1776~1953年中国传统农区人口的格网化重建

张学珍^{1,5}, 王发浩^{1,2}, 路伟东³, 李士成⁴, 郑景云^{1,5*}

1. 中国科学院地理科学与资源研究所, 中国科学院陆地表层格局与模拟重点实验室, 北京 100101;

2. 北京师范大学地理科学学部, 北京 100875;

3. 复旦大学历史地理研究中心, 上海 200433;

4. 中国地质大学(武汉)公共管理学院, 武汉 430074;

5. 中国科学院大学资源与环境学院, 北京 100049

* 通讯作者, E-mail: zhengjy@igsrr.ac.cn

收稿日期: 2020-05-24; 收修改稿日期: 2021-08-29; 接受日期: 2021-11-05; 网络版发表日期: 2021-12-17

中国科学院战略性先导科技专项(A类)项目(编号: XDA19040101)和国家重点研发计划项目(编号: 2017YFA0603301)资助

摘要 本研究利用现代人口和环境因子数据, 通过地理探测器分析辨识了人口分布的显著影响因子, 进而构建了基于随机森林回归算法的人口空间分布模型, 最后以史学界考订的历史人口数据为基础, 重建了1776~1953年6个时间断面中国传统农区10km×10km分辨率的人口空间分布, 并分析了人口分布的时空变化特征. 研究发现: (1) 显著影响中国传统农区县域之间人口密度差异的主要环境因子包括海拔、坡度、地势起伏度、距地区(州、府)级城市与省会城市的远近、距河流远近及气候湿润指数. (2) 利用2000年传统农区1934个县级人口及上述环境因子数据, 基于随机森林回归算法构建的人口空间分布模型的方差解释量(R^2)为0.81, 88.4%的县(区)的模型预测误差小于50%. (3) 1776~1953年研究区人口总体呈增加趋势, 1851年之前主要表现为长三角地区人口持续增长, 该地区500人/km²以上的网格数由292增至683个; 1851~1953年则以多地区人口普遍增加为主要特征, 华北平原500人/km²以上的网格数由36增至88个, 珠江三角洲地区由4增至35个. 各时期人口分布的空间聚集特征不一, 其成因涉及经济重心转移、传统观念、战乱、灾荒、移民政策等多方面. (4) 本文重建与HYDE数据集在数据源、环境因子遴选和建模方法上均有所不同, 与本文重建结果相比, 1776~1851年HYDE的人口总体上东部偏少, 西部偏多, 1851年后城镇偏多, 乡村偏少.

关键词 人口, 传统农区, 历史时期, 格网化重建

1 引言

近三百年来, 人类活动是全球环境变化的重要驱动力, 深刻地改变了陆表景观格局和大气成分(IPCC, 2007). 据估算, 过去300年间, 全球陆地面积的42~68%

均不同程度的受人类活动的扰动(Hurt等, 2006). 人类活动主导的土地利用/覆被变化以及农业生产导致的碳排放量约占全球温室气体排放总量的20~25%(Searchinger等, 2018). 人类活动强度与人口密度密切相关, 人口密度的空间差异决定了人类活动对环境扰

中文引用格式: 张学珍, 王发浩, 路伟东, 李士成, 郑景云. 2022. 1776~1953年中国传统农区人口的格网化重建. 中国科学: 地球科学, 52(3): 528~539, doi: 10.1360/SSTe-2020-0133

英文引用格式: Zhang X, Wang F, Lu W, Li S, Zheng J. 2022. Gridded reconstruction of the population in the traditional cultivated region of China from 1776 to 1953. Science China Earth Sciences, 65(2): 365~378, <https://doi.org/10.1007/s11430-020-9866-2>

动强度的空间格局. 在历史土地利用/覆被变化(Kaplan等, 2011; 何凡能等, 2019; 方修琦等, 2020)、历史水资源开发和利用(Qin等, 2019)研究中, 人口空间分布均被作为重要的基础数据. 历史时期的人口空间分布重建成为历史地理、环境演变及全球变化领域的重要基础研究.

人口统计资料通常以行政区作为统计单元, 难以刻画画区内部的空间差异, 并且与栅格形式的自然环境数据匹配性差, 因而阻碍了对人口与资源、环境空间配置关系的深入认识. 为解决这一问题, 人口分布的格网化模拟, 作为一种降尺度方法被广泛应用(柏中强等, 2013; Sorichetta等, 2015; Leyk等, 2019). 尽管人口格网化方法极其多样, 但是其数学基础均是空间统计与建模, 具体是基于统计分析理论在空间域建立人口密度(因变量)与环境要素(自变量)的映射关系. 各种方法之间的差异主要体现在两个方面: (1) 数学模型; (2) 环境要素(自变量). 就数学模型而言, 既有线性模型(Yang等, 2013; Tan等, 2018), 也有非线性模型(Linard等, 2017; Li等, 2018); 就自变量类别而言, 既有仅以地理坐标作为自变量的模型(王平等, 2010), 也有以多类型的环境要素作为自变量的模型(Klein Goldewijk等, 2010; Fang和Jawitz, 2018); 就自变量数量而言, 既有采用单一环境要素作为自变量的模型(Deville等, 2014), 也有采用多种环境要素的多变量模型(Tatem, 2017; Han等, 2019; Yang等, 2019). 鉴于人口分布受多种要素的制约, 既有自然环境要素, 也有人文环境要素, 并且各种要素对人口分布的影响是极其复杂的非线性关系. 因而, 多元非线性模型在人口数据的格网化中应用相对较多.

