]。例如：从石家庄人口矢量数据集出发，下推得到街区人口矢量数据集，进而下推得到最小粒度人口密度数据集（图 3a）。依托最小粒度人口密度数据集，利用圆形滤波算法，开展人口密度尺度上推计算，可系统探讨多尺度人口密度推绎中的 MAUP 问题[47]，并定性建构区域人口分布的扎根理论。该模型的缺点是，当栅格面积较大时，栅格内部的人口密度异质性特征被掩盖；模型没有考虑人口密度与影响因子之间的量化关系，无法定量解释人口密度影响机制。

3.1.2　距离衰减模型

距离衰减模型适合刻画城镇及周边区域的人口分布特征，但若忽略模型的适用条件，将模型延伸用于计算区域（含乡村）人口密度，则会大幅降低人口密度图的效度。此类模型属于插值算法，虽考虑了距离因子，但无法定量解释人口密度影响机制。

3.1.3　线性回归模型

CIESIN 等编制的 GRUMP 数据集[14]、JRC 和 CIESIN 联合编制的 GHS-POP 数据集[15]、ORNL 编制的 LandScan 数据集[18]、ESRI 编制的 WPE 数据集[16]、PBL 编制的 HYDE 数据集[17]和中国科学院地理科学与资源研究所编制的 CnPOP 数据集[20–21]均是采用线性回归模型生产得到。

线性回归模型引入了人口密度影响因子，不同模型引入的因子种类、因子数量、引入方式均不相同，各自生产的栅格人口密度数据集差别较大（表 2）。线性回归模型借助影响因子数据集，适合生产缺少人口普查数据国家或地区的人口密度数据集，但此类模型的准则效度（Criterion Validity，即采用不同人口密度模型测算人口密度时，将其中一个人口密度数据集设定为准则数据集，其他人口密度数据集与准则数据集作比较，所得相关系数称为准则效度）不高。利用线性回归模型编制的人口密度数据集普遍存在过低估计城市区域人口密度、过高估计乡村区域人口密度的问题[37, 63]，这说明将人口密度与影响因子之间设定为线性依赖关系值得商榷。2015 年，联合国可持续发展目标（Sustainable Development

Goals，SDGs）认为，在整合人口、资源、环境数据的过程中，人口密度模型的信度（Reliability，即采用同一人口密度模型，测量结果的一致性或稳定性）和效度（Validity，即人口密度模型能够测出人口密度的准确程度）依然亟待改进[6]。

3.1.4 随机森林人口密度模型

近5年来，国内外学者利用随机森林模型，结合覆盖全球的地理大数据，研究了网格人口密度模型的非线性建模方法，显著改善了人口密度数据集的质量。

随机森林模型属于机器学习的集成算法，善于非线性计算[64]。随机森林中的“树”是利用训练子集构建的CART树（Classification And Regression Tree），训练每株CART树所需的训练子集均是从总训练样本数据集中借助放回抽样（Bootstrap）随机构造而成。每株CART树将制图区域分割成了若干类型区，每个类型区的人口分布概率值为落入该区内所有样本人口密度的平均值。假设随机森林生成了n株CART树，就相当于对人口密度制图区域形成了n套相互独立的空间分割方案，即每个栅格得到n个人口分布概率值，再计算每个栅格人口分布概率的算术平均值（即每个栅格的人口分布概率预测值），显然人口分布概率预测值总会介于总训练样本数据集中人口密度最高值和最低值之间。随机森林模型还利用袋外数据（Out-Of-Bag，OOB）度量人口密度影响因子的重要性，若某个影响因子的重要值越大，则说明该影响因子重要性越高。由于随机森林模型能准确刻画人口密度与影响因子之间的非线性关系，且能对人口密度影响因子的重要性进行排序，因此，随机森林模型具备遴选人口密度影响因子的潜力。随机森林模型已成为生产栅格人口密度数据集的主流算法。英国南安普顿大学就利用随机森林模型编制了覆盖全球的WorldPop人口密度数据集[19]。

3.2 自下而上的人口调查数据估计算法

若假设没有开展人口普查国家的人口分布规律与周边开展过人口普查国家的人口分布规律相仿，则将没有开展人口普查国家的影响因子数据集，代入到邻近开展过人口普查国家的基于随机森林的人口密度模型中，就能编制缺少人口普查工作国家或地区的人口密度数据集。基于随机森林的人口密度模型催生了“自下而上的人口调查数据估计算法”。

“自下而上的人口调查数据估计算法”就是在没有开展人口普查的国家或地区，通过开展若干典型微型社区人口调查，结合典型微型调查社区的人口密度影响因子数据集，以微型调查社区为采样单元，构造人口密度训练样本数据集，训练得到随机森林模型，并借助影响因子数据集，生产缺失人口普查工作国家或地区的人口密度数据集[52]。

3.3 遴选人口密度影响因子的困惑

随着地理大数据的不断发展，栅格人口密度模型引入的影响因子数据集逐步增多（表2）。在生产LandScan、WorldPop和HYDE数据集时，运用了年均温、年降水、DEM、地形起伏度、土地覆被、水体等数据集，考虑了自然条件对人口密度的影响；生产LandScan、WPE和WorldPop数据集时，考虑了道路对人口密度的影响；生产WorldPop数据集时，运用了兴趣点（Point of Interesting，POI）数据集[65]，考虑了生产、生活和消费设施对人口密度的影响。由于不同的栅格人口密度模型引入的影响因子数据集种类不同、数量不同、引用方式不同，故不同模型生产的人口密度数据集存在显著差异。如果同一项目引用不同的人口密度数据集，其研究结论往往很不一致[36, 66–69]。

