人口密度随机森林模型优化实验研究

李玲玲[1],刘劲松[1,2,3,4],李 智[1,2,3,4],温佩璋[1],李艳成[1],刘 艺[1]
(1. 河北师范大学地理科学学院,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. 河北师范大学地理计算与规划研究中心,石家庄 050024;
4. 河北省环境演变与生态建设实验室,石家庄 050024)

摘要: 随机森林模型是精准刻画区域人口分布规律和影响机制的主流研究方法。本文以石家庄为实验区,以综合禀赋分区为建模单元,在公顷网格粒度上分区开展分层采样,系统进行了递增式人口密度影响因子遴选实验,全流程(分区建模、分层采样、因子遴选、加权输出)优化了人口密度随机森林模型。研究表明:① 分区建模抑制了模型混淆人口分布法则问题;在栅格粒度上采样,不仅使训练样本数据质量摆脱了MAUP的困扰,而且在形式上尝试降低区群谬误的不良影响;分层采样确保了样本数据集中人口密度标签值的分布稳定性。② 利用人口密度随机森林模型,分区开展人口密度影响因子遴选实验,逐步提升了模型的拟合优度R^2;距聚落距离是各区人口密度的主要影响因子;各区的人口分布影响机制存在显著差异,创新禀赋因子对城镇地区人口密度有较强影响,自然禀赋因子对乡村地区人口密度有较强影响。③ 对人口密度预测数据集进行优化组合,显著提高了模型的鲁棒性。④ 所获人口密度数据集具有多尺度叠加特征,大尺度上呈现出平原人口密度高于山区,小尺度上呈现出城镇人口密度高于乡村的"核心—边缘"特征。人口密度随机森林模型优化方案为揭示地方性人口分布规律和人口分布影响机制提供了统一的技术框架。

关键词: 人口密度;随机森林模型;禀赋分区;分层采样;因子遴选;加权输出
DOI: 10.11821/dlxb202305015

1 引言

人口密度是单位面积上的人口数量,是表征区域人口分布特征的定量指标[1]。高分辨率人口密度数据集是揭示人口分布规律的基础依据。为在栅格尺度整合人口、资源、环境数据集,推动全球变化的定量研究工作,20世纪90年代初,HDP(The Human Dimensions of Global Environmental Change Programme)第3工作组倡议研制全球人口密度栅格数据集[2],"自上而下的人口普查数据分解算法"[3](含面积加权[4-8]和线性回归[9-22]两类人口密度模型)率先得到了发展,GPW[7]、GHS-POP[8]、WPE[9]、HYDE[16]和LandScan[19]均是利用此类算法生产的全球人口密度栅格数据产品。2015年联合国可持续发展目标(Sustainable Development Goals, SDGs)认为,栅格人口密度模型的信度和效

收稿日期: 2022-11-28; **修订日期:** 2023-03-15
基金项目: 国家自然科学基金项目(42071167, 42201197, 40871073); 第二次青藏高原综合科学考察研究(2019QZKK0406); 河北省自然科学基金项目(D2007000272) **[Foundation:** National Natural Science Foundation of China, No.42071167, No.42201197, No.40871073; The Second Tibetan Plateau Scientific Expedition and Research Program, No.2019QZKK0406; Natural Science Foundation of Hebei Province, No.D2007000272]
作者简介: 李玲玲(1996-), 女, 河北石家庄人, 博士生, 研究方向为人口地理学。E-mail: lingling@stu.hebtu.edu.cn
通讯作者: 刘劲松(1970-), 男, 辽宁凌源人, 教授, 博导, 中国地理学会会员(S110004975M), 长期从事人口地理研究。E-mail: liujinsong@hebtu.edu.cn

度亟待改进[23]，与之相呼应，近年来随机森林模型在“自上而下的人口普查数据分解算法”[24-27]和“自下而上的人口调查数据估计算法”[28-29]中得到了广泛应用。

然而在构建人口密度随机森林模型时，下列问题并未得到妥善解决，制约了人口密度随机森林模型的信度和效度。① 训练样本的数据质量仍受可塑性面积单元问题（Modifiable Areal Unit Problem, MAUP）困扰[30-32]。人口密度属于定比量化指标，改变统计单元的形状或面积，人口密度值将发生变化。人口密度随机森林模型通常以人口普查区[24-27]或人口调查区[28-29]为单位开展采样，此时只能借助聚合运算才能获得建模所需的训练样本数据（含人口密度和影响因子），受MAUP困扰，样本数据质量存疑[33-34]。② 模型存在区群谬误问题（Ecological Fallacy）[35]。由于人口密度随机森林模型的输入单元多为人口普查区或人口调查区，输出单元多为公里网格或公顷网格，模型的输入单元粒度远大于输出单元粒度，故模型隐含区群谬误问题。③ 模型存在混淆人口分布规律问题。以中国为例，在地域辽阔的国土中，存在众多地理区划单元[36]，各区划单元的人口分布规律和影响机制存在显著差异[37-39]。仅用一套训练样本构建覆盖中国的人口密度随机森林模型[24, 27]，会混淆不同区域（例如平原和山区）的人口分布规律[40]。④ 忽视分区遴选人口密度影响因子。由于人口密度随机森林模型属于监督模型，故引入不同的影响因子，计算所得的人口密度数据集存在显著差异。“千篇一律”的样本模式，不仅存在引入错误影响因子的风险，而且会阻碍探讨“各美其美，美美与共”的人口分布法则和影响机制[34]。

为系统破解上述问题，本文提出了一套人口密度随机森林模型优化方案。发扬地理学中国学派的区划传统，将石家庄划分为平原城镇、平原乡村、山区城镇、山区乡村4个综合禀赋区，通过分区采样，分区建模，克服人口密度随机森林模型混淆人口分布规律的问题；以公顷网格为采样单元，开展分层采样，统一模型输入单元和输出单元的粒度，规避聚合运算，避免训练样本受到MAUP问题困扰，消除模型隐含的区群谬误问题；以模型的平均拟合优度为衡量标准，系统开展递增式影响因子遴选实验；通过对10组独立的人口密度预测数据集的优化组合，提高人口密度栅格数据集的稳定性。

2　材料与方法

2.1　研究区概况

石家庄市是河北省省会，位于37°27′N~38°47′N，113°30′E~115°30′E之间，地势西高东低（图1）。全市（含辛集市）下辖8个区、11个县，3个县级市，总面积14464 km^2。截至2020年11月1日，石家庄市常住人口为1123.51万人[41]。

2.2　数据来源

文中所用主要数据集详见表1。村人口数据集为2007年4月30日24时石家庄市户籍人口分村统计数据，村界、聚落数据集取自第二次全国土地调查数据集，依托上述3个数据集，利用二元加权模型，计算获得聚落人口密度数据集，是文中建模所需的人口密度标签数据集。

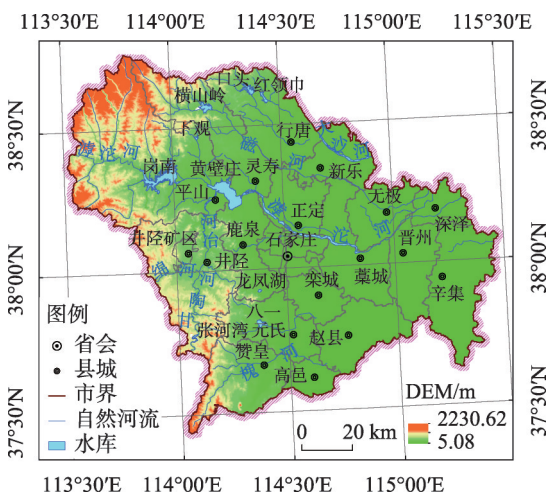

图1　研究区行政区划与地形

Fig. 1　Administrative divisions and terrain of the study area

表1 主要数据集

Tab. 1 The main datasets

目标数据集	原始数据集	数据来源	处理方法
聚落人口密度数据集	村人口数据集	石家庄市公安局	二元加权模型①
	村界数据集	第二次全国土地调查数据集	
	聚落数据集	第二次全国土地调查数据集	
自然禀赋因子数据集	DEM数据集	地理国情监测云平台1∶25万DEM数据	投影转换和重采样[42]
	地形起伏度数据集		Focalmean函数[43]
	坡度数据集		坡度函数[42]
	年均气温数据集	1971—2000年 河北省及周边气象台站气象监测数据	Kriging插值[6]
	年均降水数据集		Spline插值[44]
	距河流距离数据集	2015年1∶100万全国基础地理数据库	欧氏距离
	距自然河流距离数据集		欧氏距离
经济禀赋因子数据集	距POIs距离数据集	2012年百度中国POIs数据集	欧氏距离[34]
	距聚落距离数据集	聚落数据集	欧氏距离
创新禀赋因子数据集	POIs核密度数据集	2012年百度中国POIs数据集	核密度[34]
	聚落核密度数据集	聚落数据集	核密度
	夜光影像数据集	2007年DMSP/OLS夜光影像数据集	重采样(像元大小为100 m)
分区训练样本	分区训练样本数据集	通过分层采样,每区获得10套训练样本数据集	分层采样