相比于现代人口数据的格网化研究, 历史人口数据的格网化研究较为薄弱, 特别是过去300年中国传统农区人口格网化研究几近空白. 就其原因, 一方面受制于历史时期的人口统计数据. 历史时期不存在现代意义的人口普查, 通常以户、丁等为计量单位, 且历经政区沿革, 以及史料数据的真伪等一系列问题. 因而, 要得到与现代人口普查可以互相对比的数据, 需要大量细致的考订工作. 另一方面受限于技术方法. 如上所述, 多元非线性模型应用较为广泛, 这类模型通常需要较多的自变量, 而大多数自变量(例如: 植被指数、不透水面的面积、夜间灯光指数等)在历史时期难以获取. 近年来, 薛樵风等(2019)通过构建历史人口分布适应性模型重建了1776和1820年苏州府人口空间

分布. 但是, 其模型采用了精细的历史时期城镇空间规模数据, 这对于中国传统农区地理范围而言几乎是不可获取的数据; 另外, 其模型中采用了大量的经验参数, 这些参数具有较强的区域特征, 能否推广至更大的地理空间也需要开展大量研究. Wang等(2020)基于一种非线性模型——随机森林模型, 采用城市、地形、气候、河流要素作为指示人口空间分布的环境因子, 针对甘肃省人口格网化进行了探索性研究, 发现这一方法不仅误差较小, 而且变量易获取, 对于历史人口的格网化重建而言具有较强可操作性; 进而王发浩(2020)开展了中国中、东部清代17个行省的人口格网化重建.

中国传统农区一直是人口密度大、人类活动强烈的典型地区. 本文拟在前述研究的基础上, 以清代中国传统农区(辖18个行省)为研究区, 在量化分析人口分布与环境因子关系的基础上构建基于随机森林回归算法的人口格网化模型, 最后以史学界考订的人口数据为基础, 在10km×10km尺度上开展1776~1953年人口空间分布格网化重建, 并分析其演变的时空特征, 以期为历史时期人口空间分布研究与人-地关系及其演变研究提供基础数据.

2 数据与方法

2.1 研究区概况

本文研究区范围大致包括内蒙古高原以南、青藏高原以东的地区, 辖18个行省(不包括台湾府, 图1). 作为中国耕作文明的主要分布区, 该区自古至今都是人口分布最为密集的地区. 研究区主体位于温带、亚热带, 以湿润和半湿润性季风气候为主, 雨热同期, 年降水量为400~2000mm. 地势西高东低, 东部以平原和低山丘陵为主, 包括华北平原、长江中下游平原、东南丘陵等, 西部以高原、盆地为主, 包括黄土高原、云贵高原、四川盆地等. 大部分地区海拔不超过2000m.

2.2 数据来源

本文采用的主要数据包括: 古、今人口统计数据与政区数据, 以及地形、气候和河流三类环境因子数据.

(1) 人口统计数据. 现代人口数据来自中国人口分布标准化数据库(杨艳昭, 2016), 系第五次全国人口普

图1 研究区的地理范围(1820年)

来自中国历史地理信息系统

查(2000年)数据,统计单元为区、县(县级市)。历史人口数据涉及册载资料相对详细的6个时段,依次为1776年、1820年、1851年、1880年、1910年、1953年,以州、府(地级市)为基本统计单元,涵盖清代主要人口增长期、民国时期以及中华人民共和国成立初期等三个典型时段(曹树基, 2001)。该数据以“人口”为计量单位,与清早期(1776年之前)以“丁(户)”为计量单位的数据相比,更具现代统计意义(路伟东, 2014)。此外,经史学界考订与点校,该数据的统计单元为1820年《嘉庆重修大清一统志》的府级行政区。

(2) 政区数据。现代政区采用2000年行政区划数据,来自1:100万全国基础地理数据集(<http://www.web-map.cn/>)。历史时期行政区划数据采用《嘉庆重修大清一统志》政区,根据《中国近现代政区沿革表》(张在普, 1987)的行政驻点沿革信息,校正各时段行政驻点,由中国历史地理信息系统(<http://www.people.fas.harvard.edu/~chgis/>)提供。

(3) 地形数据。包括海拔、坡度、地势起伏度三个指标,源数据系30m×30m空间分辨率的ASTER GDEM数据(<https://www.nasa.gov/>)。利用GIS软件提取海拔和坡度,并以5km×5km范围内最高与最低海拔之差,作为地势起伏度。

(4) 气候和河流数据。气候数据指气候湿润指数,来自中国气象背景数据集。该数据为多年平均的地面

水分收入(降水)与支出(蒸发、径流)之比,来自中国科学院资源环境科学数据中心(徐新良和张亚庆, 2017)。现代河流数据采用中国主要水系空间分布数据,来自国家地球系统科学数据中心(<http://www.geodata.cn/>)。历史水系数据采用史学家考订的清代中国河湖数据,来自中国历史地理信息系统。

2.3 研究思路与方法

本研究首先利用2000年县域尺度的数据,采用地理探测器方法,量化了各环境因子对人口空间分布的影响程度,遴选了影响显著的环境因子。然后,以这些因子为自变量,基于随机森林回归算法建立人口密度预测模型,并对模型精度进行检验。之后,再基于不同历史时段环境因子数据及经过考订的册载历史人口数据,开展历史时期人口空间分布重建。最后,通过空间自相关分析,揭示1776~1953年研究区人口空间分布特征与变化过程;同时,通过对比分析,揭示本文重建与HYDE人口数据集的异同。