在栅格人口密度模型的研发过程中，长期受到数据短缺的困扰，为了生产全球人口密度数据集（任务驱动），每个研究机构都不得不遵守数据可得性原则。显然早期栅格人口密度模型引用哪些影响因子数据集，必然是任务进度和数据可得之间相互妥协的结果。随着地理空间大数据的逐步完善和空间位置服务的快速发展，LandScan 和 WorldPop 初步摆脱了影响因子数据集短缺的局面（表 2）。但栅格人口密度模型究竟应该引入哪些影响因子？如何引入影响因子？不仅缺乏理论指导，而且缺乏遴选手段，突出表现为：应该用什么理论指导遴选人口密度影响因子？究竟哪些影响因子是人口密度的关键影响因子？不同的影响因子是否在全球范围内均有影响效应？是否应该分区遴选人口密度的影响因子？如果需要分区，应该依据什么进行分区？

表 2　典型人口密度数据集所选用的影响因子（参考[4]，有改动）

栅格人口数据集	人口密度	人口密度影响因子								
		道路	土地覆被	建筑物结构	城区	夜光影像	基础设施	保护区[a]	自然条件[b]	水体
GPW	x							x		x
GRUMP	x				x	x		x		x
LandScan	x	x	x	x	x		x	x	x	x
GHS-POP	x			x						
WPE	x	x	x		x					x
WorldPop	x	x	x	x	x	x	x	x	x	x
HYDE	x								x	x
CnPOP	x		x							x

注：a 陆上保护区没有被排除在计算掩膜之外，但保护区人口数经常为 0 或无数据；
注：b 自然条件包括气候、地形、海拔高度等。

4　基于随机森林的人口密度模型的优化试验研究

笔者以石家庄市为例，开展了基于随机森林的人口密度模型的优化试验研究。

4.1　研究区概况

石家庄市（含辛集市）包括 12 个县、5 个县级市和 1 个国家级高新技术开发区；208 个乡镇，64 个街道办事处；192 个居委会，4,317 个行政村。户籍人口 943 万，其中市区户籍人口 232 万（含井陉矿区，9.73 万）。

村人口数据集由石家庄市公安局提供，汇总时间为 2007 年 5 月 1 日 0 时。户籍人口统计信息包括：村名、总户数、总人口、男性人口、女性人口、非农业人口、农业人口。村界和街区矢量数据集由河北省国土资源厅提供，为 2006 年版。地名数据由石家庄市地名办公室提供。

4.2 基于随机森林的人口密度模型的探索试验研究

最小粒度人口密度图（图 3a）是利用面积加权模型，通过尺度下推获得（图 1）。R=12（滤波半径为 1.2 km）的人口密度图（图 3b）是基于最小粒度人口密度图，借助圆形滤波模型，通过尺度上推获得（图 1）。

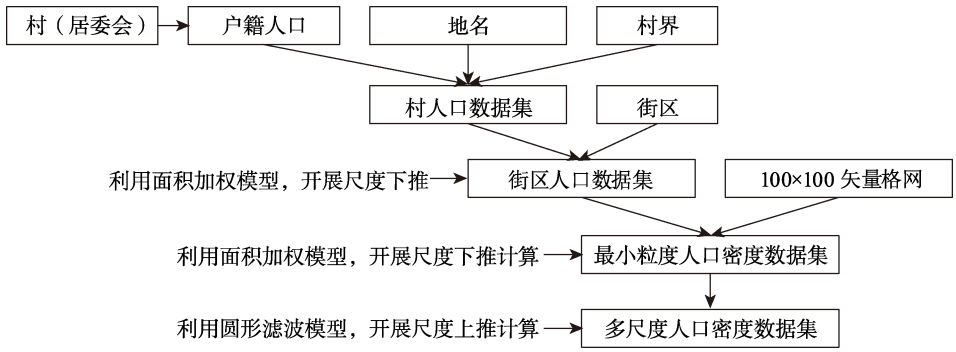


图 1 多尺度人口密度数据集计算流程图

根据图 3b，归纳出石家庄市人口分布扎根理论。即：(1) 自然河流对周边人口密度有显著影响作用，其中，为躲避洪水灾害，平原自然河流两侧往往是人口密度低值区；为取水方便，山区自然河流两侧往往是人口密度高值区。(2) 人工河流（减河、灌渠等）对周边人口密度没有显著影响。(3) 平原人口密度高，山区人口密度低。(4) 城市人口密度高，乡村人口密度低。(5) 在平原地区，明清时期的废弃古自然河道两侧依然是人口密度低值区。

图 3c、3d、3e 均是采用 WorldPop 推荐的基于随机森林的石家庄人口密度模型（图 2）计算获得。区别在于，图 3c、3d 是项目组基于 2007 年户籍人口数据，按乡采样，获得的计算结果[71]；图 3e 是 WorldPop 利用 2010 人口普查数据，按县采样，获得的计算结果[19]。

依照石家庄市人口分布扎根理论，审视利用随机森林模型测算的人口密度图（图 3d），发现存在混淆人口分布法则的现象。例如：在大沙河、磁河的平原部分，自然河流两侧本应是人口密度低值区，但竟然呈现人口密度高值特征。这是构建随机森林模型所用训练样本数据集将山区训练样本和平原训练样本混在一起，所导致的必然结果。如果对山区和平原分别采样，分区构建随机森林模型，则理论上能够克服此类问题。

如果按照县、乡等行政区划单元采样，训练样本中的人口密度和影响因子数据就必须通过聚合运算获得（图 2），样本数据就必然会受到 MAUP（例如：人口密度属于空间定比数据，人口密度值往往随着统计单元面积的变化而变化，这种现象称为人口密度的可塑性面积单元问题）困扰。

