基于随机森林模型的西藏人口分布格局及影响因素

王 超[1],阚瑷珂[2],曾业隆[3],李国庆[4],王 民[1],次仁[5]
(1. 北京师范大学地理科学学部,北京 100875; 2. 成都理工大学地球物理学院,成都 610059;
3. 中国科学院遥感与数字地球研究所,北京 100101; 4. 鲁东大学资源与环境工程学院,
烟台 264025; 5. 西藏自治区科技信息研究所,拉萨 850000)

摘要: 在乡镇尺度下厘清人口分布格局及其影响因素与区域差异,对在生态脆弱区制定可持续发展政策具有重大指导意义。基于2010年西藏自治区的乡镇尺度人口普查数据,提取人口密度和空间因子,利用空间统计方法分析了人口分布的疏密特征和集聚特征,对比运用多元线性回归方法和随机森林回归方法探索该地区人口分布的影响因素及其区域差异。结果表明:① 西藏乡镇人口密度在空间上表现出极强的非均衡性,其总体趋势是东南高西北低,高密度区与大江大河及主要交通干线具有较强的空间耦合性;② 大致以波绒乡(聂拉木县)—岗尼乡(安多县)为西藏的人口分界线,人口集聚的"核心—边缘"特征明显;③ 多元线性回归方法中,人造地表指数对人口分布的影响程度最大,随后依次为夜间灯光指数和路网密度;④ 利用随机森林方法进行的人口密度预测比多元线性回归方法精度高,可以用来对影响因子的重要性进行排序;排序在前六位的影响因子由高到低依次为夜间灯光指数、人造地表指数、路网密度、工业总产值、GDP和多年平均气温,它们与人口密度均呈正相关关系;地形地貌要素中以海拔和坡度的贡献率最大且与人口密度均呈负相关关系;⑤ 西藏人口分布格局的影响因素及其相互作用呈现出明显的区域差异特征,河谷是西藏地区人口的集聚区,主要分布在拉萨河谷、年楚河谷以及三江河谷;⑥ 通过随机森林回归分析,可以利用概念模型来表达人口分布影响因素,将主导因素概括为土地利用结构、道路通达度及城镇化水平。

关键词: 人口分布;影响因素;乡镇尺度;随机森林;概念模型
DOI: 10.11821/dlxb201904004

1 引言

人口分布是指某一特定时间段内某一空间范围的人口集散状态,同当地的地形、气候、资源等地理要素紧密相关,是自然因素和社会经济因素综合作用的结果,也是人类活动区域差异的客观反映[1]。不同人口分布状态下的人类活动对土地覆被、生态过程乃至全球变化产生重大影响[2-4]。揭示人口分布规律及其影响因素,对深刻理解人地关系,协调区域人口、资源与环境之间的矛盾[5-6],制定可持续发展政策具有重要指导意义。

西藏是青藏高原的主要组成部分,地形高差巨大、地质构造复杂、气候独特[7],是中

收稿日期: 2017-08-31; **修订日期:** 2019-03-11
基金项目: 西藏自治区自然科学基金项目(XZ2017ZRG-100, 2015ZR-13-56); 国家科技支撑计划(2014BAL07B02-2); 中国清洁发展机制基金赠款项目(2014058) [**Foundation:** Natural Science Foundation of Tibet Autonomous Region, No.XZ2017ZRG-100, No.2015ZR-13-56; The National Key Technology R&D Program of China, No.2014BAL07B02-2; Grants Program of China Clean Development Mechanism Fund, No.2014058]
作者简介: 王超(1992-), 女, 山西晋城人, 硕士生, 研究方向为区域可持续发展。E-mail: lrningxiang@163.com
通讯作者: 阚瑷珂(1980-), 男, 四川什邡人, 博士后, 讲师, 研究方向为人文地理与地理信息系统。
E-mail: kanaike@qq.com

国的“江河源”和“生态源”，更是中国的生态安全屏障[8]。然而，在人类活动和全球变化的综合影响下，该地区出现生态系统稳定性降低、资源环境压力增大等问题，突出表现为冰川退缩显著、水土流失严重、自然灾害增多、土地退化形势严峻（冻土退化、土地沙化及草地退化）等[9-11]，严重制约着该地区的可持续发展。人口分布研究的深入开展有利于把握人类活动规律，对加强区域认知与可持续发展具有重要作用。目前在探讨青藏高原人口分布格局及影响因素上，众多学者探讨了人口分布与海拔、土地利用、交通、河流水系、畜牧业等因素的关系[12-15]。另外，人口分布空间定量化也得到一定的发展[16-18]，进而利用1 km分辨率的人口分布栅格图探析人口分布作用于环境的定量关系[19-20]。

值得注意的是，目前对于西藏地区的人口分布研究还主要依赖于县级尺度的统计数据；虽然利用遥感技术可以实现公里网格尺度下的人口空间定量化，但是由于遥感固有的尺度依赖关系，其可塑性面积存在较大的不确定性[21-22]；另外由于区域人口分布规律独特，其影响因素是复杂、非线性的，已有研究多利用相关系数和线性回归的方法讨论其影响因素，未能深入刻画其关系[21, 23-24]。尽管公里网格的人口分布数据能展现更详细的空间细节，但由于东西部的自然条件及人口地域差异过大引起西部地区人口空间化存在较大误差[22]；而乡镇尺度是中国人口统计数据公开发布的最小统计单元，较县级尺度反映更多的空间异质性，故在青藏高原地区选择乡镇尺度进行人口分布研究更具优势。在影响因素探讨上，现有研究表明，随机森林不需要顾虑一般回归分析面临的多元共线性的问题，不需要做变量选择，便于计算变量的非线性作用，而且可以评估自变量的重要性[25]。因此，本文选取西藏自治区乡镇尺度最近一次人口普查数据（第六次人口普查，2010年）及相应年份的影响因子数据，利用空间统计方法分析人口分布的疏密特征和集聚特征，对比运用多元线性回归方法和随机森林回归方法探索该地区人口分布的影响因素及其区域差异，以期为人口分布精细化模拟与协调人口、资源与环境关系提供科学依据和决策支持。

2　研究区概况

西藏是青藏高原的主体，平均海拔在4000 m以上，地势起伏显著，地形以高原和山地为主；气候复杂多样，以高原高山气候为主；水系发达，湖泊广布。区域内受地形、气候等因素的影响，植被区系丰富，森林、湿地、草原及冰川等自然景观多样，为农牧业和旅游业的发展提供了良好的自然条件，但生态环境脆弱[26]。复杂的地形地貌及高原气候制约着区域的经济发展，同时也造成该地区人口分布不均衡。

3　研究方法与数据来源

3.1　影响因素选取

人口分布的影响因素具有复杂多变性，依据其属性分为自然地理环境和社会经济两大类[27]，但其构成要素各异，没有统一的标准。本研究假设任何构成自然地理环境和社会经济且作用于人口的要素皆能成为影响因素，根据前人相关研究及能够被空间量化的原则，总结了影响人口分布的七大因素：地形地貌、气候、植被覆盖、土地利用、河流、道路和经济发展。表1简要描述了这些影响因素，在影响机制上既有单向的因果关系（例如当人类改造客观世界的能力有限时，地形地貌成为了限制因素），也有双向的互