2.3.1 环境因子遴选

人口空间分布的影响因素复杂多样,以往在遴选环境因子时通常采用单因素评价或相关性分析的方法,忽视了环境因子之间复杂的交互关系,从而难以遴选出主要且相对独立的环境因子。本文基于地理探测器方法进行环境因子遴选。该方法是一种基于空间分异理论探测地理要素空间分异规律成因机制的统计学方法。以自变量与因变量空间分布相似程度来量化自变量的影响力,用 q 值表示,取值范围为[0,1],越接近1,表示自变量的解释力越强(王劲峰和徐成东, 2017)。该方法不但可以探测数值型数据,而且可用于定性数据的统计分析,且可以实现因子间交互关系的探测,被广泛应用于影响因子选取与因子间交互作用研究(Luo等, 2016; 刘彦随和李进涛, 2017)。

基于三方面的考虑:(1) 对人口空间分布有显著影响;(2) 近300年内该因子相对稳定;(3) 可以被量化。本文将两类因子纳入地理探测器分析。一类是地形、气候、水系等自然地理因子,包括海拔、坡度、地势起伏度、湿润指数、距1~3级河流的远近等指标;另一类是以距中心城市远近为代表的城市因子,包括距地级市(府城)的远近、距省会城市(省城)的远近等指标。

2.3.2 基于随机森林回归算法的人口空间分布建模及检验

随机森林回归算法隶属于机器学习算法的范畴, 其实现方式是分类回归树(CART)(Breiman, 2001). 因其能较好地表达人口与环境因子之间的非线性关系, 而被广泛用于现代人口格网化分布模拟(Stevens等, 2015; Ye等, 2019; 王超等, 2019). 本研究的建模过程采用了2000年研究区1934个县区的统计数据, 以县区为统计样本, 人口密度为因变量, 经地理探测器筛选后的环境因子为自变量.

随机森林建模的第一步是构建样本子集. 具体而言, 本研究采用有放回抽样(Bootstrap)的方式从样本库中随机抽取一定数量的样本构建子集, 子集样本量大小与样本库一致. 第二步是建立决策树, 采用二叉树树形结构, 从根节点至叶节点递归生长(“分叉”), 其不经剪枝. 这一步的关键点是确定每一个节点“分叉”的主控因子. 本研究从所采用的环境因子中随机抽取2个环境因子, 利用CART方法依据方差最小原则从中选取一个环境因子作为该节点的分叉主控因子(Rodriguez-Galiano等, 2014; 宋建国等, 2016), 并建立分叉准则.

重复上述过程, 本研究建立了300棵决策树. 据此, 给定一组环境参数, 可以获得300个预测结果. 最后, 根据大样本无偏估计的思想, 计算300个预测结果的平均值, 将其作为最终预测结果.

在模型精度和稳定性检验环节, 本研究采用了留一法(Vehtari等, 2017). 具体的, 从1934个县区的(样本)的数据集中, 预留其中一个县区的数据, 利用其余1933个县区的开展随机森林建模, 之后利用该模型模拟预留县区的人口密度. 如此循环迭代, 经过1934次建模和模拟试验, 可获得每个县区的人口密度预测值, 通过与普查人口数据进行比对即可检验模型的精度和稳定性. 采用方差解释量(R^2)、相对误差、均方根误差三个指标定量表征模型的精度, 采用平均有效系数(CE)和误差缩减值(RE)两个参数定量表征模型稳定性. 其中, RE和CE的取值范围为 $(-\infty, 1)$, 越接近1, 表明模型的稳定性越高(Gou等, 2015).

2.3.3 1776~1953年传统农区人口分布格网化重建

将各时期10km×10km网格尺度的环境因子数据

输入随机森林模型, 即可模拟各格网的人口密度. 不过, 考虑到模型训练过程采用的是2000年的统计数据, 其系数表征的是2000年人口密度与各环境因子的数量关系; 本研究时段内研究区总人口经历了大幅增加, 因而模拟结果与历史人口密度相差甚大. 然而, 由于网格人口密度的相对高低受控于环境因子所决定的“宜居性”, 因此模拟结果可以用于表征历史时期网格之间人口密度的相对高低. 故本研究在州(府)单元上利用古今人口总数之比作为系数, 对模型预测的格网人口密度进行了调整(式(1)).

在格网化过程中, 针对位于行政区边界线上同属于两个及以上行政区的网格, 人口密度通过面积权重法计算(式(2)). 同时, 针对河流湖泊等水域, 在“无土地无人口”的原则下, 根据历史时期河流和湖泊的分布, 将大型湖泊水域和1~2级河流的人口分配权重设定为0, 在州府单元内对网格人口按实际土地面积再次修正(林珊珊等, 2008).

$$P'_{ij} = P_{ij} \times \frac{S_i}{\sum_{j=1}^n D_{ij} P_{ij}}, \quad (1)$$

式中, P_{ij} 和 P'_{ij} 分别为政区*i*内网格*j*修正前和修正后的人口密度; S_i 为历史时期政区*i*的人口总数; D_{ij} 为网格*j*属于政区*i*的部分的面积.

$$P''_j = \frac{\sum_{i=1}^n P'_{ij} D_{ij}}{D_j}, \quad (2)$$

式中, j 为网格编号, 其土地分属于 $i(i \geq 2)$ 个行政区, P'_j 表示网格*j*人口密度的修正值, D_{ij} 为网格*j*内隶属于行政区*i*的土地面积, D_j 为网格*j*的总面积.