随机森林模型的采样单元多是县、乡等行政区划单元（图 2），而输出单元多是公里栅格或公顷栅格，理论上讲，这种用一种集群的分析单位收集资料，而用另一种集群的分析单位下结论的现象，通常会产生区群谬误（ecological fallacy）问题。

基于随机森林的人口密度模型（图 2）并没有讨论如何引入和为什么引入影响因子数据集的问题。①若将人工河流和自然河流笼统作为河流因子，则计算结果在人工河流两侧就会出现人口密度或低或高的特征（图 3d、3e），这不符合图 3b 揭示的人口分布扎根理论。②若将夜光影像作为影响因子，如何克服夜光影像的“灯撒效应”（Blooming Effect）就成了难题。③当仅把 24 种 POI 距离因子作为人口密度影响因子时，则所获得的人口分布概率图（图 3c）定量描述了人口分布的点轴分布特征，说明 POIs 距离组合数据集能够很好地表征创新禀赋特征，如用其替代夜光影像，有望克服“灯撒效应”的不良影响。

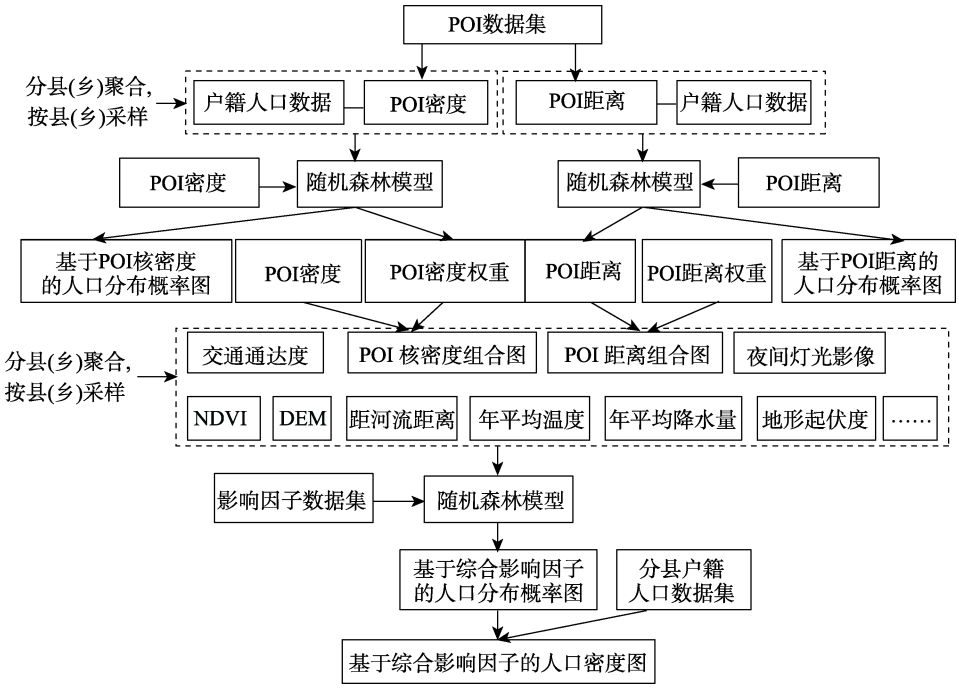


图 2　基于随机森林的人口密度模型的计算流程图

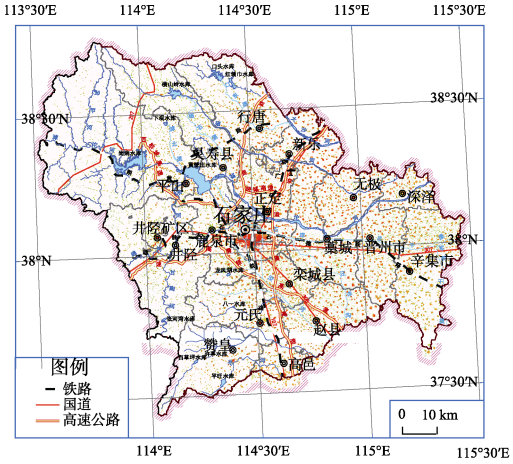
5　讨论与结论

5.1　讨论

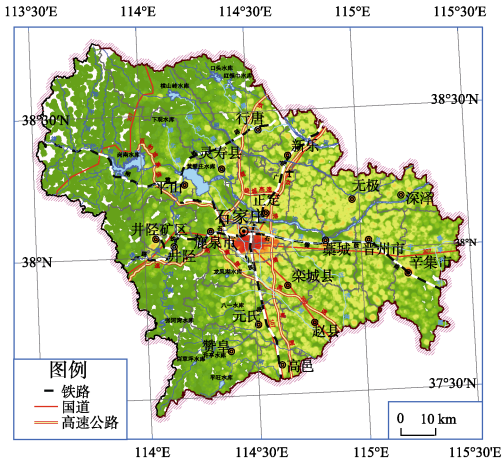
5.1.1　分区构建栅格人口密度模型问题

事实上，早在 20 世纪 80 年代，胡焕庸先生就已明确指出，不仅要通过综合研究探索人口分布规律，而且要分区探讨人口分布和人口发展规律。他认为：人口、资源、经济、科技之间存在相互作用关系，它们之间必须综合、协调，单独计算某一类要素意义不大。他将中国划分为三带、八区，并分区进行了人口发展展望[71–73]。胡先生在人口地理研究中的分区思想与人地关系地域系统理论[74–77]、演化经济地理学理论[78–80]不谋而合，为深化人口密度研究指明了方向和路径。优化基于随机森林的人口密度模型，优选人口密度影响因子应当借鉴分区思想。