注：① 处理方法参考未公开发表中文期刊文献:李艳成, 温佩璋, 刘劲松. 基于聚落的人口统计数据空间分解算法。

文中从自然禀赋、经济禀赋和创新禀赋3个维度选取人口密度的候选影响因子。其中，自然禀赋因子包括海拔高度、地形起伏度、坡度、年均温、年均降水量、距河流距离（包括自然和人工河流）、距自然河流距离；经济禀赋因子包括距POIs距离、距聚落距离；创新禀赋因子包括POIs核密度、聚落核密度、夜光影像。上述数据集全部采用Albers伪圆锥等积投影。各栅格数据集均为GeoTiff格式，栅格尺寸统一为100 m×100 m，统一了各数据集的四至点坐标。

2.3 研究方法

人口密度随机森林模型的优化方案如图2所示，包括综合禀赋分区和影响因子分区遴选实验2大关键逻辑环节。综合禀赋分区强调将中国地理学的区划传统嫁接到人口密度随机森林模型的建模之中，目的在于分区建模，分区揭示人口分布规律，分区认识人口分布影响机制，以实现“各美其美，美美与共”的人口密度建模目标。人口密度影响因子分区遴选实验包括分层采样、人口密度预测数据集的优化组合、分区人口制图、模型效度检验等关键技术环节。

2.3.1 综合禀赋分区 以100 m等高线为界，保持村界的完整性，将研究区分为平原和山区；将含有201城和202镇两类聚落的村和居委会设定为城镇，将只含有203村聚落的村或居委会设定为乡村；将地貌分区数据集与城乡分区数据集叠加，把研究区划分为平原城镇、平原乡村、山区城镇、山区乡村4类禀赋区，生成分区计算掩膜（图3）。综合禀赋分区划定了揭示人口分布规律和影响机制的建模单元，规定了人口密度随机森林模型的建模尺度。基于行政区划数据集（县、乡）和研究区计算掩膜，得到县级分区密度制图掩膜和乡级人口密度准则效度检验掩膜。

2.3.2 分层采样 借助分层采样，构造训练样本。分层采样是依据禀赋区聚落人口密度数据集的均值和标准差，按照公式（1），将禀赋区分为若干层。

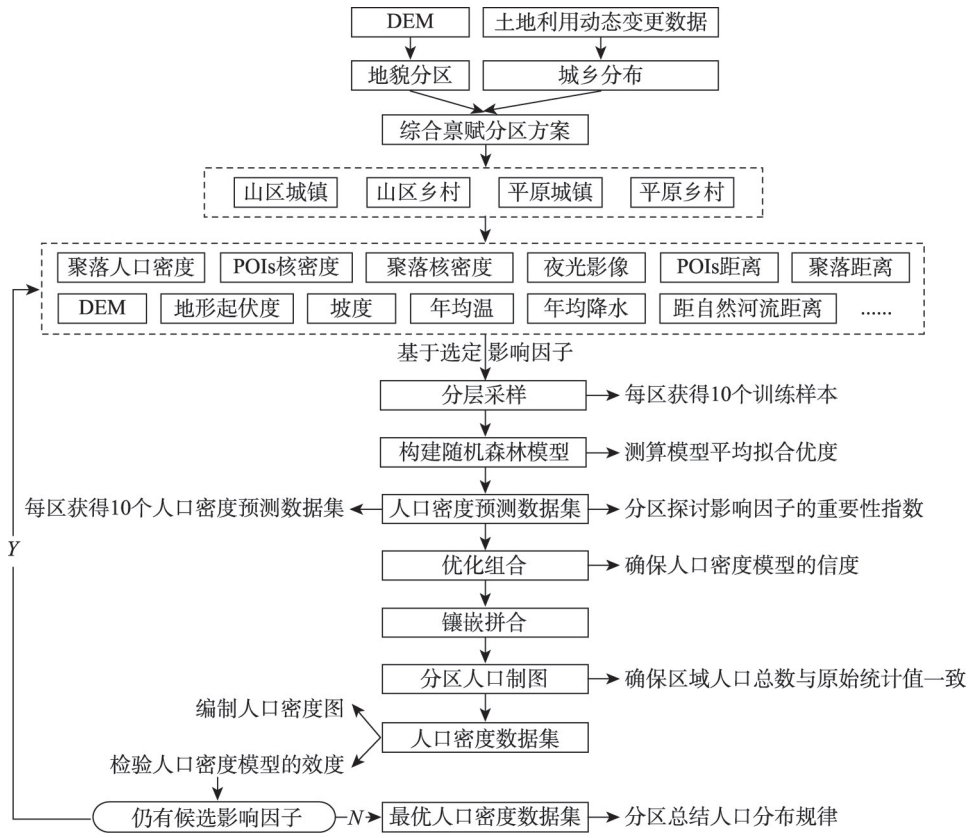

图2　人口密度随机森林模型优化流程

Fig. 2　Flow chart of the optimization of population density using a random forest model

$$Z_{grid}^{i}=j\begin{cases}\rho_{grid}^{i}\in\left[0,Z_{ave}^{i}\right), & j=1\\ \rho_{grid}^{i}\in\left[Z_{ave}^{i}+(j-2)Z_{std}^{i},\ Z_{ave}^{i}+(j-1)Z_{std}^{i}\right), & j\in[2,n]\end{cases}\tag{1}$$

式中：i是禀赋区编号，$i=1$代表山区乡村，$i=2$代表山区城镇，$i=3$代表平原乡村，$i=4$代表平原城镇；Z_{grid}^{i}表示第i个禀赋区的第$grid$个栅格的所属层序号；j代表层序号；ρ_{grid}^{i}表示第i个禀赋区的第$grid$个栅格的人口密度值；Z_{ave}^{i}和Z_{std}^{i}分别代表第i个禀赋区聚落人口密度数据集的平均值和标准差。

根据训练样本的采样规模和层面积占比，确定层采样规模。若第i个禀赋区第j层的面积为S_j^i，则第i个禀赋区第j层的采样规模NUM_j^i计算公式为：

$$NUM_j^i=\frac{S_j^i}{S^i}\times T^i,\qquad j=1,2,\cdots,n\tag{2}$$

式中：S^i为第i个禀赋区的总面积；S_j^i为第i个禀赋区第j层总面积；T^i为第i个禀赋区的采样规模（600）。若n满足（3）式条件，则表示已完成对第i个禀赋区的分层采样工作。

$$\sum_{j=1}^{n}NUM_j^i=T^i\tag{3}$$

各禀赋区随机进行10次分层采样，获取10组相互独立的训练样本。

2.3.3 拟合优度 R^2 拟合优度 R^2 是检验模型效度的标准化测度指标，其计算公式如下：

$$R^2 = 1 - \frac{\sum_{i=1}^{m}\left(\rho_i - \hat{\rho}_i\right)^2}{\sum_{i=1}^{m}\left(\rho_i - \bar{\rho}\right)^2} \tag{4}$$

式中：ρ_i 是人口密度真实值；$\hat{\rho}_i$ 是人口密度预测值；m 代表检验样本容量；R^2 越大，表示拟合优度越好；$\bar{\rho}$ 是人口密度预测的平均值。