2.3.4 基于格网化数据的历史时期人口分布集聚模式分析

在空间分析领域, 通常采用空间自相关分析方法刻画地理现象的空间非随机分布特征, 揭示地理现象的空间集聚模式(Legendre, 1993). 其度量指标为莫兰指数(Moran's I), 是反映空间邻接和相邻单元属性相似度的统计量, 取值范围为 $[-1, 1]$. 其值越接近1, 表示相邻单元的空间属性的相似度越高; 越接近-1, 表示相邻单元空间属性的差异性越大; 接近于0, 表示相邻单元不存在空间相关关系(刘涛等, 2015; 王胜今和杨鸿儒, 2019). 根据空间范围尺度大小, 分为全局莫兰指

数(Global Moran's I ; 式(3))和局部莫兰指数(Local Moran's I ; 式(4)).

$$I = \frac{n \sum_{i=1}^n \sum_{j=1}^n W_{ij} (X_i - \bar{X})(X_j - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n \sum_{j=1}^n (W_{ij})}, \quad (3)$$

式中, X_i 和 X_j 分别为格网 i 和 j 的人口密度, $\bar{X}$ 为研究区人口密度平均值, n 表示研究区内格网的数量, W_{ij} 表示空间权重矩阵的相应元素(其中, 空间相邻为1, 不相邻为0).

$$I_i = Z_i \sum_{j=1}^n W_{ij} Z_j, \quad (4)$$

式中, Z_i 和 Z_j 为格网 i 和 j 标准化后的观测值. 在一定的显著性水平下, 根据 I_i 和 Z_i 的取值, 分为如下情况: 当 I_i 和 Z_i 均为正数, 则格网 i 处在H-H象限(高-高集聚); 当 I_i 为正数而 Z_i 为负数, 则表示格网 i 处在L-L象限(低-低集聚); 当 I_i 为负数而 Z_i 为正数, 则格网 i 处在H-L象限(高-低集聚); 当 I_i 和 Z_i 均为负数, 则格网 i 处在L-H象限(低-高集聚).

3 结果分析

3.1 环境因子遴选结果

经地理探测器分析发现, 地形、城市、气候、河流四类因子均对人口空间分布有显著影响, 其中地形、城市、气候因子的显著性水平达0.01, 河流因子的显著性水平为0.1. 如表1所示, 影响力最大的环境因子为地形, 位居第二位的是城市, 之后是气候和河流因子. 从 q 值来看, 地形因子的 q 值为0.37~0.55, 明显高于城市因子的 q 值, 0.21~0.37, 气候与河流因子的 q 值仅为0.02, 低于地形与城市因子的 q 值1个数量级. 这一结果

表1 地理探测器单因子探测 q 值^{a)}

环境因子	q	p
海拔	0.55***	0
坡度	0.37***	0
地势起伏度	0.37***	0
距地级市远近	0.37***	0
距省会远近	0.21***	0
湿润指数	0.02***	0.01
距河流远近	0.02*	0.09

a) ***表示在0.01信度水平下显著, *表示在0.1信度水平下显著

说明在中国传统农区地域空间内, 人口密度与地形因子地域分布规律的一致性较强, 人口密度与城市辐射能力地域分布规律的一致性相对较弱, 地形对人口密度空间分布差异的解释力要强于城市因子.

从各地形因子 q 值对比来看, 海拔的 q 值最高, 达0.55, 坡度与地势起伏度的 q 值大致相当, 均为0.37, 说明海拔对人口密度空间差异的解释力强于坡度和地势起伏度; 就两个城市因子而言, 距地级市远近的 q 值为0.37, 略高于距省会城市远近的 q 值0.21, 说明地市级城市的辐射能力对人口密度空间差异的解释力强于省会城市. 气候湿润指数和距河流的远近对人口密度空间分异也有显著影响, 气候较为湿润的地方和距离河流近的地方, 人口较为稠密, 反之亦然. 但是, 这二者对人口密度空间分异的影响力较低, 这表明研究区内的气候湿润度变幅处于人类生产与生活的可适应范围, 对人口分布的影响较弱, 距离河流的远近也未能深度影响人口分布. 河流影响力较弱的主要原因可能是研究区基本位于湿润和半湿润区, 水系较为发达, 绝大多数县与前三级河流的距离较小, 难以分辨出明显差异.

表2 地理探测器交互作用探测 q 值

	海拔	坡度	地势起伏度	距地级市远近	距省会远近	湿润指数	距河流远近
海拔	0.55						
坡度	0.64	0.37					
地势起伏度	0.64	0.38	0.37				
距地级市远近	0.69	0.61	0.62	0.37			
距省会远近	0.64	0.47	0.48	0.46	0.21		
湿润指数	0.62	0.41	0.43	0.40	0.25	0.02	
距河流远近	0.58	0.39	0.40	0.41	0.28	0.09	0.02

再从双因子组合的交互作用看(表2), 绝大多数双因子组合的交互作用明显大于单因子的影响力. 其中影响力最大的是距地级市(州、府城)的远近与海拔组合($q=0.69$), 其次是距省会(省城)的远近与海拔组合($q=0.64$)及坡度与海拔($q=0.64$)组合. 由此来看, 地形因子与城市因子组合的影响力明显大于地形与气候、河流单一因子的影响力. 这进一步印证了传统农区人口空间分布受自然与人文多重因素的综合影响. 基于地理探测器分析的结果, 本文在人口格网化建模过程中采用的环境因子包括: 地形因子(海拔、坡度、地势起伏度)、城市因子(距地级市的远近、距省会城市的远近)、气候因子(湿润指数)和河流因子(距1~3级河流远近).

3.2 基于随机森林的人口格网化模型检验

基于上述四类共7个因子建立的随机森林模型可基本再现2000年县域人口密度的空间分布特征. 如图2所示, 模型预测的人口密度与普查数据之间显著正相关, 相关系数为0.9($p<0.001$), 意味着模拟预测对普查数据的方差解释量达81%. 并且, 平均有效系数(CE)和模型误差缩减值(RE)分别为0.72和0.75, 表明模型是基本稳定的. 不过模型预测低估了人口高密度区的数值, 高估了人口低密度区的数值, 从而使得模型预测的人口密度空间差异略低于普查数据(图2).