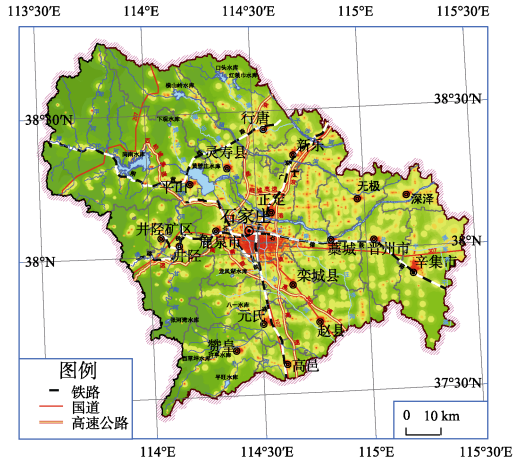
自然地理区划单元已成为利用地理空间大数据发现知识的基础对象。自然地理区划不仅是区域划分的结果，也是认识地理特征和发现地理规律的科学方法[81]。区划把地理特征相同的地方归在一起，相异的地方另划一区，既可避免支离破碎的弊病，也可省去重复论述的麻烦[82]。在信息时代，自然地理区划单元已成为利用地理空间大数据发现知识的对象与基础[81]。由于人口分布是区域人地要素协同演化的历史积累，区域综合禀赋不同，人口发展路径不同，不同的社会发展阶段会呈现不同的人口密度影响机制，因此，优化基于随机森林的人口密度模型，应该继承和发扬我国自然地理区划的学术传统，分区构建人口密度模型，分区编制人口密度图，分区阐述人口分布规律和人口发展规律[83]。



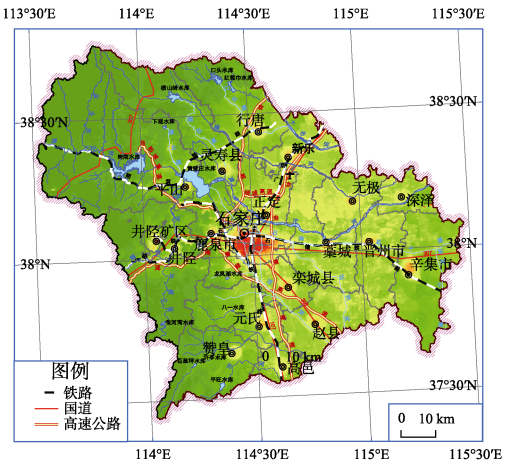
(a) 最小粒度人口密度图[47]



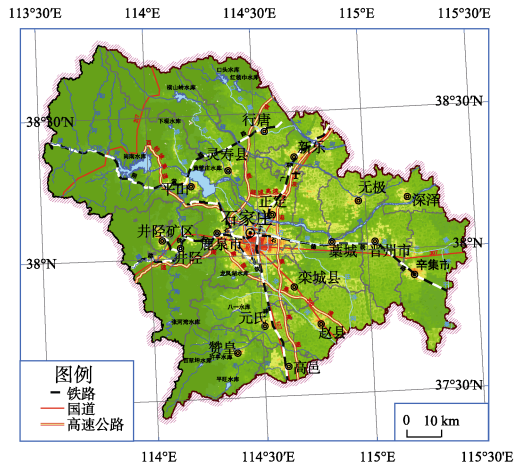
(b) R=12人口密度图[47]



(c) 基于POI距离因子的人口分布概率图[70]



(d) 基于综合禀赋因子的人口密度图[70]



(e) WorldPop人口密度图[19]

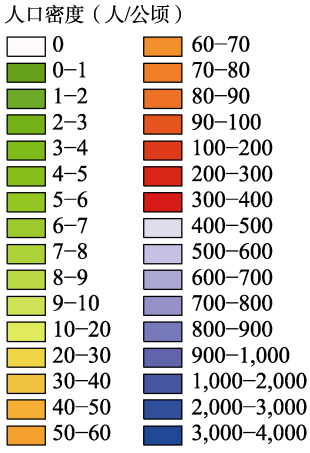


图 3 石家庄市基于不同禀赋条件的人口密度图和人口分布概率图

5.1.2　依据演化经济地理学理论，遴选人口密度影响因子问题

演化经济地理学认为：在农业文明时代，自然禀赋（海拔、地表崎岖度、水资源适宜度、农业生产潜力、年均温、年降水、年蒸发等）是区域发展的主导因素；在工业文明时代，经济禀赋（交通区位和城市区位等）是区域发展的主导因素；在信息文明时代，创新禀赋（知识、技术、网络、政策、制度等）是区域发展的主导因素[80]。胡焕庸线[7]是农业文明时代自然禀赋对区域发展锁定效应的必然结果，是自然禀赋（尤其是水资源）控制人口分布格局的经典案例；工业文明时代，中国西北半壁的交通比较成本并未降低，水资源短缺约束并未得到缓解；信息文明时代，中国西北半壁的创新禀赋仍呈点轴分布格局，这些禀赋的综合约束是过去 70 多年来“胡焕庸线整体稳定、东西半壁局部调整”的根本原因[84–90]。基于随机森林的人口密度优化模型应继承综合区划经典成果[91–92]和演化地理学的理论成就，结合区域所处发展阶段，围绕自然禀赋、经济禀赋和创新禀赋，采用定性与定量相结合的方法，分区遴选人口密度影响因子。融合多种研究范式，开展复杂地理研究[93–94]，中国人口地理学家将在优化基于随机森林的人口密度模型中发挥积极作用。