表1 影响人口分布的因素

Tab. 1 Influencing factors of population distribution

影响因素	影响因子	描述
地形地貌(A)	海拔(A_1) 坡度(A_2) 坡向(A_3) 地形起伏度(A_4)	地形地貌是影响人口分布的最基本因素之一：随着海拔高度的增加，地形起伏度呈现逐渐升高趋势，人口也随之减少[20]；在中国人口大多居住在坡度小于15°的地区[19]；而坡向是构成地形地貌因素的重要定量指标，通过调节各自然要素的分配影响人口分布，已有研究将其作为乡镇尺度人口分布的影响因子[28]。
气候(B)	多年平均降雨量(B_1) 多年平均气温(B_2)	人类对不同水热条件影响下的资源和环境条件具有选择偏好，在较广的范围内，人口密度与气温、降水量呈显著的正相关关系[29]。
植被覆盖(C)	NDVI(C)	以NDVI为代表的植被指数能反映植被覆盖情况，一方面显示了人类生产生活所依赖的植被资源分布，另一方面则展示了不适宜居住区的空间分布(如沙漠和密林)[30]。
土地利用(D)	土地利用指数(D_n)	根据“无土地则无人口”原则，人口分布受特定的土地利用类型影响，且土地利用面积与人口分布具有强相关性[22]。
河流(E)	河网密度(E_1) 距河流距离(E_2)	河流一方面为人们提供充足而稳定的水源，另一方面造就沿岸地势低平、土壤肥沃的格局，为人们居住生活、基础设施建设和发展生产提供适宜的空间[20]。
道路(F)	路网密度(F_1) 距道路距离(F_2)	基于交通条件的区域可达性与人口分布关系密切，该因素对欠发达地区人口集聚所起的作用远大于发达地区[15]。
经济发展(G)	夜间灯光指数(G_1) GDP(G_2) 工业总产值(G_3) 农林牧渔业产值(G_4)	夜间灯光亮度是城镇化的一个解释性指标，能反映城镇化水平，根据灯光影像的亮度值和人口距离衰减定律可以估算人口总数和人口分布[31-32]；也可以在一定程度上反映GDP，但两者的定量关系仍存在较大不确定性，并且夜间灯光指数反映的是消费而非生产[33]，不能完全替代经济统计数据；另外它还能反映区域的总体能源消耗量[34-35]。一般来说，区域人口分布与经济发展之间具有较强的一致性，产业结构的差异也会影响人口分布[36-37]。

为因果关系（例如经济发展程度能影响人口发展，反过来人口的发展状况也能影响经济发展）；虽然某些因素（如夜间灯光指数和GDP）存在共线性特征，但在影响人口分布方面体现了不同的侧重点，允许它们同时存在。通过运行相关模型筛选出主要的影响因子，进而厘清西藏人口分布的影响因素。

3.2 数据来源与预处理

按照原始数据的属性，本文的数据集可分为属性数据和空间数据。属性数据需要通过公共字段关联到空间数据中，空间数据则统一采用Albers等积投影坐标系，栅格空间分辨率为1 km。

人口统计数据来源于《中国2010年人口普查分乡、镇、街道资料》（西藏自治区）数据集，乡镇边界数据、河流水系与道路数据均来源于西藏自治区科技信息研究所。由于行政区域变迁、边界数字化误差等因素影响，乡镇边界和人口普查数据未能完全匹配，有必要对其进行核查与合并。调整一致后，利用ArcGIS将属性数据与对应的乡镇空间数据进行关联，得到624个行政单元，常住总人口为3002165人。需要说明的是，乡镇边界缺少7个县的数字化结果，以县级边界代替。

经济统计数据来源于《西藏统计年鉴》（2011）。为了匹配乡镇边界数据，进一步描述经济差异，将县级经济数据平均至对应的乡镇，得到乡镇尺度的农林牧渔业产值、工业总产值和GDP，并将其关联到对应的空间数据。

DEM数据来源于地理空间数据云网站（http://www.gscloud.cn/）SRTM数据集，空间分辨率为90 m；2010年夜间稳定灯光强度数据来源于NOAA网站（https://ngdc.noaa.gov/eog/dmsp/downloadV4composites.html），空间分辨率为30″。

气象数据来源于国家生态系统观测研究网络科技资源服务系统（http://rs.cern.ac.cn/data/）2000-2012年全国1 km空间分辨率气温和降水栅格数据集[38]。该数据集采用日降雨和日均温两个要素，对其进行读取、合并、检查、统计、空间插值批处理代码的生成等操作，最后由ANUSPLINE软件插值生成1 km空间分辨率的年均气温和年降雨栅格数据。本文选取2001-2010年气温和降雨栅格数据并对其求平均，得到多年平均气温和多年平均降雨量。

NDVI数据（2010全年）来源于NASA官网（https://ladsweb.nascom.nasa.gov）MOD13A3数据集，空间分辨率为1 km，覆盖西藏全境。该数据集采用MVC方法得到的逐月NDVI能有效去除云雾的影响，广泛应用于植被覆盖、生态监测等领域[39-40]。运行MRT软件对NDVI数据进行批量重投影、重采样、裁剪，再计算NDVI年平均值。

2010年土地覆盖数据来源于全球地表覆盖产品GlobeLand 30（http://www.globallandcover.com），空间分辨率为30 m，覆盖西藏全境。该数据集是全球首套30 m全球地表覆盖数据，全球总体分类精度达80%以上[41-43]，包括人造地表、耕地、林地等10大类型，在土地利用变化、人口密度估算、环境监测等领域拥有广阔的应用前景[44-46]。

3.3 研究方法

3.3.1 人口分布格局分析方法　首先计算研究区各乡镇的人口密度，叠加河流、道路等空间信息对其进行人口密度分级制图，并统计不同人口密度等级的人口总数与面积，分析人口分布的疏密特征；再利用ArcGIS计算人口密度的莫兰指数（Moran′s *I*），分析人口分布的集聚特征。

莫兰指数是一个衡量空间相关性（集聚）的重要指标，包括全局性莫兰指数和局部莫兰指数，详细计算过程参考已有研究成果[47-48]。Moran′s *I*取值在−1~1之间，＞0表示正自相关，即表示要素具有包含同样高或同样低的属性值的邻近要素；＜0表示负自相关，即要素具有包含不同值的邻近要素；等于0则表示属性值是随机分布的。通过统计学上的显著性检验（$P \leqslant 0.05$）的局部Moran′s *I*指数有4种输出模式：高值（H−H）聚类、低值（L−L）聚类、高值主要由低值围绕的异常值（H−L）以及低值主要由高值围绕的异常值（L−H）。全局莫兰指数用来判断是否存在空间集聚现象，而局部莫兰指数则可以探测集聚现象和异常值的空间分布。

3.3.2 影响因子提取　河流因子与道路因子。运行ArcGIS的Euclidean Distance工具，得到距河流距离与距道路距离；再根据公式（1）计算河网密度与路网密度。

$$RD = L/A \tag{1}$$

式中：*RD*是河网密度或路网密度（km/km^2）；*L*是河流或道路的总长度（km）；*A*为统计格网面积（km^2），由Fishnet工具生成。

地形地貌因子。利用ArcGIS提取研究区的海拔、坡度、坡向和地形起伏度。其中，采用4 km^{2}的普适性采样单元分别计算90 m空间分辨率DEM的最大值和最小值，再求两者的差值经重采样后得到地形起伏度[49]。

土地利用指数因子。研究表明土地利用数据的面积信息在人口分布方面有很强的相关关系，在人口模拟估算方面得到广泛利用[28, 50-51]。本文采用格网统计的方法计算土地利用指数：首先生成拥有唯一识别字段的1 km^{2}网格；再利用分区统计功能统计每一格网下30 m空间分辨率的各土地利用类型面积，并将其基于公共字段与网格进行属性连接；然后计算该土地利用类型面积占网格面积的比例；最后将网格数据转换成栅格数据。根据此方法，计算得到冰川和永久积雪指数（D_1）、草地指数（D_2）、耕地指数（D_3）、灌木指数（D_4）、裸地指数（D_5）、人造地表指数（D_6）、森林指数（D_7）、湿地指数（D_8）、水体

指数（D_9）和苔原指数（D_{10}）。

运用ArcGIS的Zonal Statistics工具对夜间灯光指数、土地利用指数等22个栅格影响因子进行分区统计，得到乡镇尺度下各影响因子的平均值。

3.3.3 影响因素分析方法 以人口密度为因变量、影响因子为自变量，对比运用多元线性回归方法与随机森林回归方法进行拟合验证，探索研究区人口分布的影响因素。

首先利用SPSS 20.0计算人口密度与各影响因子（共计25个）的相关关系，再选择与人口分布相关关系较强的因子（$r \geq 0.6$）作为自变量进行多元线性回归，采用标准化回归系数描述自变量对人口分布的影响程度。