从模型预测误差的统计分布形态来看, 误差总体呈正态分布(图3a), 约88.4%的模型预测误差低于50%, 仅4.1%的模型预测误差高于80%, 表明大多数预测结果的误差较小, 而极少数预测结果误差较大, 符合统计模型误差分布的一般特征, 这进一步说明了模型的稳定性. 图3b展示了误差的空间分布特征, 低误差主要出现在黄淮海平原及长江沿江低地地带, 高误差主要出现在研究区西北部 and 西南部山区, 以及东南沿海、关中平原、江南丘陵与成都平原. 从误差的正、负分布看, 正误差主要分布在西北部 and 西南部山区及江南丘陵, 表明在这些地区模型预测的人口密度高于实际值, 而负误差主要分布在关中地区、东南沿海和成都平原, 表明在这些地区模型预测的人口密度低于实际值. 其主要成因是模型未考虑宏观经济区位因素对人口分布的影响. 关中平原和成都平原地形平坦, 农业发达, 交通便利, 均为区域经济中心; 东南沿海虽然多为低山丘陵地貌, 但其海运便利, 其经济水平也高

图2 2000年县域尺度上随机森林回归模型预测人口与普查数据对比散点图

于临近地区, 因而人口聚集, 人口密度较大. 然而, 根据地理探测器分析结果, 高海拔和地形复杂地区的人口密度较低. 关中及成都平原海拔(约500~600m)远高于华北平原(0~50m), 东南沿海地区多为低山丘陵地貌, 因而模型预测的人口密度低于实际值.

3.3 1776~1953年传统农区格网化的人口分布格局

从重建结果来看, 1776~1953年中国传统农区人口空间分布格局总体较为稳定, 主要呈现东多西少的分布特征. 长江中下游平原、华北平原、关中平原、四川盆地、珠江三角洲地区的人口密度较高, 而川西高原、云贵高原、东南丘陵、黄土高原、内蒙古高原及河西走廊以北的人口密度相对较低. 在网格尺度上, 全局莫兰指数 I 值均大于0($p<0.01$)(表3), 说明传统农区人口空间分布的聚集特征明显, 但不同时段特征不尽相同.

1776~1851年间, 随着通商口岸开辟和运河商贸发展, 长三角地区成为传统农区人口增幅最大的地区(图4a~4c). 在此期间长三角地区人口密度达500人/ km^2 以上的网格数由292增至683个, 增幅高达134%, 苏州府城所在网格人口密度为全区最高, 超过5000人/ km^2 . 在东部各省, 省城和府城是人口聚集的主要地区, 人口密度明显较高, 顺天府、开封府、济南府、南昌府等府城所在网格人口密度皆高达1500人/ km^2 以上. 东部平原腹地广大地区人口虽有差异, 但均保持较高水平,

审图号: GS(2021)7277号

图3 随机森林回归模型相对误差分布直方图(a)和空间分布图(b)

表3 1776~1953年传统农区10km×10km网格尺度人口密度全局Moran's I指数与Z值

年份	Moran's I	Z
1776	0.85	247.86
1820	0.85	245.52
1851	0.85	245.58
1880	0.83	242.25
1910	0.81	236.68
1953	0.82	236.86

由此形成人口的高密度集聚特征,使其成为主要的人口高密度集聚区(图5a~5c)。在此期间,东部城镇人口密度虽有明显上升,但人口高密度集聚区域面积增幅甚微,主要原因是当时农业生产水平下人口几近饱和,同时受故土难迁的传统观念影响,农民不愿离开他们的祖籍地(裴卿, 2017),因而未出现明显的扩散效应。而同时期西部地区四川盆地与云贵高原的城镇人口持续增加,同时向外围扩散,成都平原上城镇聚落开始发育,人口集聚特征不断加强。至1851年,成都府城及周边地区成为人口高密度集聚区,府城所在网格人口密度达1770人/km²。这一现象的主要成因是康乾

时期移民大量迁入导致的人口高增长惯性。康乾时期近百万人移入四川,奠定了清中后期该地区人口大幅增加的基础,人口空间集聚特征也随之加剧。

1851~1880年间,研究区人口密度变化呈现鲜明的区域差异(图4d)。一方面,因战乱和灾荒,苏南、浙北、皖南、陕中、甘南等地区城镇人口剧减。因太平天国运动,长江中下游地区人口损失严重。至1880年,长江中下游地区人口密度超过500人/km²的网格数下降至145个,仅约为1851年的20%。其中,苏州府城所在格网人口密度由5235人/km²降至1894人/km²,降幅高达64%。因“丁戊奇荒”和同治西北战争,甘南、豫北及山陕境内的大同-西安-乾州-凤翔一线人口严重减少,各府城所在网格人口密度出现200~500人/km²不等的降幅。由于城镇地区人口损失,人口的空间集聚特征随之而变,长江中下游人口高密度集聚区域呈现向北扩展向东收缩的趋势,关中地区人口高密度集聚区几近消失。另一方面,随着近代工商业的发展,城市的功能与地位逐渐强化,战乱和灾荒影响较小的华北平原、四川盆地、珠三角地区城市发育,人口一直维持增加态势并且呈现出明显的空间集聚特征(图5d)。至1880年,华北地区冀南、山东各府城所在网格人口密度普

图 4 1776~1953 年中国传统农区人口密度空间格局(10km×10km 网格)