5.1.3　基于随机森林的人口密度模型优化方案

在开展过综合农业区划工作的国家或地区，鼓励利用综合农业区划图与城乡分布图叠加，形成综合禀赋分区方案。在没有开展综合农业区划工作的国家或地区，可通过地貌区划图与城乡分布图叠加，形成替代性的综合禀赋分区方案。

依托最小粒度人口密度图（图 3a）、自然禀赋（DEM、地形起伏度、距自然河流距离、NDVI、年均温、年降水量等）、经济禀赋（交通通达度、POIs 核密度组合图）、创新禀赋数据集（POIs 距离组合图），以公顷栅格为采样单元，分区开展随机采样，克服区群谬误问题和 MAUP；分区构建训练样本数据集，分区训练随机森林模型，分区测算人口密度，避免混淆人口分布规律。

开展系列人口密度制图实验，通过增删训练样本数据集中某一个人口密度影响因子或调整样本规模，分区筛查人口密度影响因子，分区明确最佳采样规模，是提升模型信度的可行途径。同时，结合现场考察，采用定性与定量相结合的对比分析方法，可以提升模型效度。此外，通过综合制图试验成果，最终可以确定基于随机森林的人口密度模型优化方案的技术细节（图 4）。

5.2　结论

近 30 年来，随着全球地理大数据不断完善和机器学习算法的不断进步，栅格人口密度模型逐步从面积加权、距离衰减等插值模型，进化为线性回归模型和随机森林模型。利用基于随机森林的人口密度模型编制的人口密度图精度更高，但基于随机森林模型的人口密度模型仍受到区群谬误、MAUP、混淆人口分布规律、遴选影响因子欠缜密等 4 类不确定性问题困扰。

将自然区划、地貌区划与城乡分布相融合，编制综合禀赋分区图；基于村户籍人口数据，利用面积加权模型，编制最小粒度人口密度图；利用滤波模型，开展尺度上推计算，形成区域人口密度扎根理论；从自然、经济和创新等禀赋因子中，遴选人口密度影响因子；以公顷栅格为采样单元，分区开展随机采样，克服样本采集单元面积显著大于输出单元面