随机森林（Random Forest）模型是由Breiman在2001年提出来的一种基于分类树的机器学习算法，该模型是利用bootsrap重抽样方法从原始样本中抽取多个样本，对每个bootsrap样本进行决策树建模，然后组合多棵决策树的预测，通过投票得出最终预测结果[52-53]。随机森林模型由于本身在算法上具有明显而独特的优势，可以用来做聚类、判别、回归和生存分析，同时可以评估变量的重要性。本文是在R语言平台上进行的随机森林回归，参数设置如下：ntree = 5000，mtry = 3，其他默认。

由于缺少更大比例尺的人口统计数据，在使用人口密度统计实测值来验证多元线性回归和随机森林回归的精度时，一般进行必要性检验[50, 54]，评价指标包括决定系数（R^2）和平均绝对误差（MAE）。

$$R = \frac{\sum_{i=1}^{n}\left(P_i - \bar{P}\right)\left(O_i - \bar{O}\right)}{\sqrt{\sum_{i=1}^{n}\left(P_i - \bar{P}\right)^2}\sqrt{\sum_{i=1}^{n}\left(O_i - \bar{O}\right)^2}} \tag{2}$$

$$MAE = \frac{\sum_{i=1}^{n}\left|P_i - O_i\right|}{n} \tag{3}$$

式中：O_i为第i个乡镇的统计实测人口密度；P_i为第i个乡镇的拟合人口密度；$\bar{P}$为所有乡镇拟合人口密度的平均值；$\bar{O}$为所有乡镇的统计实测人口密度的平均值。R^2值越大，MAE值越小，模型解释精度越高。

4 结果与分析

4.1 人口分布格局

4.1.1 人口分布总体特征 利用ArcGIS对西藏乡镇尺度的人口密度进行分级（图1）。总体而言，西藏平均人口密度为2人/km^2；人口分布呈现“东南—西北”模式，即东南部人口密度高于西北部；最高值出现在拉萨市的城关区街道（2928.8人/km^2），最低值出现在山南地区的玉麦乡，仅为0.009人/km^2。人口密度高值区主要分布在“一江两河”（雅鲁藏布江及其支流拉萨河与年楚河）的河谷平原区和怒江—澜沧江沿岸，大部分乡镇的人口密度介于10~50人/km^{2}之间，且多条国道经过该地区；人口密度低值区位于研究区中部以北的藏北高原以及藏东南的中印边境地区，区域内的大河分布稀疏，主要的交通干线较少。表2统计了不同人口密度分级的乡镇人口总数与面积，并计算了它们的比例结构。结果显示，不同密度等级的人口和面积比例严重失衡：人口密度小于1人/km^{2}的区域面积最大，比例高达63.82%，人口总量却只占西藏人口总数的7.86%；人口密度大于10人/km^{2}的区域仅占总面积的4.48%，人口总量却超过了45%。

4.1.2 人口分布空间自相关分析 将各乡镇的人口密度值作为输入变量进行全局性莫兰指数的计算，得到Moran′s I = 0.112（Z = 29.803，P = 0），说明西藏的人口分布存在集聚

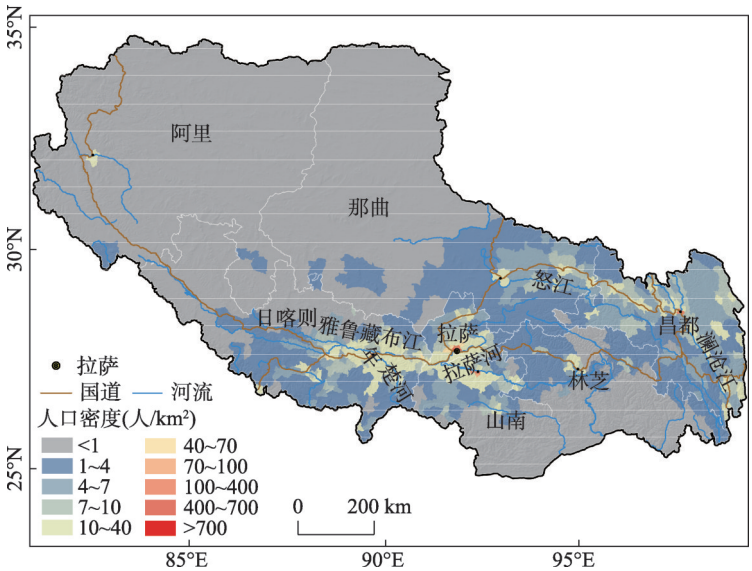


图1　2010年西藏乡镇尺度人口密度分布

Fig. 1　The population density at township level in Tibet in 2010

表2　西藏不同人口密度分级的乡镇人口总数与面积

Tab. 2　The total population and area of each population density range at township level in Tibet

人口密度分级(人/km²)	统计单元个数(个)	总人口(人)	人口占比(%)	总面积(km²)	面积占比(%)
0~1	102	235880	7.86	768471	63.82
1~4	189	550149	18.33	244664	20.32
4~7	115	482070	16.06	91516	7.60
7~10	75	378282	12.60	45631	3.79
10~40	126	870345	28.99	51935	4.31
40~70	7	43936	1.46	918	0.08
70~100	1	6621	0.22	74	0.01
100~400	6	142181	4.74	722	0.06
400~700	2	93542	3.12	172	0.01
>700	1	199159	6.63	68	0.01

现象，但集聚程度较低。为了检测人口集聚的空间分布状况，进行局部莫兰指数计算。初步的计算结果显示仅有极少数乡镇存在集聚现象，难以反映该地区的人口集聚状况，因此，有必要进行多次局部莫兰指数计算直至集聚状态稳定为止。本文进行了4次局部莫兰指数计算后结果趋于稳定（图2），需要指出的是，后一次局部莫兰指数计算的输入变量为前一次集聚结果中去除H-H模式后剩余的乡镇人口密度。

从图2可知，西藏大致以波绒乡（聂拉木县）—岗尼乡（安多县）为人口分界线，该线东南部人口密度较大，西北部人口密度较小。形成了拉萨—乃东一级核心区、日喀则二级核心区、山南—昌都三级核心区与日喀则—昌都四级核心区（H-H模式）：一级核心区主要位于拉萨市区与乃东县，是西藏人口分布最为密集的地区；二级核心区位于日喀则的雅鲁藏布江—年楚河沿岸地带，其人口密度为15.15~520.06人/km²，平均值为45.74人/km²，远大于全区的平均水平（2人/km²）且分布极不均衡；三级核心区主要位

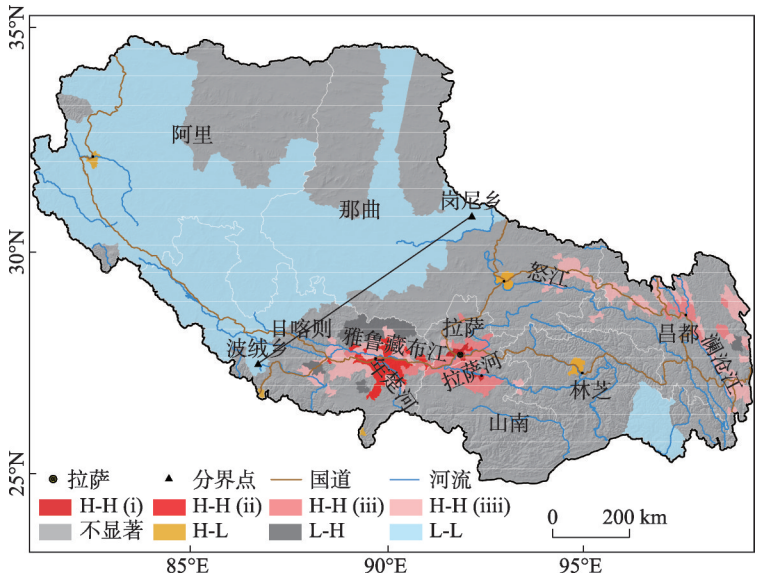


图2 研究区人口分布的空间集聚特征

Fig. 2 Spatial clustering characteristics of population distribution in the study area

于雅鲁藏布江—拉萨河沿岸以及澜沧江及其支流沿岸，平均人口密度为20.61人/km^2；四级核心区主要分布在三级核心区外围及317、318国道两侧，其人口密度最小值为6.54人/km^2。在人口密度低值区出现了5处次一级的人口集聚中心，即H–L模式，均有河流或国道经过。另外，在人口集聚核心区外围出现了低值中心，即L–H模式。