遍增长50~200人/km², 形成了以山东、冀南、豫东北为主的人口高密度集聚区域。同时, 四川盆地及珠三角地区人口高密度集聚区初具规模。

1880~1953年间, 传统农区平原和东南沿海地区人口密度普遍增长(图4e~4f)。至1953年, 华北平原人口密度达500人/km²以上的网格数增至88个, 珠江三角洲增至35个, 分别约是1851年的2.4倍和8.75倍。北京、上海等城区人口密度达4000人/km²以上。至此, 以华北地区和长江中下游地区为代表的东部人口高密度集聚区基本形成。此外, 因农业活动范围向北扩展, 华北平原北部人口集聚区持续北扩, 低密度区域向北退缩。西部地区, 在经历了长期的人口增长和扩张后, 四川盆地内部人口密度显著上升。至1953年, 成都平原人口密度达500人/km²以上的网格数增至60个, 重庆地区增至12个, 逐渐形成了以四川中东部和重庆为核心的人口高密度集聚区域(图5e~5f)。云贵高原地区人口虽有增长, 但增幅较小, 未显现出明显的空间集聚特征。

3.4 与HYDE人口数据集的对比

目前重建时序最长的全球格网化历史人口数据集是HYDE 3.2人口数据集(Klein Goldewijk等, 2017), 在时、空维度上与本文重建结果部分重叠。对于部分不是完全一致的年份, 本文选择了二者最为接近的年份进行比对。如图6所示, 本文重建与HYDE数据集展示的人口空间分布格局基本一致, 但在部分区域存在明显差异。主要表现为: 1776~1851年间, HYDE数据集在东部平原地区人口密度低于本文重建, 在长江中下游地区以及各省省城、府城等城镇地区尤为明显, 其中江南部分城镇所在网格人口密度偏低幅度超过500人/km²。而HYDE数据集在研究区北部和西部高原山地地区的人口密度较本文重建整体偏高。偏高最为严重的地区为四川盆地, 除成都平原外的大部分地区网格偏高50~200人/km²。1851~1953年间, HYDE数据在研究区东部地区较本文重建仍然偏低, 偏低幅度较先前减小, 然而二者之差呈现出了新的特征——城乡差异, HYDE数据在城镇地区偏高, 在乡村地区偏低; 在

图 5 1776~1953年传统农区人口分布LISA聚类图(10km×10km网格)

城镇发展滞后乡村占据绝对主导地位的西部地区, HYDE数据普遍低于本文重建结果。据此可知, HYDE低估了1851~1953年间乡村地区的人口, 而高估了城镇地区的人口。

本文重建与HYDE数据的差异主要源于数据源和分配方法的不同(表4)。HYDE3.2的数据源是Populstat估算的亚国家级(省级)人口数据, 而本文重建的数据源是以历史文献记载为基础并经过史学专家考证的府级人口数据。两套数据的人口总量不同, 造成了HYDE数据与本文重建人口密度的区域性整体差异。这可能是1776~1851年HYDE数据在西部地区较本文重建偏高, 在东部地区则偏低, 以及1851~1953年西部多地偏低的主要原因。在分配方法上, HYDE参照LandScan数据集的现代人口空间分布型进行历史人口空间分配, 由此可能导致城镇地区赋予较高的人口权重, 从而高估城镇地区的人口。这是因为工业化之前, 中国社会仍然是以农业为主题, 城镇虽然具有一定的吸引力,

但是城乡人口密度之差远小于当前。本文重建则以经过地理探测器分析遴选的人口分布主要影响因子为依据, 在人口空间分布模型中考虑了历史时期动态的城市等级及位置, 体现了多因子对人口分布的共同影响及因时而变的城市体系对人口分布的影响。分配方法的上述差异可能是1851~1953年间HYDE东部城镇人口较本文重建偏高而乡村人口偏低的主要原因。

4 讨论

本文基于中国史学界严格考订的府级人口数据, 通过格网化重建, 精细地揭示了1776~1953年传统农区人口空间分布的变化规律。人口空间分布建模是重建过程的关键环节, 模型精度对重建结果具有深刻影响。为认识重建结果的不确定性, 在此对模型的不足及其对重建结果的可能影响进行讨论。

首先, 在环境因子选取上, 受历史时期环境因子数

图 6 本文重建与HYDE重建的格网化人口密度对比(HYDE减本文重建)

表 4 本文重建与HYDE人口数据集的数据源及格网化分配模型采用的环境因子比较

数据集	原始人口数据	模型环境因子
本文重建	基于历史文献的分府人口	海拔、坡度、地势起伏度、距城市远近、气候湿润指数、距河流远近
HYDE	Populstat分省人口	LandScan人口分布格局、土壤生产力、坡度、距河流远近

据可获取性的限制,可供本文选用的环境因子较少,且历史时期部分环境因子数据欠缺,而采用现代数据代替之。例如,由于历史时期高精度空间格网化的气候湿润指数数据欠缺,本文采用了现代气候湿润指数。根据历史气候重建结果,18~19世纪正处于小冰期,温度较现代偏低0.3~0.6℃(Ge等, 2013),并且较现代偏湿,华北地区尤甚(Zheng等, 2006)。但是,由于现有重建大多是区域性的,不足以揭示州府(地市)范围内的空间差异,因而尚难以评估湿润指数的古今替代对格网化人口数量模拟结果的影响。一般认为地形复杂地区气候变化的空间差异也较大,据此可推断湿润指数古今替代对地形复杂地区重建结果的影响大于地形平坦地区。

其次,本文以距城镇的远近作为城市因子影响的

衡量标准,但在不同社会形态下产业结构不同,历史时期城镇对人口的吸引力要远弱于现代,特别是改革开放后,随着城市化进程的加速,城镇地区的人口集聚效应也在不断增强。因而,基于现代数据构建的人口格网化模型,可能导致历史时期城市及周边地区人口预测值偏高。此外,模型中未考虑经济区位因素,不过由于本文重建采用的是州府级人口数据,在数据层面上已包含国家级经济区位的影响,因此对宏观尺度人口分布影响可能较小。