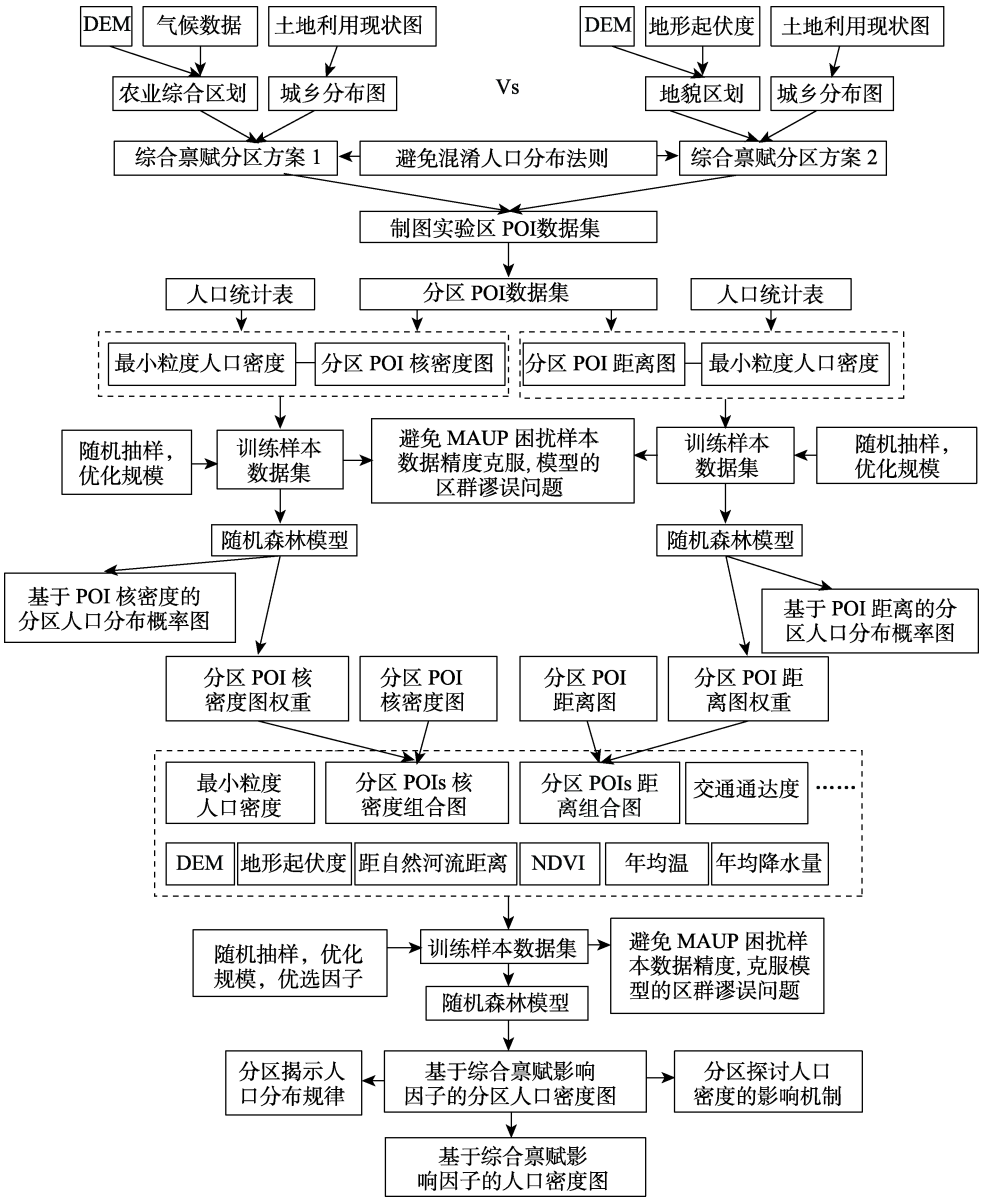


图 4 基于随机森林的人口密度模型的优化方案流程图

积的区群谬误问题，避免训练样本数据因聚合运算产生可塑性面积单元问题，将会改善基于随机森林的人口密度模型的训练样本数据质量；借助基于随机森林的人口密度模型，开展分组对照实验，分区筛选人口密度影响因子，分区推敲随机采样规模，分区编制人口密度图，避免在人口密度图中混淆人口分布法则，将进一步改善基于随机森林的人口密度模型的信度和效度。基于随机森林的人口密度模型优化方案将推动人口密度影响机制、人口分布规律、人口演化规律等基础理论研究工作。

参考文献

[1] 吴文恒, 牛叔文. 人口数量与消费水平对资源环境的影响研究[J]. 中国人口科学, 2009, 23(2): 66–73.

[2] Goldewijk, K. K., Ramankutty, N. Land cover change over the last three centuries due to human activities: the availability of new global data sets [J]. *GeoJournal*, 2004, 61(4): 335–344.

[3] Goldewijk, K. K., Beusen, A., Drecht, G. V., *et al*. The HYDE 3.1 spatially explicit database of human-induced global land-use change over the past 12,000 years [J]. *Global Ecology & Biogeography*, 2011, 20(1): 73–86.

[4] Leyk, S., Gaughan, A. E., Adamo, S. B., *et al*. 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.

[5] 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.

[6] Tatem, A. J. WorldPop, open data for spatial demography [J]. *Scientific Data*, 2017, 4(1): 1–4.

[7] 胡焕庸. 中国人口之分布——附统计表与密度图[J]. 地理学报, 1935, 2(2): 33–74.

[8] 张从宣. 用经纬网格单元编制人口密度图——以京津唐地区为例[J]. 中原地理研究, 1985, 4(2): 57–66.

[9] Openshaw, S. The Modifiable Areal Unit Problem [M]. Norwich, UK: Geobooks, 1983.

[10] 杨小唤, 江东, 王乃斌等. 人口数据空间化的处理方法[J]. 地理学报, 2002, 57(z1): 70–75.

[11] Liu, J. S., Wang, W., Xiang, H. B. The computational model of multi-scale population density [C]. *International Conference on Geoinformatics*, 2011: 1–4.

[12] Clarke, J. I., Rhind, D. W., Becket, C., *et al*. Population data and global environmental change [Z]. Barcelona Spain International Social Science Council Human Dimensions of Global Environmental Change Programme, 1992, 3(2): 147.

[13] Center for International Earth Science Information Network (CIESIN), Columbia University. Documentation for the gridded population of the world, version 4 (GPWv4), revision 11 data sets [EB/OL]. NASA Socioeconomic Data and Applications Center (SEDAC). https://sedac.ciesin.columbia.edu/downloads/docs/gpw-v4/gpw-v4-documentation-rev 11.pdf.

[14] Center for International Earth Science Information Network (CIESIN, Columbia University), International Food Policy Research Institute (IFPRI), The World Bank, *et al*. Global rural-urban mapping project, version 1: population density grid [EB/OL]. NASA Socioeconomic Data and Applications Center (SEDAC). https://sedac.ciesin.columbia.edu/data/set/ grump-v1-population-density, 2018-11-28.

[15] Florczyk, A. J., Corbane, C., Ehrlich, D., *et al*. GHSL data package 2019 [EB/OL]. European Commission Joint Research Center. https://ghsl.jrc.ec.europa.eu/documents/GHSL_Data_Package_2019.pdf?t=1478q532234372.

[16] Frye, C., Wright, D. J., Nordstrand, E., *et al*. Using classified and unclassified land cover data to estimate the footprint of human settlement [J]. *Data Science Journal*, 2018, 17(4): 1–12.

[17] Goldewijk, K. K., Beusen, A., Janssen, P. Long-term dynamic modeling of global population and built-up area in a spatially explicit way: HYDE 3.1 [J]. *The Holocene*, 2010, 20(4): 565–573.

[18] Oak Ridge National Laboratory. LandScan datasets [EB/OL]. UT-Battelle for the Department of Energy. https://landscan.ornl.gov/landscan-datasets.

[19] WorldPop, University of Southampton. Population density [EB/OL]. School of Geography and Environmental Science, University of Southampton; Department of Geography and Geosciences, University of Louisville; Departement de Geographie, Universite de Namur, Center for International Earth Science Information Network (CIESIN), Columbia University. https://www.worldpop.org/project/categories?id=18.

[20] 付晶莹, 江东, 黄耀欢. 中国公里网格人口分布数据集[J]. 地理学报, 2014, 69(s1): 41–44.

[21] 付晶莹, 江东, 黄耀欢. 中国公里网格人口分布数据集[J/DB/OL]. 全球变化数据仓储电子杂志, 2014. https://doi.org/10.3974/geodb.2014.01.06.V1.

[22] Smith, K. We are seven billion [J]. *Nature Climate Change*, 2011, 1(7): 331–335.

[23] World Resources Institute (WRI). World resources 2010–2011: Decision making in a changing climate — adaptation challenges and choices [J]. *Executive Summary*, 2011, 4(6): 305.