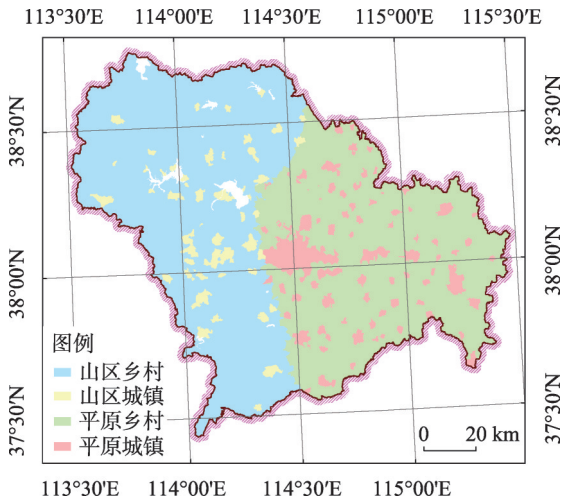

图3 分区计算掩膜

Fig. 3 Zonal computational mask

R^2 在文中有2个作用。一是人口密度随机森林预测模型的 R^2，其取值范围介于负无穷和1之间。随机森林模型利用袋外样本，测算人口密度预测模型的 R^2。利用10个人口密度随机森林预测模型的 R^2 的算术平均值，比较不同实验组别人口密度预测模型的拟合效果，是遴选人口密度影响因子的重要宏观量化指标。二是利用乡（或街道）真实人口统计数与人口密度数据集乡（或街道）的人口汇总数构建一元线性回归方程（公式（5）），并用 R^2 测度人口密度数据集的准则效度，其取值范围为0~1，R^2 越大，表示人口密度数据集的准则效度越高，并据此与国际著名人口密度模型进行准则效度对比。

$$\hat{\rho}^i = \beta^i \rho^i + \varepsilon^i \tag{5}$$

式中：i 是禀赋区编号，$i=1$ 代表山区乡村，$i=2$ 代表山区城镇，$i=3$ 代表平原乡村，$i=4$ 代表平原城镇，$i=0$ 代表研究区；$\hat{\rho}^i$ 代表第 i 区各乡人口预测总数；ρ^i 代表第 i 区各乡原始人口统计总数。

2.3.4 人口密度预测数据集的优化组合 由于人口密度随机森林模型高度依赖训练样本，若仅用一套分层采样获得的训练样本集构建人口密度随机森林模型，则无法保证人口密度预测数据集无偏。为确保人口密度预测数据集的再测信度，根据正态分布理论，优化人口密度预测模型的输出环节。首先通过分层采样，独立抽取10个训练样本集，对应构建10个人口密度随机森林模型，获得10个（$n=10$）相互独立的人口密度预测数据集。

分区计算10个人口密度预测模型的平均拟合优度值，并将其作为人口密度预测模型准则效度的评价指标和构造人口密度数据集优化组合系数的评判阈值，其计算公式如下：

$$AVG\left(R_i^2\right) = \frac{1}{10}\sum_{j=1}^{10} R_{i,j}^2 \tag{6}$$

式中：i 是禀赋区序号，$i \in \{1, 2, 3, 4\}$；$R_{i,j}^2$ 是第 i 个禀赋区第 j 个人口密度预测模型的拟合优度。

若从各区10个人口密度预测数据集中随机抽取6个人口密度预测数据集，则理论上每个禀赋区均存在210种人口密度预测数据集的组合方案。实验表明，各分区人口密度预测数据集的210种组合的拟合优度平均值符合正态分布。当某一组合中6个人口密度预测模型的拟合优度的平均值大于10个人口密度预测模型的平均拟合优度值时，则给该组合所含各人口密度预测数据集分别累计1次，否则不累计。按这个原则，将210种组合逐个评估一遍，则某个人口密度预测数据集的优化组合系数 W_i=该人口密度预测数据的有

效累计次数/10个人口密度预测数据的有效累计次数之和。通过加权平均，获得各禀赋区人口密度预测数据集，将各禀赋区人口密度预测数据集拼接在一起，则获得研究区人口密度预测数据集。

$$\rho = \sum_{i=1}^{10} W_i \times \rho_i \tag{7}$$

式中：ρ是通过加权平均获得的某禀赋区人口密度预测最终数据集；W_i是基于某禀赋区第i个训练样本获得的人口密度预测数据集的权重；ρ_i是基于某禀赋区第i个训练样本所获的人口密度预测数据集。

2.3.5 分区密度制图（Dasymetric Mapping） 分区密度制图是将人口密度预测数据集转换为人口密度数据集的国际通行计算方法，借助人口密度预测数据集获得每个栅格的分配权重，从而确保人口密度数据集中每个人口统计单元（县）的人口总数与原始人口统计汇总数据相等，分区密度制图公式如下[27]：

$$Pop_{grid} = Pop_{county} \times \frac{W_{grid}}{W_{county}} \tag{8}$$

式中：Pop_{grid}是人口密度数据集中栅格*grid*的人口数（即人口密度值）；Pop_{county}是人口密度数据集中栅格*grid*所属县*county*的原始人口统计汇总数；W_{grid}是人口密度预测数据集中栅格*grid*的人口密度预测值；W_{county}是人口密度预测数据集中栅格*grid*所属县*county*在人口密度预测数据集中的人口汇总数；W_{grid}/W_{county}实质就是*county*县人口统计汇总数在栅格*grid*的分配权重。

3 人口密度模型优化实验

3.1 遴选人口密度影响因子

3.1.1 遴选人口密度影响因子的实验方案 采用递增式遴选方法，系统开展人口密度影响因子的遴选实验（表2），遴选实验共开展8轮。其中，将增加1个影响因子或替换某个影响因子的行为称为实验刺激（Experimental Stimulus），将给予实验刺激之前的实验称为前测（Pretest）实验，将给予实验刺激之后的实验称为后测（Posttest）实验[35]。

例如，表2中01组实验为02组实验的前测实验，02组实验为01组实验的后测实验，02组实验较01组实验增加了距河流距离因子F_1，距河流距离因子F_1在此充当实验刺激。每组实验均采用分层抽样策略，每个禀赋区独立抽取10个训练样本数据集，构建10个随机森林模型，产生10个人口密度预测数据集。通过比较前测和后测所获10个人口密度预测数据集的平均拟合优度（表3）和人口密度预测优化数据集（图5），半定量半定性评价实验刺激对人口密度预测模型所产生的影响效果[23]，并决定是否保留新引入的影响因子。

依据演化经济地理学理论，农业文明时代人口分布受自然河流显著影响，山区人口呈现逐水而居的特征，平原人口则避水而居[24]，因此在01组实验（即前测实验1）中，将A（海拔高度）、B（地形起伏度）、C（坡度）、D（年均气温）、E（年均降水）作为人口密度随机森林模型的影响因子，模拟构建自然禀赋因子影响下的人口密度预测模型。随着农耕文明的不断发展，人类修建了许多人工河流（减河、运河、灌渠等），在02组实验（即后测实验1）和03实验（即后测实验2）中，分别增加了F_1（距河流距离，含自然河流和人工河流）、F_2（距自然河流距离），尝试回答人工河流是否对人口分布有显著影响。

表2 遴选人口密度影响因子的实验方案

Tab. 2 Experimental scheme for selecting the factors that influence population density

实验组别	前测控制组					后测实验组		
	引入部分自然禀赋因子					河流距离	创新禀赋	经济禀赋
01	A	B	C	D	E			
02	A	B	C	D	E	F_1		
03	A	B	C	D	E	F_2		
04	A	B	C	D	E	F_m	G_1	
05	A	B	C	D	E	F_m	G_2	
06	A	B	C	D	E	F_m	G_3	
07	A	B	C	D	E	F_m	G_n	H_1
08	A	B	C	D	E	F_m	G_n	H_2

注：A表示DEM；B表示地形起伏度；C表示坡度；D表示年均气温；E表示年均降水；F_1表示距河流距离；F_2表示距自然河流距离；F_m表示各区选中的河流因子（m=1或2）；G_1表示夜光影像；G_2表示POIs核密度；G_3表示聚落核密度；G_n为各区选中的创新禀赋因子（n=1或2或3）；H_1表示距POIs距离；H_2表示距聚落距离。

信息文明时代，要素集聚是人口再分布的重要驱动力，是创新禀赋的重要表现。夜光影像、POIs核密度和聚落核密度均能反映设施集聚特征，故将这3个数据集选做创新禀赋候选影响因子。设计04（后测实验3）、05（后测实验4）、06（后测实验5）等实验，分别增加G_1（夜光影像）、G_2（POIs核密度）、G_3（聚落核密度）等实验刺激，尝试回答聚落核密度是否能够替代POIs核密度？不同禀赋区选用哪个创新禀赋因子更好？等系列问题。

工业文明时代，城乡区位显著影响人口就业和生活，是人口再分布的重要驱动力。设计07（后测实验6）、08（后测实验7）两组实验，分别增加H_1（距POIs距离）、H_2（距聚落距离）等实验，尝试回答聚落距离和POIs距离哪个因子更合适的问题。

3.1.2 各区人口密度影响因子的遴选结果 随着逐步引入不同的禀赋因子，各禀赋区人口密度预测模型的平均拟合优度值不断提升（表3）。在平原城镇和山区城镇区，当引入创新禀赋因子时，人口密度预测模型的平均拟合优度值有小幅提升，说明城镇地区人口分布已受到创新禀赋因子的影响。当引入距聚落距离因子后，各禀赋区的人口密度预测模型的平均拟合优度值均得到显著提升，说明城乡区位是影响人口分布的关键因子。