4.2 人口分布的影响因素分析

4.2.1 多元线性回归与随机森林回归的人口密度验证 由表3可知，西藏人口分布与各影响因子的相关系数差异较大，且通过显著性检验的影响因子较少，与人口密度显著相关的仅有8个（与人口总数显著相关的有12个）。计算表3中与人口密度和与人口总数均通过显著性检验的相关系数平均值，可以发现：除海拔与人口分布呈负相关关系外，其他因子均呈正相关关系；人造地表指数、夜间灯光指数、路网密度与人口分布具有很强的相关性，均超过0.6；河网密度、工业总产值、GDP、多年平均气温对人口分布的影响较小。在相关性分析的基础之上，选择通过显著性检验且相关系数$R \geqslant 0.6$的因子作为自变量进行多元线性回归，其拟合公式为：

$$y = 0.778D_6 + 0.142G_1 + 0.081F_1, \quad (n = 624, R^2 = 0.69, P < 0.001) \tag{4}$$

式中：由于采用标准化回归系数，没有常数项；y为拟合人口密度；D_6、G_1、F_1分别为人造地表指数、夜间灯光指数和路网密度。式（4）显示，人造地表指数、夜间灯光指数和路网密度与西藏地区的人口密度均呈正相关关系。人造地表指数的标准化回归系数最大，说明其对人口分布的影响程度最大，随后依次为夜间灯光指数和路网密度。

将全部影响因子作为自变量，对人口密度进行随机森林回归；分别计算经过两种回归方法得到的拟合值与统计实测人口密度值的R^2与MAE（图3）。由图3可知，多元线性回归的R^2为0.90，比随机森林回归小（R^2= 0.98）；多元线性回归的MAE大于随机森林方法，说明随机森林的预测值偏离误差小于多元线性回归；另外，多元线性回归的人口密度预测值中存在1/3的负值，与人口密度的实际意义相悖，而随机森林的拟合值中并未发现负值。因此，随机森林的人口密度预测较多元线性回归拥有更大的R^2，更小的偏离误差且具有实际意义，预测精度更高，可以用随机森林方法来分析人口分布的影响因素。

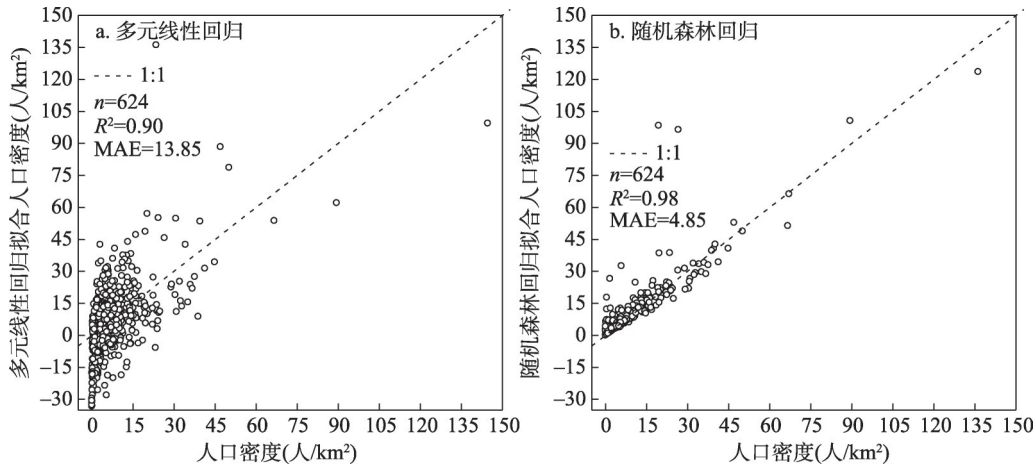
表3　人口分布与各影响因子的相关系数

Tab. 3　The correlation coefficient between population distribution with each influencing factor

影响因子	序号	与人口密度相关系数	与人口总数相关系数	相关系数平均值
海拔	A_1	−0.137**	−0.172**	−0.155
坡度	A_2	−0.058	−0.058	−
坡向	A_3	0.013	0.016	−
地形起伏度	A_4	−0.043	−0.058	−
多年平均降雨量	B_1	−0.057	−0.02	−
多年平均气温	B_2	0.171**	0.197**	0.184
NDVI	C	−0.001	0.047	−
冰川和永久积雪指数	D_1	−0.044	−0.047	−
草地指数	D_2	−0.043	−0.008	−
耕地指数	D_3	0.078	0.088*	−
灌木指数	D_4	0.067	0.049	−
裸地指数	D_5	−0.05	−0.100*	−
人造地表指数	D_6	0.942**	0.810**	0.876
森林指数	D_7	−0.025	−0.009	−
湿地指数	D_8	−0.015	−0.02	−
水体指数	D_9	−0.023	−0.021	−
苔原指数	D_{10}	−0.005	0.072	−
河网密度	E_1	0.498**	0.437**	0.468
距河流距离	E_2	−0.045	−0.071	−
路网密度	F_1	0.664**	0.630**	0.647
距道路距离	F_2	−0.048	−0.089*	−
夜间灯光指数	G_1	0.902**	0.792**	0.847
GDP	G_2	0.398**	0.490**	0.444
工业总产值	G_3	0.453**	0.452**	0.453
农林牧渔业产值	G_4	0.036	0.327**	−

注：*在0.05水平上显著相关，**在0.01水平上显著相关。

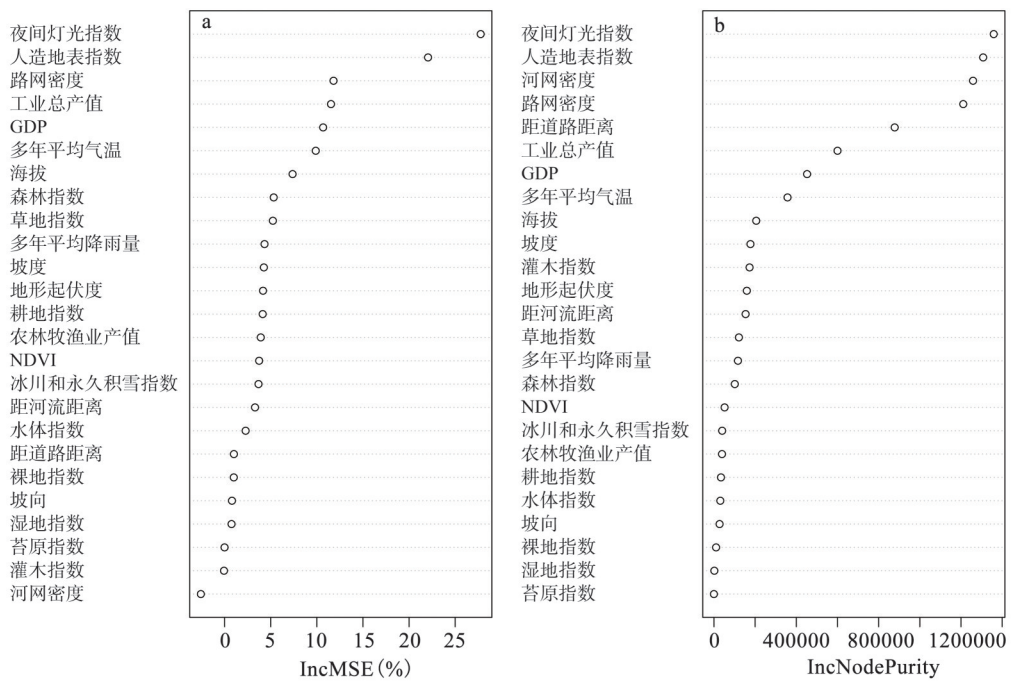
4.2.2 影响因素的重要性排序　图4显示了随机森林方法下影响因子的重要性排序情况，重要性排序靠前的因子与显著相关系数、多元线性标准化回归系数的排序趋于一致，但略有不同。以IncMSE方法为例，分析前6个影响因素对人口密度分布的作用（影响强度）。由图5可知，除多年平均气温（图5f）对人口密度的影响呈先升后降的趋势外，夜间灯光指数（图5a）、人造地表指数（图5b）、路网密度（图5c）、工业总产值（图5d）、GDP（图5e）对人口密度的影响均呈阶梯状上升的趋势：当夜间灯光指数≥ 42、人造地表指数≥ 0.38时，二者对人口密度的影响达到最大并保持不变；当路网密度在1.5~1.6 km/km^{2}时，对人口密度分布的影响程度急剧增加，大于1.6 km/km^{2}时，影响程度达到最大值并保持平稳不变；人口密度受GDP和工业总产值影响较大，且呈正相关关系，拉萨城关区所辖乡镇的GDP和工业总产值在西藏全区位居前列，其人口密度也最高正说明了这一点；对于气温条件（图5f），人口主要分布在年平均气温在7 ℃以上的地区，随着气温的升高，人口密度也随之增加，并在7.9 ℃时达到峰值。青藏高原地区的人口分布受地形限制较大，不可避免地要讨论该影响因素（海拔与坡度，地形地貌因素中IncMSE重要