[24] Gleeson, T., Wada, Y., Bierkens, M. F. P., *et al*. Water balance of global aquifers revealed by groundwater footprint [J]. *Nature*, 2012, 488(7410): 197–200.

[25] Kibret, S., Lautze, J., Mccartney, M., *et al*. Malaria impact of large dams in sub-Saharan Africa: maps, estimates and predictions [J]. *Malaria Journal*, 2015, 14(1): 339.

[26] Goldewijk, K. K. Three centuries of global population growth: a spatial referenced population (density) database for 1700–2000 [J]. *Population and Environment*, 2005, 26(4): 343–367.

[27] Seto, K. C., Guneralp, B., Hutyra, R. Global forecasts of urban expansion to 2030 and direct impacts on biodiversity and carbon pools [J]. *Proceedings of the National Academy of Sciences of the United States of America*, 2012, 109(40): 16083–16088.

[28] Melchiorri, M., Florczyk, A. J., Freire, S., *et al*. Unveiling 25 years of planetary urbanization with remote sensing: perspectives from the global human settlement layer [J]. *Remote Sensing*, 2018, 10(5): 768.

[29] Sorichetta, A., Bird, T. J., Ruktanonchai, N. W., *et al*. Mapping internal connectivity through human migration in malaria endemic countries [J]. *Scientific Data*, 2016, 3(1): 5–11.

[30] Ouma, P. O., Maina, J., Thuranira, P. N., *et al*. Access to emergency hospital care provided by the public sector in sub-Saharan Africa in 2015: a geocoded inventory and spatial analysis [J]. *The Lancet Glob Health*, 2018, 6(3): 342–350.

[31] Thomson, D. R., Linard, C., Vanhuysse, S., *et al*. Extending data for urban health decision-making: a menu of new and potential neighborhood-level health determinants datasets in LMICs [J]. *Journal of Urban Health-bulletin of the New York Academy of Medicine*, 2019, 96(4): 514–536.

[32] Gaston, K. J., Blackburn, T. M., Goldewijk, K. K. Habitat conversion and global avian biodiversity loss [J]. *Proceedings of the Royal Society B*: *Biological Sciences*, 2003, 270(1521): 1293–1300.

[33] Houweling, S., VanderWerf, G. R., Goldewijk, K. K., *et al*. Early anthropogenic CH_4 emissions and the variation of CH_4 and $^{13}CH_4$ over the last millennium [J]. *Global Biogeochemical Cycles*, 2008, 22(1): 1–21.

[34] Ellis, E. C., Goldewijk, K. K., Siebert, S., *et al*. Anthropogenic transformation of the biomes, 1700 to 2000 [J]. *Global Ecology & Biogeography Letters*, 2010, 19(5): 589–606.

[35] Maisels, F., Strindberg, S., Blake, S., *et al*. Devastating decline of forest elephants in Central Africa [J]. *PLoS One,* 2013, 8(3): 1–17.

[36] 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): 1363–1378.

[37] 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.

[38] Wang, F. H., Guldmann, J. M. Simulating urban population density with a gravity-based model [J]. *Socio-economic Planning Sciences*, 1996, 30(4): 245–256.

[39] Lo, C. P. Modeling the population of China using DMSP operational linescan system nighttime data [J]. *Photogrammetric Engineering & Remote Sensing*, 2001, 67(9): 1037–1047.

[40] 冯健. 杭州市人口密度空间分布及其演化的模型研究[J]. 地理研究, 2002, 21(5): 635–646.

[41] 江东, 杨小唤, 王乃斌等. 基于RS、GIS的人口空间分布研究[J]. 地球科学进展, 2002, 17 (5): 734–738.

[42] 刘纪远, 岳天祥, 王英安等. 中国人口密度数字模拟[J]. 地理学报, 2003(1): 17–24.

[43] 金君, 李成名, 印洁等. 人口数据空间分布化模型研究[J]. 测绘学报, 2003, 32(3): 278–282.

[44] 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.
[45] Yue, T. X., Wang, Y. A., Liu, J. Y., *et al*. Surface modeling of human population distribution in China [J]. *Ecological Modelling*, 2005, 181(4): 461–478.
[46] 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.
[47] 刘劲松. 人口密度尺度推绎中可塑性面积单元问题的地理学解释[D]. 石家庄: 河北师范大学, 2009.
[48] 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.
[49] 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.
[50] 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.
[51] 王超, 阚瑷珂, 曾业隆等. 基于随机森林模型的西藏人口分布格局及影响因素[J]. 地理学报, 2019, 74(4): 664–680.
[52] Wardrop, N. A., Jochem, W. C., Bird, T. J., *et al*. Spatially disaggregated population estimates in the absence of national population and housing census data [J]. *Proceedings of the National Academy of Sciences of the United States of America*, 2018, 115(14): 3529–3537.
[53] Tobler, W., Deichmann, U., Gottsegen, J., *et al*. World population in a grid of spherical quadrilaterals [J]. *International Journal of Population Geography*, 1997, 3(3): 203–225.
[54] Clark, C. Urban population densities [J]. *Journal of the Royal Statistical Society*, 1951, 114(4): 490–496.
[55] Sutton, P. Modeling population density with night-time satellite imagery and GIS [J]. *Computers Environment & Urban Systems*, 1997, 21(3/4): 227–244.
[56] Sutton, P., Roberts, D., Elvidge, C., *et al*. A comparison of nighttime satellite imagery and population density for the Continental United States [J]. *Photogrammetric Engineering and Remote Sensing*, 1997, 63(11): 1303–1313.
[57] Dobson, J. E., Bright, E. A., Coleman, P. R., *et al*. LandScan: a global population database for estimating populations at risk [J]. *Photogrammetric Engineering and Remote Sensing*, 2000, 66(7): 849–857.
[58] Amaral, S., Monteiro, A. M. V., Câmara, G., *et al*. DMSP/OLS night-time light imagery for urban population estimates in the Brazilian Amazon [J]. *International Journal of Remote Sensing*, 2006, 27(5): 855–870.
[59] Bhaduri, B., Bright. E., Coleman, P., *et al*. LandScan USA: a high-resolution geospatial and temporal modeling approach for population distribution and dynamics [J]. *GeoJournal*, 2007, 69(1/2): 103–117.
[60] 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.
[61] 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): 1–22.
[62] 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(1): 1–11.
[63] Yang, X. C., Yue, W. Z., Gao, D. W. Spatial improvement of human population distribution based on multi-sensor remote-sensing data: An input for exposure assessment [J]. *International Journal of Remote Sensing*, 2013, 34(15): 5569–5583.
[64] Breiman, L. Random Forests [J]. *Machine Learning*, 2001, 45(1): 5–32.
[65] 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.
[66] Hay, S. I., Noor, A. M., Nelson, A., *et al*. The accuracy of human population maps for public health application [J]. *Tropical Medicine & International Health*, 2005, 10(10): 1073–1086.
[67] Balk, D. L., Deichmann, U., Yetman, G., *et al*. Determining global population distribution: methods, appli-