对比前测实验与后测实验所获人口密度预测模型的平均拟合优度，形成人口密度预测模型影响因子的最终遴选方案（表4）。如果逐列看表3和表4，人口密度预测模型的平均拟合优度自左向右，逐步提升。3个创新禀赋因子可相互替代，山区城镇和平原城镇倾向选用POIs核密度，平原乡村倾向选用聚落核密度，山区乡村倾向选用夜光影像；山

表3 各禀赋区影响因子遴选实验人口密度预测模型的平均拟合优度对照表

Tab. 3 Average goodness of fit of the population density prediction model in the experiment to determine influencing factors in each endowment area

禀赋分区	自然禀赋因子	河流因子		创新禀赋因子			经济禀赋因子	
	前测因子	距河流距离	距自然河流距离	夜光影像	POIs核密度	聚落核密度	POIs距离	聚落距离
山区城镇	0.135	0.148	0.142	0.172	0.180	0.156	0.185	0.525
山区乡村	−0.093	−0.079	−0.084	−0.063	−0.068	−0.075	−0.068	0.533
平原城镇	0.164	0.178	0.220	0.236	0.263	0.240	0.265	0.515
平原乡村	−0.109	−0.089	−0.097	−0.092	−0.087	−0.080	−0.080	0.711

表4　各禀赋区人口密度影响因子的遴选结果

Tab. 4　Selected factors affecting population density in each endowment area

禀赋分区	DEM	地形起伏度	坡度	年均温	年降水	河流因子		创新禀赋因子			经济禀赋因子	
						距河流距离	距自然河流距离	夜光影像	POIs核密度	聚落核密度	聚落距离	POIs距离
山区城镇	√	√	√	√	√	√	–	◎	√	O	√	–
山区乡村	√	√	√	√	√	√	–	√	◎	O	√	–
平原城镇	√	√	√	√	√	–	√	O	√	◎	√	–
平原乡村	√	√	√	√	√	√	–	O	◎	√	√	–

注：–代表不选影响因子；√代表首选影响因子；◎代表二选影响因子；O代表三选影响因子。

区城镇、山区乡村和平原乡村倾向选用距河流距离（含人工河流），平原城镇倾向选用距自然河流距离；距聚落距离因子明显好于距POIs距离，各区均倾向选择距聚落距离。如果逐行看表4，各禀赋区遴选的人口密度影响因子存在细微差异，这反映出不同禀赋区的人口分布影响机制确实存在区域差异。

3.1.3 人口密度影响因子的空间异质性分析　以08组实验为例，利用人口密度随机森林预测模型的重要性指数（Importance Index），说明影响因子的空间异质性特征（图4）。山区城镇人口密度主要受聚落距离（0.418）、POIs核密度（0.103）、海拔高度（0.099）、地形起伏度（0.081）等因子的影响；平原城镇人口密度主要受聚落距离（0.297）、POIs核密度（0.230）等因子影响；山区乡村人口密度主要受聚落距离（0.551）、年均温（0.072）、年降水（0.069）等因子的影响；平原乡村人口密度主要受聚落距离（0.710）的影响。总的来看，人口分布与聚落之间有很强的空间依附关系，距聚落距离（经济禀赋因子）是影响人口分布的最重要因子。聚落核密度（创新禀赋因子）已成为山区城镇、平原城镇、平原乡村人口分布的第二重要影响因子。自然禀赋因子对现代人口分布影响较小，其中城镇地区气温、降水等对人口分布影响较小，但气温、降水对乡村地区

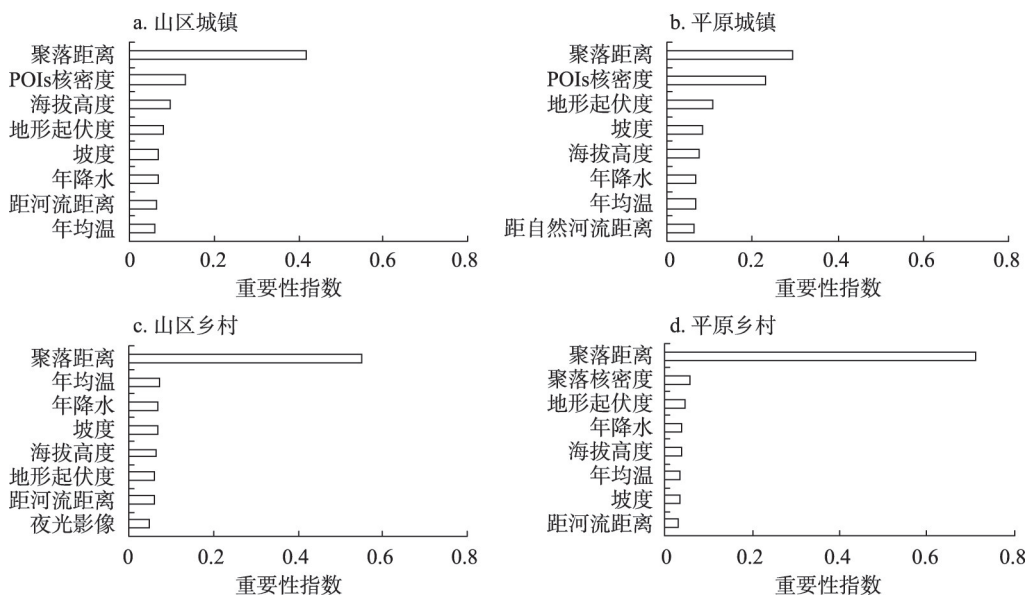

图4　分区人口密度影响因子的重要性指数(实验08)

Fig. 4　Importance index of the factors affecting population density in different areas (Experiment 08)

人口分布影响较大；由于聚落供水设施的不断完善，距河流距离或距自然河流距离对人口分布影响较小。

随着社会进步（从农业文明奔向工业文明或信息文明），在华北平原和太行山区过渡地带，自然禀赋因子作为一种外部因子，对人口分布的影响持续减弱。经济禀赋和创新禀赋因子已成为该区人口分布的主要影响因素。以石家庄为例，聚落距离（反映城乡区位）是影响人口密度分布的最重要因子；创新禀赋因子（聚落核密度）也逐步发挥了重要影响作用，尤其在城镇地区，创新禀赋因子是影响城镇地区人口分布的第二重要因子。只有在经济欠发达、自然禀赋条件较差的山区乡村，自然禀赋因子仍是影响人口分布的重要因素。

3.2 计算结果和模型检验

3.2.1 人口密度预测数据集的优化组合输出结果 利用2.3.4所述方法，获得每个人口密度预测数据集的优化组合系数；通过线性组合，获得分区人口密度预测数据集；通过镶嵌拼接，获得01~08组实验的石家庄人口密度预测数据集（图5）。从视觉效果看，01~07