5 结论

通过上述分析,可以发现地形因子(海拔、坡度、地势起伏度)是造成中国传统农区县域之间人口分布

差异的主要环境因子, 其次是城市因子(距离省会和地级市的远近), 再次是气候因子(湿润指数)和河流因子(距1~3级河流远近); 以上述因子为自变量, 基于随机森林回归算法构建的人口格网化模型的方差解释量为0.81($p<0.01$), 表明该模型较好地再现了人口的空间分布格局。利用该模型, 结合经考订的历史人口数据, 重建了1776~1953年6个时间断面分辨率为10km×10km的人口空间分布。据此发现, 在此期间中国传统农区人口总体呈增加趋势, 1851年之前主要表现为长三角地区人口持续增长, 1851~1953年则主要表现为人口普增, 以平原地区和东南沿海地区为主; 同时, 人口分布存在显著的空间集聚特征, 但在不同时段集聚特征不尽相同, 其成因涉及经济重心转移、传统观念、战乱、灾荒、移民政策等多方面。与本文重建相比, 1776~1851年间HYDE数据集低估了东部地区人口, 高估了西部地区人口, 1851~1953年间高估了东部城镇人口, 低估了乡村人口, 其原因是本文重建与HYDE的数据源、环境因子遴选和建模方法均不同。

本文重建结果为研究传统农区历史时期资源环境与人类社会相关问题, 诸如气候变化对人类社会的影响、区域环境演变与人类活动的相互作用、重大历史事件的背景与影响、疫疾风险分析、社会资源承载力评估以及人类活动驱动的土地利用/覆被变化及其气候生态效应提供重要的数据支持。不过, 人口空间格网化模拟是对人口空间真实分布的“近似”, 受模型框架与历史时期环境因子数据可获取性的影响, 重建结果尚有预定的不确定性, 有待于进一步完善。最后, 20世纪50年代以后人口数据格网化采用的因变量、模型结构、网格尺度均不同于本文, 因而其与本文重建结果之间存在系统性差异, 如何消除该差异成为本领域面临的一个重要问题。

参考文献

柏中强, 王卷乐, 杨飞. 2013. 人口数据空间化研究综述. 地理科学进展, 32: 1692-1702

曹树基. 2001. 中国人口史(第五卷, 清时期). 上海: 复旦大学出版社. 691-719

方修琦, 赵婉一, 张成鹏, 张頔颐, 魏学琼, 邱修理, 叶瑜. 2020. 全球历史LUCC数据集数据可靠性的评估方法及评估案例. 中国科学: 地球科学, 50: 1009-1020

何凡能, 李士成, 杨帆, 李美娇. 2019. 全球历史土地利用数据集中国

草地数据可靠性评估. 中国科学: 地球科学, 49: 421-433

林珊珊, 郑景云, 何凡能. 2008. 中国传统农区历史耕地数据网格化方法. 地理学报, 63: 83-92

刘涛, 齐元静, 曹广忠. 2015. 中国流动人口空间格局演变机制及城镇化效应——基于2000和2010年人口普查分县数据的分析. 地理学报, 70: 567-581

刘彦随, 李进涛. 2017. 中国县域农村贫困化分异机制的地理探测与优化决策. 地理学报, 72: 161-173

路伟东. 2014. GIS支撑下的长时段区域人口变动规律分析——以1776年至1953年陕甘地区人口为例. 历史地理, (2): 314-324

裴卿. 2017. 自然灾害与移民: 一个中国历史上农民的被动选择. 中国科学: 地球科学, 47: 1406-1413

宋建国, 高强山, 李哲. 2016. 随机森林回归在地震储层预测中的应用. 石油地球物理勘探, 51: 1202-1211, 1052-1053

王超, 阚媛珂, 曾业隆, 李国庆, 王民, 次仁. 2019. 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, 74: 664-680

王发浩. 2020. 融合多源数据的历史时期人口空间分布建模与应用——以18~20世纪中叶中国传统农区为例. 硕士学位论文. 山东: 山东师范大学. 68

王劲峰, 徐成东. 2017. 地理探测器: 原理与展望. 地理学报, 72: 116-134

王平, 王志伟, 张学通, 李春娥, 王迅, 冯琦胜, 陈全功. 2010. 西汉时期我国人口分布空间格局及其成因探讨. 西北人口, 31: 88-90, 96

王胜今, 杨鸿儒. 2019. 我国中原经济区人口与经济空间格局演变研究. 人口学刊, 41: 35-44

徐新良, 张亚庆. 2017. 中国气象背景数据集. 中国科学院资源环境科学数据中心数据注册与出版系统

薛樵风, 金晓斌, 韩娟, 杨绪红, 周寅康. 2019. 区域历史人口空间格局精细化重建: 方法与实证. 地理科学, 39: 1857-1865

杨艳昭. 2016. 中国人口分布标准化数据库(1982、1990、2000、2010、2012年). 中国科学院地理科学与资源研究所

张在普. 1987. 中国近现代政区沿革表. 福建: 福建省地图出版社. 13-266

Breiman L. 2001. Random forests. *Mach Learn*, 45: 5-32

Deville P, Linard C, Martin S, Gilbert M, Stevens F R, Gaughan A E, Blondel V D, Tatem A J. 2014. Dynamic population mapping using mobile phone data. *Proc Natl Acad Sci USA*, 111: 15888-15893

Fang Y, Jawitz J W. 2018. High-resolution reconstruction of the United States human population distribution, 1790 to 2010. *Sci Data*, 5: 180067

Ge Q, Hao Z, Zheng J, Shao X. 2013. Temperature changes over the past 2000 yr in China and comparison with the Northern Hemisphere. *Clim Past*, 9: 1153-1160