cations and data [J]. *Advances in Parasitology*, 2006, 7(62): 119–156.
[68] Mondal, P., Tatem, A. J. Uncertainties in measuring populations potentially impacted by sea level rise and coastal flooding [J]. *PLoS One*, 2012, 7(10): 1–7.
[69] Smith, A., Bates, P. D., Wing, O., *et al*. New estimates of flood exposure in developing countries using high-resolution population data [J]. *Nature Communications*, 2019, 10(1): 1–7.
[70] 王彦. 基于随机森林模型的人口密度研究[D]. 石家庄: 河北师范大学, 2020.
[71] 胡焕庸. 中国人口的分布、区划和展望[J]. 地理学报, 1990, 45(2): 139–145.
[72] 胡焕庸. 中国八大区人口增长、经济发展的过去与未来[M]. 上海: 华东师范大学出版社, 1986.
[73] 胡焕庸. 中国东部、中部、西部三带的人口、经济和生态环境[M]. 上海: 华东师范大学出版社, 1989.
[74] 吴传钧. 论地理学的研究核心——人地关系地域系统[J]. 经济地理, 1991, 11(3): 1–6.
[75] 樊杰. 地理学的综合性与区域发展的集成研究[J]. 地理学报, 2004, 59(S1): 33–40.
[76] 樊杰. “人地关系地域系统”学术思想与经济地理学[J]. 经济地理, 2008, 28(2): 870–878.
[77] 樊杰. 中国人文地理学 70 年创新发展与学术特色[J]. 中国科学: 地球科学, 2019, 49(11): 1697–1719.
[78] Krugman, P. First nature, second nature, and metropolitan location [J]. *Journal of Regional Science*, 1993, 33(2): 129–144.
[79] 陆大道. 中国区域发展的新要素与新格局[J]. 地理研究, 2003, 22(3): 261–271.
[80] 夏海斌, 王铮. 中国大陆空间结构分异的进化[J]. 地理研究, 2012, 31(12): 2123–2138.
[81] 郑度, 欧阳, 周成虎. 对自然地理区划方法的认识与思考[J]. 地理学报, 2008, 6(3): 563–573.
[82] “黄秉维文集”编辑小组. 地理学综合研究: 黄秉维文集[M]. 北京: 商务印书馆, 2003.
[83] 刘昌明, 郑度, 陆大道等. 地理学研究的发展方向——地理学期刊主编笔谈[J]. 地理学报, 2005, 60(4): 531–545.
[84] 胡焕庸. 论中国人口之分布[M]. 上海: 华东师范大学出版社, 1983.
[85] 丁金宏, 刘振宇, 程丹明等. 中国人口迁移的区域差异与流场特征[J]. 地理学报, 2005, 60(1): 106–114.
[86] 葛美玲, 封志明. 基于GIS的中国2000年人口之分布格局研究——兼与胡焕庸1935年之研究对比[J]. 人口研究, 2008, 32(1): 51–57.
[87] 刘盛和, 邓羽, 胡章. 中国流动人口地域类型的划分方法及空间分布特征[J]. 地理学报, 2010, 65(10): 1187–1197.
[88] 戚伟, 刘盛和, 赵美风. “胡焕庸线”的稳定性及其两侧人口集疏模式差异[J]. 地理学报, 2015, 70(4): 551–566.
[89] 陆大道, 王铮, 封志明等. 关于“胡焕庸线能否突破”的学术争鸣[J]. 地理研究, 2016, 35(5): 805–824.
[90] 陈明星, 李扬, 龚颖华等. 胡焕庸线两侧的人口分布与城镇化格局趋势——尝试回答李克强总理之问[J]. 地理学报, 2016, 71(2): 179–193.
[91] 周立三. 中国农业区划的理论与实践[M]. 合肥: 中国科学技术大学出版社, 1993.
[92] 刘彦随, 张紫雯, 王介勇. 中国农业地域分异与现代农业区划方案[J]. 地理学报, 2018, 73(2): 203–218.
[93] 宋长青. 地理学研究范式的思考[J]. 地理科学进展, 2016, 35(1): 1–3.
[94] 宋长青, 程昌秀, 史培军. 新时代地理复杂性的内涵[J]. 地理学报, 2018, 73(7): 1204–1213.