注：为了展现更多的细节，只截取150人/ km²以下的区域进行显示。

图3 多元线性回归(a)与随机森林回归(b)人口密度验证散点图

Fig. 3 Scatter plot of population density verification in Multiple Linear Regression (a) and Random Forest Regression (b)



注：IncMSE为精度平均减少值，指将变量随机取值后进行随机森林模型估算误差相对于原来误差的升高幅度；IncMSE值越大，说明该变量越重要。IncNodePurity为节点不纯度平均减少值，是指该变量对各个决策树节点的影响程度；IncNodePurity值越大，说明该变量越重要。

图4 影响因子重要性排序

Fig. 4 The importance ranking of influencing factors

性排序前两位）对人口密度分布的作用。由图可知，海拔（图5g）和坡度（图5h）均与人口密度分布呈负相关关系，影响程度呈三级阶梯状下降的趋势：当海拔在2100~3800 m、坡度小于8°时，对人口密度分布的影响达到最大并保持不变；当海拔在3800~4100 m、

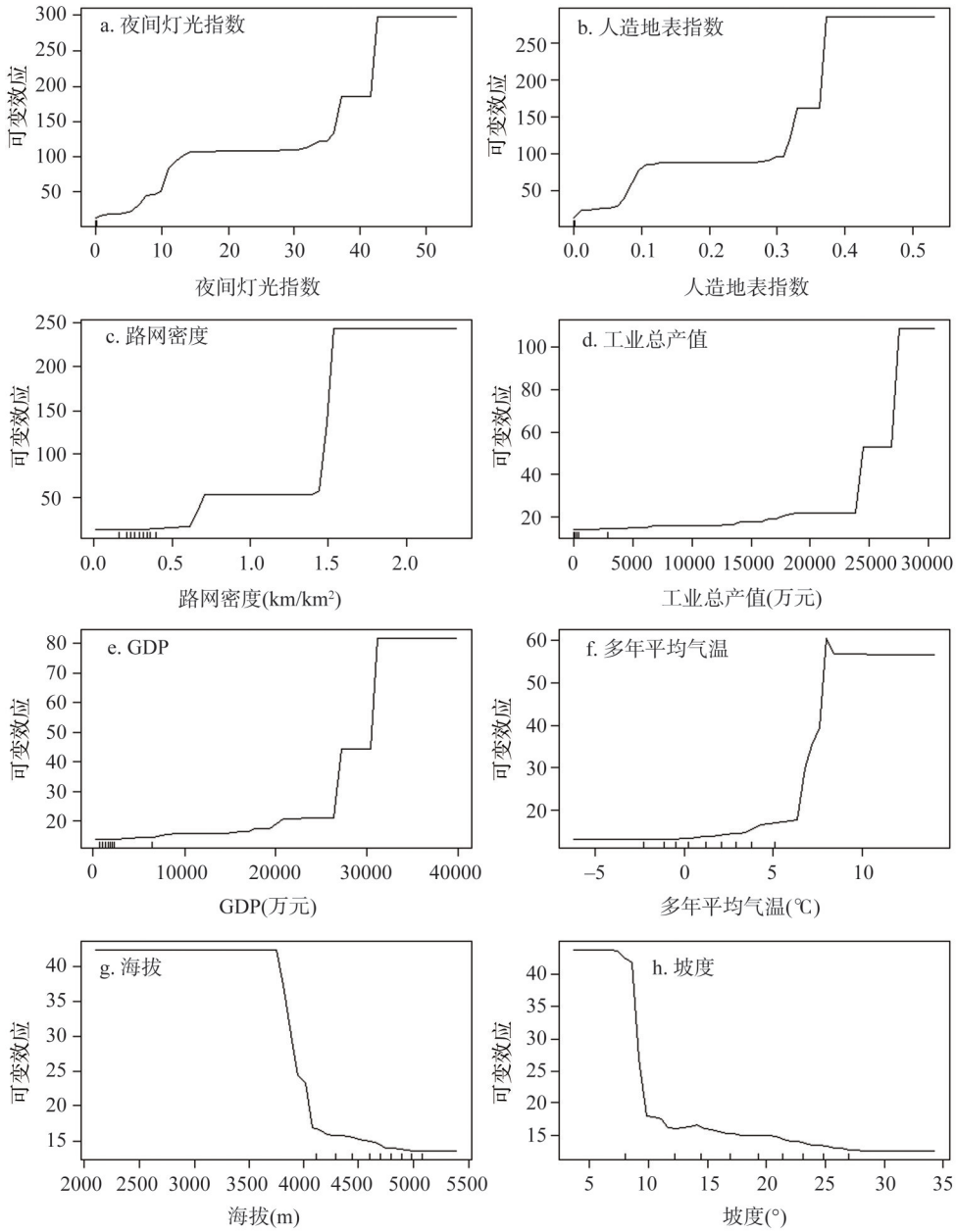


图5　夜间灯光指数(a)、人造地表指数(b)、路网密度(c)、工业总产值(d)、GDP(e)、多年平均气温(f)、海拔(g)、坡度(h)对人口密度的影响强度

Fig. 5　Impact of the night light index (a), artificial surface index (b), road density (c), industrial output (d), GDP (e), annual average temperature (f), elevation (g) and slope (h) on population density

坡度在8°~10°时，对人口密度分布的影响程度急剧下降；当海拔大于5100 m时、坡度大于28°时，对人口密度分布的影响程度达到最小并保持不变。

4.2.3 影响因素的区域差异分析　为了探索人口密度影响因素的区域差异，对各地级市的因子进行随机森林回归拟合与验证，并列出重要性排序在前6位的影响因子（表4）。表4

表4 西藏各地区人口密度拟合R^2、MAE与影响因子排序

Tab. 4 Population density fitting R^2, MAE and their rankings of influencing factors among different regions in Tibet

地区	R^2	MAE	排序1	排序2	排序3	排序4	排序5	排序6
拉萨市	0.98	50.84	G_1	D_6	B_2	G_3	A_1	D_2
昌都地区	0.96	1.94	A_1	D_6	G_1	B_2	G_2	D_3
山南地区	0.95	3.36	D_3	B_2	A_1	D_6	G_1	D_1
日喀则地区	0.97	3.53	D_6	G_1	D_3	F_1	B_2	A_1
那曲地区	0.92	0.60	C	A_1	B_1	B_2	F_1	D_6
阿里地区	0.98	0.30	D_6	G_1	D_3	B_2	F_1	D_1
林芝地区	0.87	1.16	G_1	D_6	F_1	F_2	D_8	A_2