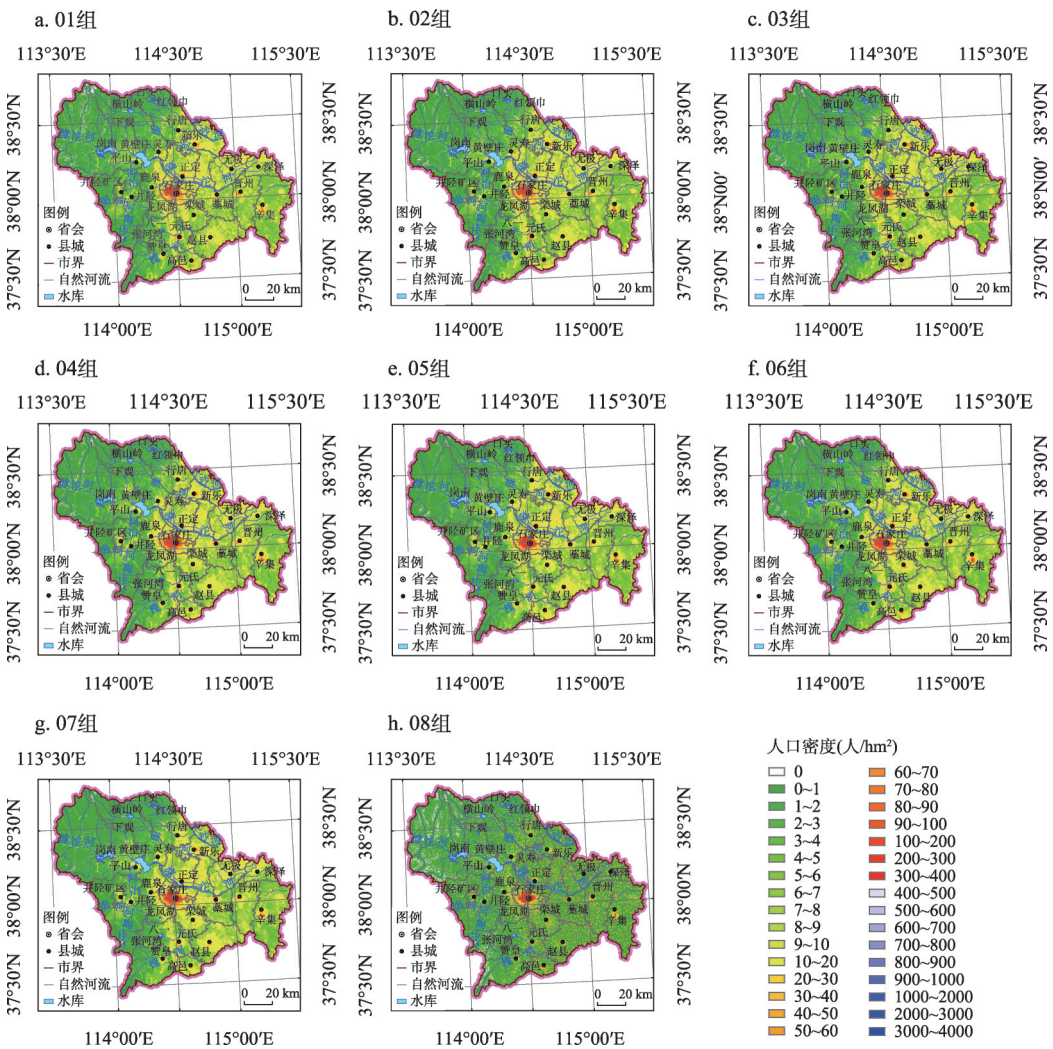

图5 01~08组实验所获石家庄人口密度预测数据集

Fig. 5 Datasets of the predicted Shijiazhuang population density for experiments from groups 01 to 08

组实验所获石家庄人口密度预测数据集没有显著差异，但08组实验所获石家庄人口密度预测数据集反映人口明显向聚落集聚。

表5列出了01组~08组实验所获石家庄人口密度预测数据集的关键统计特征，结果表明：① 当引入恰当的影响因子时，预测数据集的人口密度最大值会有所提升，说明所获人口密度预测数据集能更好刻画城乡人口密度差异。② 当引入恰当的影响因子时，预测数据集的人口密度平均值有所下降，人口密度预测数据集获得的研究区人口预测总数与真实人口总数越发接近，说明模型的预测精度得到了改善。③ 当引入恰当的影响因子时，预测数据集的人口密度标准差持续变大，说明人口密度预测数据集的空间异质性特征得到了持续改善。

表5　石家庄人口密度预测数据集关键统计指标对照（人/hm²）

Tab. 5　Key statistical indicators of the Shijiazhuang population density prediction datasets (people per hectare)

实验组别	最大值	平均值	标准差	实验组别	最大值	平均值	标准差
01	335.858	7.259	14.143	05	362.073	7.396	15.926
02	319.384	7.385	14.136	06	335.565	7.425	15.027
03	329.105	7.360	14.466	07	342.875	7.406	16.043
04	354.606	7.436	15.676	08	405.284	6.967	24.585

3.2.2 分区密度制图结果　利用2.3.5所述方法，以县为单位开展分区密度制图，获得了01~08组实验石家庄人口密度数据集（图6）。从08组实验所获的人口密度数据集看，石家庄人口分布极不均衡，总体趋势是中部高四周低，人口集聚的“核心—边缘”特征明显；山区聚落少、小、稀，人口密度低；平原聚落多、大、稠，人口密度高。山区人口密度“零值区”呈碎斑状连片分布；平原地区没有人口密度“零值区”。河流平原段周边人口密度低，聚落稀。滹沱河出山后，其北岸由于有设防水平高的河堤，聚落离河流干流距离较近，且密度较大；其南岸由于缺乏设防水平高的河堤，聚落离河流干流距离较远，密度较小。

表6是石家庄人口密度数据集（图6）的关键统计指标。与表5对应实验组别的人口密度预测数据集相比，可以看出分区密度制图的若干功效：① 使绝大多数人口密度数据集的人口密度最大值得到了进一步提高，其中，08实验所获人口密度数据集的最大值提升尤为显著。② 使每组实验所获人口密度数据集的平均值相等。③ 使大部分人口密度数据集的标准差有小幅提升，但07组和08组有小幅回落。

3.2.3 人口密度数据集的准则效度检验　基于石家庄人口密度数据集汇总获得各乡（镇）人口预测总数，令其为Y，令各乡（镇）人口登记总数为X，建立一元线性回归方程，计算回归方程的拟合优度R^2，作为人口密度数据集的准则效度（表7）。从01组实验到08组实验，回归方程的拟合优度R^2越来越大，说明人口密度随机森林模型的准则效度得到了显著提升。其中，08组实验所获人口密度数据集精度最高。

4　讨论

4.1　灵活制定分区策略

实验表明，分区建立人口密度随机森林模型有利于探讨人口分布规律和影响机制，但这并不意味只有分区才能构建人口密度随机森林模型。一般来说，对一个幅员辽阔的国家或地区，通常要采用分区建模策略；而对一个面积较小或景观单一的国家或地区，往往不须采用分区建模策略，只需构建一个人口密度随机森林模型。

综合禀赋分区一般要兼顾自然禀赋分区和城乡划分。当研究区面积很大时，通常采用拿来主义策略，直接将综合自然区划、农业区划、生态环境区划、地貌区划等研究成

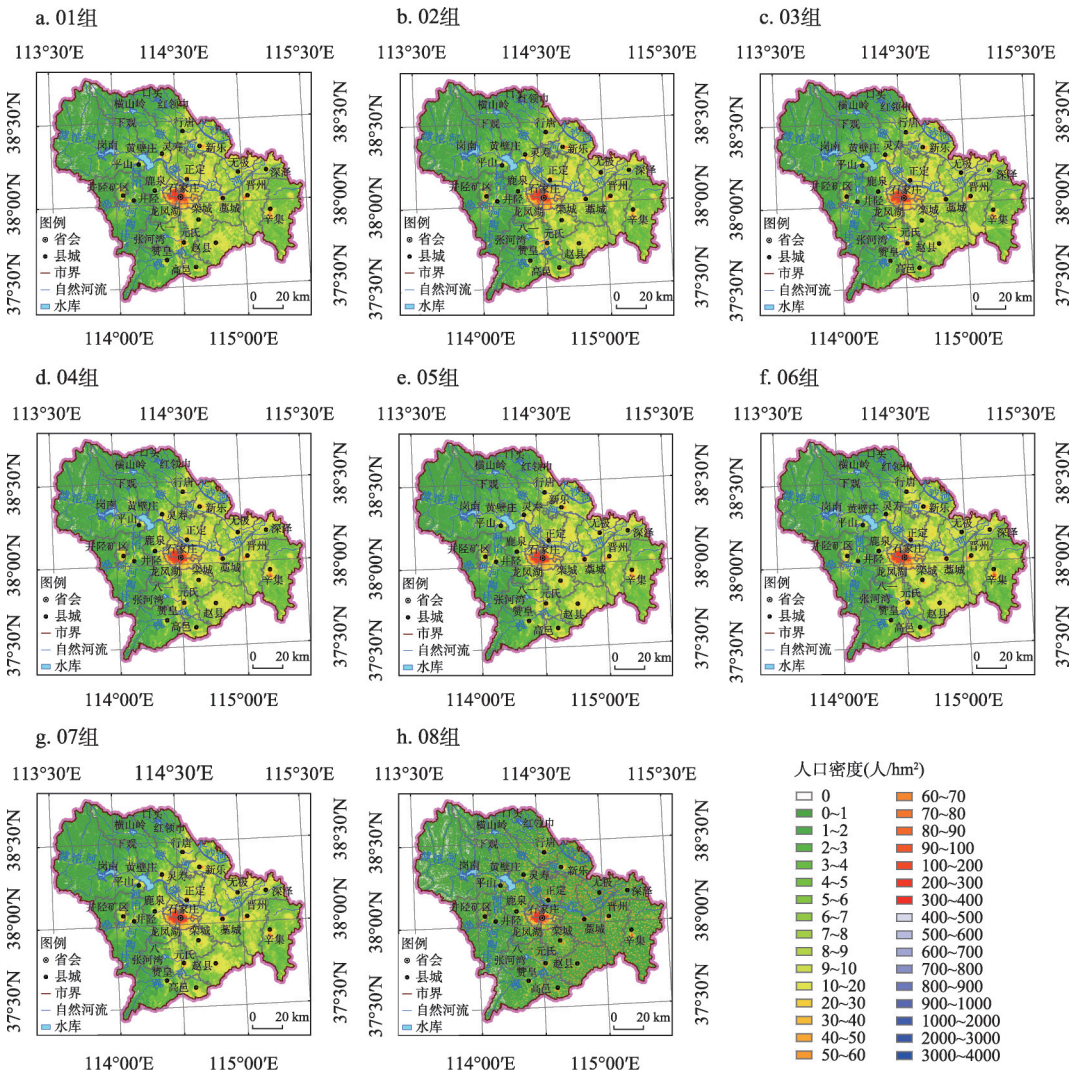

图6 01~08组实验石家庄人口密度数据集

Fig. 6 Shijiazhuang population density datasets for the experiments from groups 01 to 08