Gou X, Gao L, Deng Y, Chen F, Yang M, Still C. 2015. An 850-year tree-ring-based reconstruction of drought history in the western

- Qilian Mountains of northwestern China. *Int J Climatol*, 35: 3308–3319
- Han D, Yang X, Cai H, Xu X, Qiao Z, Cheng C, Dong N, Huang D, Liu A. 2019. Modelling spatial distribution of fine-scale populations based on residential properties. *Int J Remote Sens*, 40: 5287–5300
- Hurt G C, Frolking S, Fearon M G, Moore B, Shevliakova E, Malyshev S, Pacala S W, Houghton R A. 2006. The underpinnings of land-use history: Three centuries of global gridded land-use transitions, wood-harvest activity, and resulting secondary lands. *Glob Change Biol*, 12: 1208–1229
- IPCC. 2007. Climate Change 2007: The physical science basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge and New York: Cambridge University Press
- Kaplan J O, Krumhardt K M, Ellis E C, Ruddiman W F, Lemmen C, Klein Goldewijk K. 2011. Holocene carbon emissions as a result of anthropogenic land cover change. *Holocene*, 21: 775–791
- Klein Goldewijk K, Beusen A, Doelman J, Stehfest E. 2017. Anthropogenic land use estimates for the Holocene: HYDE 3.2. *Earth Syst Sci Data*, 9: 927–953
- Klein Goldewijk K, Beusen A, Janssen P. 2010. Long-term dynamic modeling of global population and built-up area in a spatially explicit way: HYDE 3.1. *Holocene*, 20: 565–573
- Legendre P. 1993. Spatial autocorrelation: Trouble or new paradigm? *Ecology*, 74: 1659–1673
- Leyk S, Gaughan A E, Adamo S B, de Sherbinin A, Balk D, Freire S, Rose A, Stevens F R, Blankespoor B, Frye C, Comenetz J, Sorichetta A, MacManus K, Pistolesi L, Levy M, Tatem A J, Pesaresi M. 2019. The spatial allocation of population: A review of large-scale gridded population data products and their fitness for use. *Earth Syst Sci Data*, 11: 1385–1409
- Li K, Chen Y, Li Y. 2018. The random forest-based method of fine-resolution population spatialization by using the international space station nighttime photography and social sensing data. *Remote Sens*, 10: 1650
- Linard C, Kabaria C W, Gilbert M, Tatem A J, Gaughan A E, Stevens F R, Sorichetta A, Noor A M, Snow R W. 2017. Modelling changing population distributions: An example of the Kenyan Coast, 1979–2009. *Int J Digital Earth*, 10: 1017–1029
- Luo W, Jasiewicz J, Stepinski T, Wang J, Xu C, Cang X. 2016. Spatial association between dissection density and environmental factors over the entire conterminous United States. *Geophys Res Lett*, 43: 692–700
- Qin Y, Mueller N D, Siebert S, Jackson R B, AghaKouchak A, Zimmerman J B, Tong D, Hong C, Davis S J. 2019. Flexibility and intensity of global water use. *Nat Sustain*, 2: 515–523
- Rodriguez-Galiano V, Mendes M P, Garcia-Soldado M J, Chica-Olmo M, Ribeiro L. 2014. Predictive modeling of groundwater nitrate pollution using Random Forest and multisource variables related to intrinsic and specific vulnerability: A case study in an agricultural setting (Southern Spain). *Sci Total Environ*, 476–477: 189–206
- Searchinger T D, Wiersma S, Beringer T, Dumas P. 2018. Assessing the efficiency of changes in land use for mitigating climate change. *Nature*, 564: 249–253
- Sorichetta A, Hornby G M, Stevens F R, Gaughan A E, Linard C, Tatem A J. 2015. High-resolution gridded population datasets for Latin America and the Caribbean in 2010, 2015, and 2020. *Sci Data*, 2: 1–2
- Stevens F R, Gaughan A E, Linard C, Tatem A J. 2015. Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data. *PLoS ONE*, 10: e0107042
- Tan M, Li X, Li S, Xin L, Wang X, Li Q, Li W, Li Y, Xiang W. 2018. Modeling population density based on nighttime light images and land use data in China. *Appl Geogr*, 90: 239–247
- Tatem A J. 2017. WorldPop, open data for spatial demography. *Sci Data*, 4: 170004
- Vehtari A, Gelman A, Gabry J. 2017. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Stat Comput*, 27: 1413–1432
- Wang F, Lu W, Zheng J, Li S, Zhang X. 2020. Spatially explicit mapping of historical population density with random forest regression: A case study of Gansu Province, China, in 1820 and 2000. *Sustainability*, 12: 1231
- Yang X, Ye T, Zhao N, Chen Q, Yue W, Qi J, Zeng B, Jia P. 2019. Population mapping with multisensor remote sensing images and point-of-interest data. *Remote Sens*, 11: 574
- Yang X, Yue W, Gao D. 2013. Spatial improvement of human population distribution based on multi-sensor remote-sensing data: An input for exposure assessment. *Int J Remote Sens*, 34: 5569–5583
- Ye T, Zhao N, Yang X, Ouyang Z, Liu X, Chen Q, Hu K, Yue W, Qi J, Li Z, Jia P. 2019. Improved population mapping for China using remotely sensed and points-of-interest data within a random forests model. *Sci Total Environ*, 658: 936–946
- Zheng J, Wang W C, Ge Q, Man Z, Zhang P. 2006. Precipitation variability and extreme events in Eastern China during the past 1500 years. *Terr Atmos Ocean Sci*, 17: 579–592

(责任编辑: 陈建徽)