注：日喀则、昌都、林芝、山南先后于2014-2016年完成撤地设市。

显示，各地级市人口密度拟合值的R^2均大于0.85且偏离误差较小，说明该方法的预测精度较高，可以进行影响因子的重要性排序。夜间灯光指数和人造地表指数是影响拉萨、日喀则、林芝、阿里人口分布的主导因子，说明在这些地区，较高的区域发展水平对人口具有强大的吸引力，但在其他因素组成上略有差异：作为省会城市的拉萨，工业、农业的发展和基础设施的完善对人口集聚形成绝对引力；日喀则东部地区河谷农业发展迅速，人口规模较大；林芝地区旅游资源丰富且交通设施较为发达，对人口的吸引力较强；阿里地区平均海拔为西藏最高，气候寒冷干旱，人口分布极其稀疏。昌都地区贡献率第一的影响因素是海拔，其余因子与拉萨较为一致；地貌虽是山高谷深、起伏悬殊、藏东三江（金沙江、澜沧江、怒江）纵贯，但三江流域山间平原发育，农业气候资源丰富、雨热同季，又是“茶马古道”的一个重要枢纽，紧密勾连着西藏腹地与外界的交流，使之形成人口分布次级核心区。在山南地区，耕地指数和多年平均气温对于人口密度分布的影响显著高于其他因子；该区北部地处谷地，两河交汇处热量水平较高，为农业发展提供了良好条件；区域内耕地指数较高，乃东县、扎囊县、贡嘎县均是西藏非常重要的粮食生产基地，这就是山南地区北部成为人口分布核心区的主要原因。在那曲地区，自然因素对人口密度分布的影响程度显著高于社会经济因素，植被覆盖指数（NDVI）对人口密度分布的影响程度最高，其次为海拔；由于该区地处藏北高原，海拔高、热量不足、气候严寒干旱，人口分布较少（仅占15%）；高原草地覆盖面积广大，丰富的草地资源为农牧民提供了集聚条件，该地区的农牧业总产值占GDP的90%以上说明了这一点。

综合来看，河谷是西藏地区人口的集聚区，主要分布在拉萨河谷、年楚河谷以及三江河谷。土地是人口分布的载体，地势低平、相对开阔的河谷空间为人类活动的开展提供了最基础的土地条件。温和且雨热同期的气候特征以及邻近水源地使得区域内植被生长条件适宜，取用水方便且易于灌溉，农业发展条件较好，因此孕育了历史悠久的河谷农业，宜耕宜牧区人口集聚明显。交通网络体系的完善提高了区域交流沟通水平，城镇化进程的加快促使河谷区农业人口转化为非农业人口，较高的经济发展水平能进一步吸引人口的集聚，实现人口分布的再分配。

5 结论与讨论

5.1 结论

本文基于2010年西藏自治区的乡镇尺度人口普查数据，深化空间统计方法并引入了

随机森林方法，精细而客观地刻画了该地区的人口分布格局，进一步厘清了人口分布影响因素及其区域差异。主要结论为：

（1）西藏乡镇人口密度在空间上表现出极强的非均衡性，其总体趋势是东南高西北低；高值区主要分布在“一江两河”的河谷平原区和怒江—澜沧江沿岸，低值区位于藏北高原以及藏东南的中印边境地区；大江大河与主要交通干线是人口分布的主要轴线。

（2）西藏的人口分布存在集聚现象，但集聚程度较低；大致以波绒乡（聂拉木县）—岗尼乡（安多县）为人口分界线，该线东南部人口密度较大，西北部人口密度较小；人口集聚的“核心—边缘”特征明显，形成了拉萨—乃东一级核心区、日喀则二级核心区、山南—昌都三级核心区与日喀则—昌都四级核心区，在人口集聚核心区外围出现了低值中心，在人口密度低值区出现了五处次一级的集聚中心。

（3）研究区人口分布与各影响因子的相关系数差异较大，通过显著性检验的影响因子较少；多元线性回归方法中，人造地表指数对人口分布的影响程度最大，随后依次为夜间灯光指数和路网密度。随机森林的人口密度预测较多元线性回归拥有更大的R^2，更小的偏离误差且具有实际意义，预测精度更高，可以用来进行影响因子的重要性排序。重要性排序靠前的因子与显著相关系数、多元线性标准化回归系数的排序趋于一致，但略有不同，依次为夜间灯光指数、人造地表指数、路网密度、工业总产值、GDP、多年平均气温，这些因素与人口密度均呈正相关关系；地形地貌要素中以海拔和坡度的贡献率最大且与人口密度均呈负相关关系。

（4）西藏人口分布格局的影响因素及其相互作用呈现出明显的区域差异特征：夜间灯光指数和人造地表指数是影响拉萨、日喀则、林芝、阿里人口分布的主导因子，但在其他因素组成上略有差异；昌都地区贡献率第一的影响因素是海拔，其余因子与拉萨较为一致；山南地区的耕地指数和多年平均气温对人口密度分布的影响显著高于其他因子；那曲地区的NDVI对人口密度分布的影响程度最高，其次为海拔；综合来看，河谷是西藏地区人口的集聚区，主要分布在拉萨河谷、年楚河谷以及三江河谷。

（5）通过随机森林回归分析，可以利用概念模型来表达人口分布的影响因素，将主导因素概括为土地利用结构、道路通达度和城镇化水平，且三者相互影响、相互作用（图6）。区域内的地形、气候与植被组合条件不同引起土地利用结构各异，人口首先会在自然资源禀赋具有优势的地区产生集聚；道路通达度影响区域间的沟通交流水平，使包括人口在内的各资源要素产生流动，主要交通干道成为人口分布轴线；以城镇化水平（夜间稳定灯光强度）为代表的经济发展强度在资源禀赋和沟通交流水平基础上对人口分布进行再分配，形成“核心—边缘”效应。

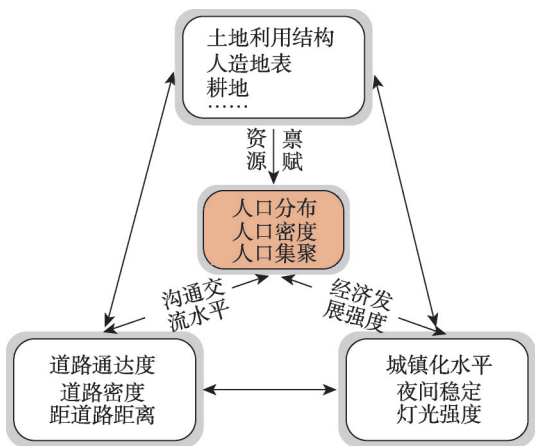


图6　人口分布概念模型

Fig. 6　Conceptual model of population distribution

本文通过多元线性回归与随机森林回归这两种方法对人口密度的拟合验证指标来评估它们对影响因素的解释精度，结果表明随机森林回归方法优于多元线性回归方法。然而，由于自然地理要素和社会经济要素对人口分布的影响是复杂非线性

的，目前关于人口分布影响因素的研究还处于深入探索阶段，统计模型仍然是其主要研究方法[21, 55-57]。由于统计原理的差异，不同的统计模型对影响因素的解释存在差异，如本研究中夜间灯光指数、人造地表指数、工业总产值、GDP在不同统计方法中具有不同的重要性排序。明确并揭示这种差异的本质是非常重要的，但从目前已有的研究看，这是相当复杂的，甚至存在极大的不确定性[55, 58]。另外，对于可验证的自然科学研究，如增温效应估算、植被生物量模拟、土壤养分空间分布等领域[59-61]，拟合预测的精度均可得到较满意的控制，从而使影响因素的解释精度较高；而对于以人类活动为主的社会科学研究领域，由于数据定量化的精度限制，目前对影响因素的解释仍处于“灰色”阶段，需要在以后的研究中深入探讨。

未来将在以下几个方面加强人口分布研究：① 对现有的人口普查乡镇数据进行系统处理，考虑从人口性质、人口迁移等不同侧面进行人口分布格局演变及其影响因素分析，并探讨不同归因方法对影响因素解释精度的差异；② 利用随机森林在分类、变量重要性排序方面的优越性，结合人口分布概念模型探索空间水平上的人口发展阶段及其类型；③ 基于人口分布概念模型，利用多种统计模型（地理加权回归、随机森林、神经网络等）进行人口空间定量化探索。

参考文献(References)

[1] Jiang Dong, Yang Xiaohuan, Wang Naibin, et al. Study on spatial distribution of population based on remote sensing and GIS. Advances 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.]

[2] Ferretti V, Pomarico S. Ecological land suitability analysis through spatial indicators: An application of the Analytic Network Process technique and Ordered Weighted Average approach. Ecological Indicators, 2013, 34: 507-519.

[3] Kleidon A. Climatic constraints on maximum levels of human metabolic activity and their relation to human evolution and global change. Climatic Change, 2009, 95(3/4): 405-431.