表6 石家庄市人口密度数据集关键统计指标对照表(人/hm²)

Tab. 6 Key statistical indicators of the Shijiazhuang population density datasets (people per hectare, pph)

实验组别	最大值	平均值	标准差	实验组别	最大值	平均值	标准差
01	420.510	6.775	15.137	05	373.971	6.775	16.278
02	453.410	6.775	15.050	06	397.708	6.775	15.522
03	408.598	6.775	15.660	07	355.833	6.775	15.890
04	350.737	6.775	15.484	08	512.187	6.775	24.492

果作为自然禀赋分区成果。当研究区面积较小时，在不打破人口普查或调查单元的情况下，即可采用主成分分析法和聚类分析法，开展综合禀赋分区；也可采用主导因素法（例如：以海拔高度为标准，将黄淮海平原分别划分为洪积平原、冲积平原、海积平原），获得自然禀赋分区。

由于城镇化是一个持续演变过程，因此城乡范围是动态演变的，城乡划分是禀赋分

区中的难点问题。对于土地利用数据变更及时的地区，可参考第二次全国土地调查或第三次全国国土调查成果数据划定城镇分布范围；对于缺乏土地利用数据的地区，可依据夜光影像，采用阈值法，确定不同时期的城乡分布范围。例如：在利用第三次至第七次人口普查数据编制青藏高原人口密度图时，就可利用夜光影像数据，界定不同时点青藏高原的城乡分布范围。

表7　乡(镇)人口预测数与人口统计数的拟合优度R^2

Tab. 7　Goodness of fit (R^2) of township (town) population forecasts and demographics

实验组别	01	02	03	04	05	06	07	08
R^2	0.890	0.882	0.919	0.936	0.934	0.923	0.941	0.967

分区建模实质上有两点好处：一是明确了模型的表达对象（禀赋区域），从而避免了混淆人口分布法则的难题；二是明确了模型的尺度（禀赋分区的范围）。分区建模为因地制宜、因时制宜揭示世界人口分布规律和影响机制，准确把握世界人口分布的过去、现在和未来，提供了科研协作框架和统一的技术框架。

4.2 进一步加强训练样本数据集的遴选工作

与随机采样相比，分层采样显著改善了训练样本数据集中人口密度标签值的分布稳定性。但针对每个禀赋区生成的10个训练样本数据集是否具有代表性？文中并未细致追究，其实这是有瑕疵的。抽样程序一般包括界定总体、制定抽样框、决定抽样方案、实际抽取样本、评估样本质量等5个环节。评估样本质量是对样本的质量、代表性、偏差进行初步检验和衡量，其目的是防止由于样本偏差过大而导致失误。通常采用比较法来评估样本质量，即将反映总体中某些重要特征及其分布的资料与样本中同类指标的资料进行比较。后续研究应利用比较法，尝试评估训练样本质量，力争提出满足分层采样策略的训练样本数据集的形式化筛选模型，为进一步开展人口密度随机森林的信度评价创造条件。

4.3 人口密度随机森林模型可能不必引入交通区位因子

相关研究表明，人口密度与交通通达度（表征交通区位的量化指标）具有显著相关性[46-47]，因此，在构建人口密度模型时，交通区位因子是常被引入的影响因子[23, 25-27, 48-51]。本文将交通通达度作为实验刺激，设计了第09组实验（后测实验8）。结果表明，在引入交通通达度后，人口密度预测数据集的最大值、人口密度预测模型的平均拟合优度R^2、人口密度数据集的标准差均出现了小幅下降的现象（即出现了影响因子边际效应），说明不宜将交通通达度引进到石家庄市人口密度随机森林模型之中。其实，1978年改革开放以来中国交通基础设施变化巨大，准确获取历史时期交通要素数据集的难度极大。如果在构建人口密度随机森林模型时不必引入交通区位因子，将显著降低各普查年份人口密度随机森林模型的构建难度。

4.4 聚落数据集在构建人口密度随机森林模型中发挥了极其重要的作用

聚落是人口生产、生活的集聚地，聚落分布与人口分布有着紧密联系。聚落数据集在创建人口密度标签数据集（聚落人口密度数据集）、经济禀赋因子（距聚落距离）、创新禀赋因子（聚落核密度）等方面都是不可或缺的核心数据集。

利用面积加权法，以矢量格式的村人口空间数据集为起点，以矢量格式的聚落和公顷网格数据集为约束，编制聚落人口密度栅格数据集，其人口密度的最大值高于GPWv4，其人口分布基尼系数大于GHS-POP和GPW，在人口分布的城乡差异和空间集聚特征方面有优异表现。聚落人口密度数据集是创建人口密度随机森林模型理想的人口标签数据集。

学者常使用距POIs距离来表征城乡区位特征，常使用POIs核密度或夜光影像来表征

要素集聚特征。但无论是POI数据，还是夜光影像数据都存在一定的局限性。以百度POI为例，POI包括生活、生产、公共服务3大类、24小类，其中，许多小类POI数据与人口分布无关。如果囫囵吞枣、不加区分，盲目引用POI数据，往往会导致人口密度随机森林模型出现偏差。另外，POI数据是大数据时代的产物，2010年以前，实际上还没有POI数据，若想构建1982年、1990年、2000年的人口密度随机森林模型，则面临无法找到POI数据的尴尬境地。夜光影像不仅自身存在“灯撒效应”（Blooming Effect），而且夜光影像的历史数据集也仅能回溯到20世纪90年代后期。

由于全球对地观测事业的不断发展，聚落数据（土地利用数据产品）具有较好的历史可回溯性。例如采用全球10 m粒度的聚落数据集和以乡为单元的人口普查数据集，项目组已经成功编制了2020年青藏高原聚落人口密度数据集，为构建青藏高原人口密度随机森林模型，探讨青藏高原人口分布规律和影响机制，奠定了扎实基础。可以预见，在构建全球不同区域的高分辨率人口密度随机森林模型时，聚落数据集已不是约束条件。考虑数据可得性，如果用距聚落距离代替距POIs距离，用聚落核密度代替POIs核密度或夜光影像，则会在保证模型效度的前提下，降低构建历史时期人口密度随机森林模型的难度。

4.5 部分实验模型与国际著名人口密度模型准则效度的比较

以乡真实人口统计数为自变量，以著名人口密度栅格数据集的乡人口汇总数为因变量，利用公式5构建一元线性回归方程；利用公式4测算不同模型在不同禀赋区的准则效度R^2，计算结果详见表8。考察研究区不同模型的准则效度，08组模型在众多随机森林模型产品中表现最好，随机森林模型（08组、WorldPOP、01组、ChinaPOP、POIPOP）整体优于线性回归模型（LandScan、GRUMP、CnPOP），线性回归模型整体优于类型赋权模型（HYDE）；二元加权和面积加权混合模型（GHS-POP）优于面积加权模型（GPW）。

表8 部分实验模型与国际著名人口密度模型准则效度（R^2）的对比

Tab. 8 Comparison of the validity (R^2) between selected experimental models and internationally famous population density models

模型	①	②	③	④	⑤	⑥	⑦	⑧	⑨	01组	08组
平原城镇	0.796	0.967	0.851	0.802	0.974	0.933	0.860	0.964	0.893	0.966	0.983
平原乡村	0.363	0.773	0.417	0.681	0.687	0.658	0.604	0.758	0.726	0.812	0.939
山区城镇	0.522	0.811	0.638	0.851	0.694	0.731	0.555	0.770	0.897	0.694	0.896
山区乡村	0.278	0.781	0.356	0.639	0.631	0.561	0.600	0.734	0.717	0.682	0.889
研究区	0.896	0.911	0.657	0.708	0.528	0.874	0.706	0.910	0.836	0.890	0.967