[4] Yang Jun, Guan Xin, Li Xiangyun, et al. Study on the relations between the LUCC and demographic factors in the past 10 years of Tarim River Basin. Journal of Arid Land Resources and Environment, 2006, 2(2): 114-117. [杨君, 关欣, 李香云, 等. 近10年塔里木河流域土地利用/土地覆被变化与人口因素关系研究. 干旱区资源与环境, 2006, 2(2): 114-117.]

[5] Feng Zhiming, Yang Yanzhao, You Zhen, et al. Research on the suitability of population distribution at the county level in China. Acta Geographica Sinica, 2014, 59(6): 723-737. [封志明, 杨艳昭, 游珍, 等. 基于分县尺度的中国人口分布适宜度研究. 地理学报, 2014, 59(6): 723-737.]

[6] Li Jiaming, Lu Dadao, Xu Chengdong, et al. Spatial heterogeneity and its changes of population on the two sides of Hu Line. Acta Geographica Sinica, 2017, 72(1): 148-160. [李佳洺, 陆大道, 徐成东, 等. 胡焕庸线两侧人口的空间分异性及其变化. 地理学报, 2017, 72(1): 148-160.]

[7] Mo Shenguo, Zhang Baiping, Cheng Weiming, et al. Major environmental effects of the Tibetan Plateau. Progress in Geography, 2004, 23(2): 88-96. [莫申国, 张百平, 程维明, 等. 青藏高原的主要环境效应. 地理科学进展, 2004, 23(2): 88-96.]

[8] Cai Yunlong, Song Changqing, Leng Shuying. Future development trends and priority areas of physieal geography in China. Scientia Geographica Sinica, 2009, 29(5): 619-626 [蔡运龙, 宋长青, 冷疏影. 中国自然地理学的发展趋势与优先领域. 地理科学, 2009, 29(5): 619-626.]

[9] Luo Lifang, Zhang Keli, Kong Yaping, et al. Temporal and spatial distribution of soil loss on Tibet-Qing Plateau. Journal of Soil and Water Conservation, 2004, 18(1): 58-62. [罗利芳, 张科利, 孔亚平, 等. 青藏高原地区水土流失时空分异特征. 水土保持学报, 2004, 18(1): 58-62.]

[10] Niu Yafei. The study of environment in the Plateau of Qing-Tibet. Progress in Geography, 1999, 18(2): 69-77. [牛亚菲. 青藏高原生态环境问题研究. 地理科学进展, 1999, 18(2): 69-77.]

[11] Sun Honglie, Zheng Du, Yao Tandong, et al. Protection and construction of the national ecological security shelter zone on Tibetan Plateau. Acta Geographica Sinica, 2012, 67(1): 3-12. [孙鸿烈, 郑度, 姚檀栋, 等. 青藏高原国家生态安全屏

障保护与建设. 地理学报, 2012, 67(1): 3-12.]

[12] Gao Zhiqiang, Liu Jiyuan, Zhuang Dafang. The relations analysis between ecological environmental quality of Chinese land resources and population. Journal of Remote Sensing, 1999, 3(1): 67-71. [高志强, 刘纪远, 庄大方. 基于遥感和GIS的中国土地资源生态环境质量同人口分布的关系研究. 遥感学报, 1999, 3(1): 67-71.]

[13] Liao Shunbao, Sun Jiulin. Quantitative analysis of relationship between population distribution and environmental factors in Qinghai-Tibet Plateau. China Population, Resources and Environment, 2003, 13(3): 65-70. [廖顺宝, 孙九林. 青藏高原人口分布与环境关系的定量研究. 中国人口·资源与环境, 2003, 13(3): 65-70.]

[14] Qin Xiaojing. The spatial-temporal patterns and coupling relationship of animals husbandry and rural laborers across Tibet [D]. Nanchong: China West Normal University, 2016. [秦小静. 西藏自治区畜牧业与乡村人口的时空格局及耦合关系[D]. 南充: 西华师范大学, 2016.]

[15] Wang Zhenbo, Xu Jiangang, Zhu Chuangeng, et al. The county accessibility divisions in China and its correlation with population distribution. Acta Geographica Sinica, 2010, 65(4): 416-426. [王振波, 徐建刚, 朱传耿, 等. 中国县域可达性区域划分及其与人口分布的关系. 地理学报, 2010, 65(4): 416-426.]

[16] Liao Shunbao, Li Zehui. Study on spatialization of population census data based on relationship between population distribution and land use: Taking Tibet as an example. Journal of Natural Resources, 2003, 18(6): 659-665. [廖顺宝, 李泽辉. 基于人口分布与土地利用关系的人口数据空间化研究: 以西藏自治区为例. 自然资源学报, 2003, 18(6): 659-665.]

[17] Liao Shunbao, Sun Jiulin. GIS based spatialization of population census data in Qinghai- Tibet Plateau. Acta Geographica Sinica, 2003, 58(1): 25-33. [廖顺宝, 孙九林. 基于GIS的青藏高原人口统计数据空间化. 地理学报, 2003, 58(1): 25-33.]

[18] Luo Yong. Study on distribution of population and the sustainable development of ecological environment in Tibet [D]. Chengdu: Chengdu University of Technology, 2010. [罗永. 西藏人口分布与生态环境可持续发展研究[D]. 成都: 成都理工大学, 2010.]

[19] Feng Zhiming, Tang Yan, Yang Yanzhao, et al. The relief degree of land surface in China and its correlation with population distribution. Acta Geographica Sinica, 2007, 62(10): 1073-1082. [封志明, 唐焰, 杨艳昭, 等. 中国地形起伏度及其与人口分布的相关性. 地理学报, 2007, 62(10): 1073-1082.]

[20] Zhao Tongtong, Song Bangguo, Chen Yuansheng, et al. Analysis of population distribution and its spatial relationship with terrain elements in the Yarlung Zangbo River, Nyangqu River and Lhasa River Region, Tibet. Journal of Geo-Information Science, 2017, 19(2): 225-237. [赵彤彤, 宋邦国, 陈远生, 等. 西藏一江两河地区人口分布与地形要素关系分析. 地球信息科学学报, 2017, 19(2): 225-237.]

[21] Bai Zhongqiang, Wang Juanle, Yang Yaping, et al. Characterizing spatial patterns of population distribution at township level across the 25 provinces in China. Acta Geographica Sinica, 2015, 70(8): 1229-1242. [柏中强, 王卷乐, 杨雅萍, 等. 基于乡镇尺度的中国25省区人口分布特征及影响因素. 地理学报, 2015, 70(8): 1229-1242.]

[22] Dong Nan, Yang Xiaohuan, Cai Hongyan. Research progress and perspective on the spatialization of population data. Journal of Geo-Information Science, 2016, 18(10): 1295-1304. [董南, 杨小唤, 蔡红艳. 人口数据空间化研究进展. 地球信息科学学报, 2016, 18(10): 1295-1304.]

[23] Song Guobao, Li Zhenghai, Bao Yajing, et al. Spatial distribution pattern of population density in vertical ridge valley region and its influencing factors. Chinese Science Bulletin, 2007, 52(Suppl.2): 78-85. [宋国宝, 李政海, 鲍雅静, 等. 纵向岭谷区人口密度的空间分布规律及其影响因素. 科学通报, 2007, 52(Suppl.2): 78-85.]

[24] Tang Jiayun. The research of population distribution pattern of space and time in Gansu province and its influencing factors [D]. Wuhan: Central China Normal University, 2015. [唐嘉韵. 甘肃省人口分布的时空格局及其影响因素研究[D]. 武汉: 华中师范大学, 2015.]

[25] Zhang Lei, Wang Linlin, Zhang Xudong, et al. The basic principle of random forest and its applications in ecology: A case study of Pinus yunnanensis. Acta Ecologica Sinica, 2014, 34(3): 650-659. [张雷, 王琳琳, 张旭东, 等. 随机森林算法基本思想及其在生态学中的应用: 以云南松分布模拟为例. 生态学报, 2014, 34(3): 650-659.]