注：①~⑨分别为GPW、GHS-POP、HYDE、CnPOP、GRUMP、LandScan、ChinaPOP、WorldPOP、POIPOP等人口密度栅格数据集，其中①、②、③、⑤、⑥为2000年数据集，④、⑦、⑧、⑨为2010年数据集。

考察禀赋区不同模型的准则效度，08组模型表现最优，尤其是在农村地区，对面积加权模型、类型赋值模型和线性回归模型均产生了碾压优势。值得注意的是，08组模型的准则效度好于二元加权和面积加权混合模型（GHS-POP），这说明采用基于聚落的人口统计数据空间分解算法（克服了局部匀质化问题）所获聚落人口密度数据集（即建模所用人口密度标签数据集）的质量要显著优于GHS-POP人口密度数据集，折射出GHS-POP所用二元加权和面积加权混合模型的局部匀质化问题对乡村和山区的人口密度值仍有不良影响。

5 结论

文中提出了人口密度随机森林模型的全流程优化方案，实验结果表明：分区建模、分层采样、因子遴选、加权输出、分区密度制图是优化方案中5个关键的改进环节。

① 按禀赋区构建人口密度随机森林预测模型，能克服人口密度预测数据集中混淆人口分布法则的错误现象；将分层采样单元定为公顷网格，能使训练样本数据集免受MAUP问题的困扰，在形式上尝试降低区群谬误问题对模型的不良影响；分层采样显著提高了训练样本数据集中人口密度分布的稳定性。② 开展递增式人口密度影响因子遴选实验，通过分区优选人口密度随机森林的影响因子，能显著改善人口密度预测数据集的拟合优度。实验证实，不同禀赋区确实存在不同的人口分布影响机制，即人口分布的影响机制存在空间差异。③ 利用本文提出的人口密度预测数据集的优化组合方法，使得组合输出的人口密度预测数据集的“零值区”显著减少，显著改善了人口密度随机森林模型的稳定性。④ 分区密度制图能够进一步提高人口密度数据集的最大值，人口密度数据集的区分度与人口密度预测数据集的区分度大致相当。⑤ 现代人口分布不仅受自然禀赋因子影响，而且更受经济禀赋和创新禀赋因子影响。在石家庄，聚落距离（经济禀赋因子）是人口分布的最主要影响因子；聚落核密度等（创新禀赋因子）是城镇地区人口分布的第二位影响因子；气候、降水等自然禀赋因子对人口分布的影响减弱，但在乡村地区仍有较大影响。⑥ 石家庄人口分布具有非均衡特征，其总体趋势是中部高四周低，人口集聚的“核心—边缘”特征明显；平原地区人口密度高于山区地区；山区聚落少、小、稀，人口密度低；平原聚落多、大、稠，人口密度高。山区人口密度“零值区”呈碎斑状连片分布。

人口密度随机森林模型优化方案为揭示地方性人口分布规律和影响机制提供了统一的技术框架，为完整认识全球人口分布时空演变规律和影响机制提供了中国方案。

参考文献(References)

[1] Zhang Shanyu. Introduction to Population Geography. Shanghai: East China Normal University Press, 2013: 183-186. [张善余. 人口地理学概论. 上海: 华东师范大学出版社, 2013: 183-186.]

[2] Clarke J I, Rhind D W, Becket C, et al. Population data and global environmental change. Paris: The International Social Science Council, 1992.

[3] 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. PNAS, 2018, 115(14): 3529-3537.

[4] Zhang Congxuan. Using latitude and longitude grid cells to compile a population density map: Taking the Beijing-Tianjin-Tangshan area as an example. Areal Research and Development, 1985, 4(2): 57-66. [张丛宣. 用经纬网格单元编制人口密度图: 以京津唐地区为例. 中原地理研究, 1985, 4(2): 57-66.]

[5] Tobler W, Deichmann U, Gottsegen J, et al. World population in a grid of spherical quadrilaterals. International Journal of Population Geography, 1997, 3(3): 203-225.

[6] Liu Jinsong. The geographical meaning about the modifiable areal unit problem in the population density scaling [D]. Shijiazhuang: Hebei Normal University, 2009. [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.]

[7] Doxsey-Whitfield E, MacManus K, Adamo S B, et al. Taking advantage of the improved availability of census data: A first look at the gridded population of the world, Version 4. Applied Geography, 2015, 1(3): 226-234.

[8] Freire S, Macmanus K, Pesaresi M, et al. Development of new open and free multi-temporal global population grids at 250 m resolution. The 19th AGILE Conference on Geographic Information Science, Helsink: Springer Cham, 2016: 14-16.

[9] Frye C, Nordstrand E, Wright D J, et al. Using classified and unclassified land cover data to estimate the footprint of human settlement. Data Science Journal, 2018, 17(20): 1-12.

[10] Dobson J E, Bright E A, Coleman P R, et al. LandScan: A global population database for estimating populations at risk. Photogrammetric Engineering and Remote Sensing, 2000, 66(7): 849-857.

[11] Lo C P. Modeling the population of China using DMSP operational linescan system nighttime data. Photogrammetric Engineering and Remote Sensing, 2001, 67(9): 1037-1047.

[12] Jiang Dong, Yang Xiaohuan, Wang Naibin. Study on spatial distribution of population based on remote sensing and GIS. Advance in Earth Science, 2002, 17(5): 734-738. [江东, 杨小唤, 王乃斌, 等. 基于RS、GIS的人口空间分布研究. 地球科学进展, 2002, 17(5): 734-738.]

[13] Goldewijk K K, Ramankutty N. Land cover change over the last three centuries due to human activities: The availability of new global datasets. GeoJournal, 2004, 61: 335-344.

[14] Tian Yongzhong, Chen Shupeng, Yue Tianxiang, et al. Simulation of Chinese population density based on land-use. Acta Geographica Sinica, 2004, 59(2): 283-292. [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. 地理学报, 2004, 59(2): 283-292.]

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

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

[17] Zhuo Li, Chen Jin, Shi Peijun, et al. Modeling population density of China in 1998 based on DMSP/OLS nighttime light image. Acta Geographica Sinica, 2005, 60(2): 266-276. [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. 地理学报, 2005, 60(2): 266-276.]

[18] Amaral S, Monteiro A M V, Camara G, et al. DMSP/OLS nighttime light imagery for urban population estimates in the Brazilian Amazon. International Journal of Remote Sensing, 2006, 27(5): 855-870.

[19] Bhaduri B, Bright E, Coleman P, et al. LandScan USA: A high-resolution geospatial and temporal modeling approach for population distribution and dynamics. GeoJournal, 2007, 69: 103-117.

[20] Briggs D J, Gulliver J, Fecht D, et al. Dasymetric modelling of small area population distribution using land cover and light emissions data. Remote Sensing of Environment, 2007, 108(4): 451-466.

[21] Zeng C Q, Zhou Y, Wang S X, et al. Population spatialization in China based on night-time imagery and land use data. International Journal of Remote Sensing, 2011, 32(24): 9599-9620.

[22] Gao Yi, Wang Hui, Wang Peitao, et al. Population spatial processing for Chinese coastal zones based on census and multiple night light data. Resources Science, 2013, 35(12): 2517-2523. [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. 资源科学, 2013, 35(12): 2517-2523.]

[23] Tatem A J. WorldPop, open data for spatial demography. Scientific Data, 2017, 4: 170004. DOI: 10.1038/sdata.2017.4.

[24] Gaughan A E, Stevens F R, Huang Z J, et al. Spatiotemporal patterns of population in mainland China, 1990 to 2010. Scientific Data, 2016, 3: 160005. DOI: 10.1038/sdata.2016.5.

[25] Tan Min, Liu Kai, Liu Lin, et al. Spatialization of population in the Pearl River Delta in 30 m grids using random forest model. Progress in Geography, 2017, 36(10): 1304-1312. [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. 地理科学进展, 2017, 36(10): 1304-1312.]

[26] Wang Chao, Kan Aike, Zeng Yelong, et al. Population distribution pattern and influencing factors in Tibet based on random forest model. Acta Geographica Sinica, 2019, 74(4): 664-680. [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. 地理学报, 2019, 74(4): 664-680.]

[27] 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. Science of the Total Environment, 2019, 658: 936-946.

[28] 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. Earth System Science Data, 2019, 11(3): 1385-1409.

[29] Hillson R, Alejandre J D, Jacobsen K H, et al. Methods for determining the uncertainty of population estimates derived from satellite imagery and limited survey data: A case study of Bo City, Sierra Leone. Plos One, 2014, 9(11): e112241. DOI: 10.1371/journal.pone.0112241.

[30] Openshaw S. The Modifiable Areal Unit Problem. Norwich: Geobooks, 1983.

[31] Yang Xiaohuan, Jiang Dong, Wang Naibin. Method of pixelizing population data. Acta Geographica Sinica, 2002, 57 (Suppl.): 70-75. [杨小唤, 江东, 王乃斌. 人口数据空间化的处理方法. 地理学报, 2002, 57(增刊): 70-75.]

[32] Wu Jianguo. Landscape Ecology: Pattern Process Scale and Hierarchy. 2nd ed. Beijing: Higher Education Press, 2007: 147-154. [邬建国. 景观生态学: 格局、过程、尺度与等级. 2版. 北京: 高等教育出版社, 2007: 147-154.]

[33] Liu Yi, Yang Xinjia, Liu Jinsong. Experimental study on optimization of population density models based on random forest. Global Change Research Data Publishing & Repository, 2020, 4(4): 402-416. [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. 全球变化数据学报(中英文), 2020, 4(4): 402-416.]

[34] Liu Yi. Experimental study on optimization of population density based on random forest model [D]. Shijiazhuang: Hebei Normal University, 2022. [刘艺. 基于随机森林模型的人口密度优化实验研究[D]. 石家庄: 河北师范大学, 2022.]

[35] Feng Xiaotian. Social Research Methods. 5th ed. Beijing: China Renmin University Press, 2018: 75-78. [风笑天, 社会研究方法. 5版. 北京: 中国人民大学出版社, 2018: 75-78.]

[36] Zheng Du, Ou Yang, Zhou Chenghu. Understanding of and thinking over geographical regionalization methodology. Acta Geographica Sinica, 2008, 63(6): 563-573. [郑度, 欧阳, 周成虎. 对自然地理区划方法的认识与思考. 地理学报, 2008, 63(6): 563-573.]

[37] Hu Huanyong. The Past and Future of Population Growth, Economic Development of China's Eight Regions. Shanghai: East China Normal University Press, 1986: 9-14. [胡焕庸. 中国八大区人口增长、经济发展的过去和未来. 上海: 华东师范大学出版社, 1986: 9-14.]

[38] Hu Huanyong. Population, Economy and Ecologic Environment of East China, Middle China and West China. Shanghai: East China Normal University Press, 1989: 62-66. [胡焕庸. 中国东部、中部、西部三带的人口、经济和生态环境. 上海: 华东师范大学出版社, 1989: 62-66.]

[39] Hu Huanyong. The distribution, regionalization and prospect of China's population. Acta Geographica Sinica, 1990, 45 (2): 139-145. [胡焕庸. 中国人口的分布、区划和展望. 地理学报, 1990, 45(2): 139-145.]

[40] Wang Yan. Study on population density based on random forest model [D]. Shijiazhuang: Hebei Normal University, 2020. [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.]

[41] Shijiazhuang Municipal Bureau of Statistics. Bulletin of the Seventh National Population Census of Shijiazhuang (No. 1). www.sjz.gov.cn/col/1596018184396/2021/05/31/1622426640444.html, 2021-05-31/2022-10-18. [石家庄市统计局. 石家庄市第七次全国人口普查公报(第一号). www.sjz.gov.cn/col/ 1596018184396/2021/05/31/1622426640444.html, 2021-05-31/2022-10-18.]

[42] Hebei Population and Family Planning Commission. Research Report on the Functional Area of Population development in Hebei province. Shijiazhuang: Hebei People's Publishing House, 2009. [河北省人口和计划生育委员会. 河北省人口发展功能区研究报告. 石家庄: 河北人民出版社, 2009.]

[43] Zhang Lei. A study of the geomorphologic forms classification based on relief: Take Beijing-Tianjin-Hebei region for example [D]. Shijiazhuang: Hebei Normal University, 2009. [张磊. 基于地形起伏度的地貌形态划分研究: 以京津冀地区为例[D]. 石家庄: 河北师范大学, 2009.]

[44] Liu Jinsong, Chen Hui, Yang Binyun, et al. Comparison of interpolation methods for annual precipitation in Hebei province. Acta Ecologica Sinica, 2009, 29(7): 3493-3500. [刘劲松, 陈辉, 杨彬云, 等. 河北省年均降水量插值方法比较. 生态学报, 2009, 29(7): 3493-3500.]

[45] Kim J. Estimation of optimality gap using stratified sampling. Applied Mathematics and Computation, 2005, 171: 710-720.

[46] Wang Zheng, Xia Haibin, Tian Yuan, et al. Big data analysis on the existence of Hu Huanyong Line: Ecological and new economic geography understanding of China's population distribution characteristics. Acta Ecologica Sinica, 2019, 39 (14): 5166-5177. [王铮, 夏海斌, 田园, 等. 胡焕庸线存在性的大数据分析: 中国人口分布特征的生态学及新经济地理学认识. 生态学报, 2019, 39(14): 5166-5177.]

[47] Bai Ying, Wang Sen, Wu Sufeng, et al. Study on the relationship between population density and traffic intensity. China Transportation Review, 2021, 43(8): 21-25, 76. [白颖, 王森, 伍速锋, 等. 人口密度与交通强度关系研究. 综合运输, 2021, 43(8): 21-25, 76.]

[48] Qiu Y, Zhao X S, Fan D Q, et al. Disaggregating population data for assessing progress of SDGs: Methods and applications. International Journal of Digital Earth, 2022, 15(1): 2-29.

[49] Qiu G, Bao Y H, Yang X C, et al. Local population mapping using a random forest model based on remote and social sensing data: A case study in Zhengzhou, China. Remote Sensing, 2020, 12(10): 1618. DOI: 10.3390/rs12101618.

[50] Wang X Y, Meng X F, Long Y. Projecting 1 km-grid population distributions from 2020 to 2100 globally under shared socioeconomic pathways. Scientific Data, 2022, 9: 563. DOI: 10.1038/s41597-022-01675-x.

[51] 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. Plos One, 2015, 10(2): e0107042. DOI: 10.1371/journal.pone.0107042.

Experimental study of population density using an optimized random forest model

LI Lingling[1], LIU Jinsong[1, 2, 3, 4], LI Zhi[1, 2, 3, 4], WEN Peizhang[1], LI Yancheng[1], LIU Yi[1]

(1. School of Geographical Sciences, Hebei Normal University, Shijiazhuang 050024, China; 2. Hebei Technology Innovation Center for Remote Sensing Identification of Environmental Change, Shijiazhuang 050024, China; 3. GeoComputation and Planning Center of Hebei Normal University, Shijiazhuang 050024, China; 4. Hebei Key Laboratory of Environmental Change and Ecological Construction, Shijiazhuang 050024, China)

Abstract: Random forest model is a mainstream research method to accurately describe the regional population distribution law and impact mechanism. Taking Shijiazhuang as the experimental area and its endowment zones as the modeling unit, we carried out stratified sampling on a hectare grid scale, and conducted a systematic experiment to determine the factors influencing the increasing population density. An optimized random forest model was applied throughout the whole process of zoning modeling, stratified sampling, factor selection, to obtain weighted outputs. Four main conclusions can be drawn as follows: (1) Zoning before modeling prevented the model from confusing the population distribution laws. Sampling at the raster unit not only freed the training samples from the modifiable areal unit problem (MAUP), but also formally reduced the negative effect of the ecological fallacy. Stratified sampling ensured the stability of the maximum population density in the training samples. (2) The experiments to determine the factors influencing population density were conducted in different zones, and the introduction of these factors significantly improved the fit (R^2) of the model. Distance to a settlement was the dominant factor influencing population density in each zone. There were significant differences in the geographical mechanisms that influenced the population distribution in different regions. Innovation endowment factors had the strongest impact on population density in urban areas, while natural endowment factors had the strongest impact in rural areas. (3) The optimized combination of the population density prediction datasets significantly improved the robustness of the model. (4) The population density datasets had the characteristics of multi-scale superposition. At the large scale, the population density in the plain area was higher than that in the mountain area, whereas at the small scale the population density in urban areas was higher than that in rural areas, which represented the characteristics of a core- periphery model. The optimized scheme of the population density random forest model provided a unified technical framework for determining the factors that control the local population distribution and the geographical mechanisms that influence population distribution.

Keywords: population density; random forest model; endowment zones; stratified sampling; factor selection; weighted output