[26] Yu Bohua, Lv Changhe. Assessment of ecological vulnerability on the Tibetan Plateau. Geographical Research, 2011, 30 (12): 2289-2295. [于伯华, 吕昌河. 青藏高原高寒区生态脆弱性评价. 地理研究, 2011, 30(12): 2289-2295.]

[27] Hu Huanyong. Distribution of China's Population. Shanghai: East China Normal University Press, 1983. [胡焕庸. 论中国人口之分布. 上海: 华东师范大学出版社, 1983.]

[28] Wang Lei, Cai Yunlong. Spatial down- scaling analysis and simulation of population density in Maotiaohe Basin,

Guizhou Province. Progress in Geography, 2011, 30(5): 635-640. [王磊, 蔡运龙. 人口密度的空间降尺度分析与模拟: 以贵州猫跳河流域为例. 地理科学进展, 2011, 30(5): 635-640.]

[29] Fang Yu, Ouyang Zhiyun, Zheng Hua, et al. Natural forming causes of China population distribution. Chinese Journal of Applied Ecology, 2012, 23(12): 3488-3495. [方瑜, 欧阳志云, 郑华, 等. 中国人口分布的自然成因. 应用生态学报, 2012, 23(12): 3488-3495.]

[30] Zhuo L, Ichinose T, Zheng J, et al. Modelling the population density of China at the pixel level based on DMSP/OLS non-radiance-calibrated night-time light images. International Journal of Remote Sensing, 2009, 30(4): 1003-1018.

[31] Sutton P, Roberts C, Elvidge C, et al. A comparison of nighttime satellite imagery and population density for the continental united states. Photogrammetric Engineering and Remote Sensing, 1997, 63(11): 1303-1313.

[32] Ma T, Zhou C H, Pei T, et al. Quantitative estimation of urbanization dynamics using time series of DMSP/OLS nighttime light data: A comparative case study from China's cities. Remote Sensing of Environment, 2012, 124(9): 99-107.

[33] Wu J S, Wang Z, Li W F, et al. Exploring factors affecting the relationship between light consumption and GDP based on DMSP/OLS nighttime satellite imagery. Remote Sensing of Environment, 2013, 134(7): 111-119.

[34] Xie Y H, Weng Q H. World energy consumption pattern as revealed by DMSP-OLS nighttime light imagery. Giscience and Remote Sensing, 2016, 53(2): 265-282.

[35] He C Y, Ma Q, Li T, et al. Spatiotemporal dynamics of electric power consumption in Chinese Mainland from 1995 to 2008 modeled using DMSP/OLS stable nighttime lights data. Journal of Geographical Sciences, 2012, 22(1): 125-136.

[36] Jia Zhanhua, Gu Guofeng. Temporal- Spatial evolution characteristics and its influence factors about population distribution in Northeast China. Economic Geography, 2016, 36(12): 60-68. [贾占华, 谷国锋. 东北地区人口分布的时空演变特征及影响因素. 经济地理, 2016, 36(12): 60-68.]

[37] Feng Zhiming, Liu Xiaona. Multi-scale studies on the space consistency between population distribution and economic development in China. Population and Economics, 2013(2): 3-11. [封志明, 刘晓娜. 中国人口分布与经济发展空间一致性研究. 人口与经济, 2013(2): 3-11.]

[38] Wang Junbang, Ye Hui, Wang Juwu, et al. 1 km spatial resolution of temperature and precipitation in China in 2000-2012 raster data set. Science Data Bank, 2016. [王军邦, 叶辉, 王居午, 等. 2000-2012年全国气温和降水1 km网格空间插值数据集 . Science Data Bank, 2016.]

[39] Nestola E, Calfapietra C, Emmerton C A, et al. Monitoring grassland seasonal carbon dynamics, by integrating MODIS NDVI, proximal optical sampling, and eddy covariance measurements. Remote Sensing, 2016, 8(3): 260-285.

[40] Li Z, Huffman T, Mcconkey B, et al. Monitoring and modeling spatial and temporal patterns of grassland dynamics using time-series MODIS NDVI with climate and stocking data. Remote Sensing of Environment, 2013, 138(11): 232-244.

[41] Jun C, Ban Y F, Li S N. Open access to Earth land-cover map. Nature, 2014, 514(7523): 434-434.

[42] Brovelli M A, Molinari M E, Hussein E, et al. The first comprehensive accuracy assessment of GlobeLand30 at a national level: Methodology and results. Remote Sensing, 2015, 7(4): 4191-4212.

[43] Chen J, Chen J, Liao A P, et al. Global land cover mapping at 30 m resolution: A POK-based operational approach. Isprs Journal of Photogrammetry and Remote Sensing, 2015, 103: 7-27.

[44] Kuang W H, Chen L J, Liu J Y, et al. Remote sensing-based artificial surface cover classification in Asia and spatial pattern analysis. Science China-Earth Sciences, 2016, 59(9): 1720-1737.

[45] Lu Nan, Zhang Weiwei, Chen Lijun, et al. Estimation of large regional urban and rural population density based on the differences of population distribution between urban and rural: Take Shandong Province as example. Geodaetica et Cartographica Sinica, 2015, 44(12): 1384-1391. [鲁楠, 张委伟, 陈利军, 等. 顾及城乡差异的大区域人口密度估算: 以山东省为例. 测绘学报, 2015,44(12): 1384-1391.]

[46] Xie H, Du L, Liu S C, et al. Dynamic monitoring of agricultural fires in China from 2010 to 2014 using MODIS and GlobeLand30 data. ISPRS International Journal of Geo-Information, 2016, 5(10): 172.

[47] Chen Peiyang, Zhu Xigang. Regional inequalities in China at different scales. Acta Geographica Sinica, 2012, 67(8): 1085-1097. [陈培阳, 朱喜钢. 基于不同尺度的中国区域经济差异. 地理学报, 2012, 67(8): 1085-1097.]

[48] Ma Xiaodong, Ma Ronghua, Xu Jiangang. Spatial structure of cities and towns with ESDA- GIS framework. Acta Geographica Sinica, 2004, 59(6): 1048-1057. [马晓冬, 马荣华, 徐建刚. 基于ESDA-GIS的城镇群体空间结构. 地理学报, 2004, 59(6): 1048-1057.]

[49] Cheng Weiming, Zhou Chenghu, Chai Huixia, et al. Quantitative extraction and analysis of basic morphological types of land morphology in China. Journal of Geo-information Science, 2009, 11(6): 725-736. [程维明, 周成虎, 柴慧霞, 等. 中国陆地地貌基本形态类型定量提取与分析. 地球信息科学学报, 2009, 11(6): 725-736.]

[50] Wang Kejing, Cai Hongyan, Yang Xiaohuan. Multiple scale spatialization of demographic data with multi-factor linear regression and geographically weighted regression models. Progress in Geography, 2016, 35(12): 1494-1505. [王珂靖, 蔡红艳, 杨小唤. 多元统计回归及地理加权回归方法在多尺度人口空间化研究中的应用. 地理科学进展, 2016, 35 (12): 1494-1505.]

[51] 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.]

[52] Breiman L. Random forests. Machine Learning, 2001, 45(1): 5-32.

[53] Fang Kuangnan, Wu Jianbin, Zhu Jianping, et al. A review of technologies on random forests. Statistics and Information Forum, 2011, 26(3): 32-38. [方匡南，吴见彬，朱建平，等.随机森林方法研究综述.统计与信息论坛，2011, 26(3): 32-38.]

[54] 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.

[55] Wang L, Feng Z M, Yang Y Z. The change in population density from 2000 to 2010 and its influencing factors in China at the county scale. Journal of Geographical Sciences, 2015, 25(4): 485-496.

[56] Liu Jinsong. A review of population geography research in China. Acta Geographica Sinica, 2014, 69(8): 1177-1189. [刘劲松. 中国人口地理研究进展. 地理学报, 2014, 69(8): 1177-1189.]

[57] Shi X L, Wang W, Shi W J. Progress on quantitative assessment of the impacts of climate change and human activities on cropland change. Journal of Geographical Sciences, 2016, 26(3): 339-354.

[58] Yao Yonghui, Zhang Baiping. MODIS-based estimation of air temperature and heating-up effect of the Tibetan Plateau. Acta Geographica Sinica, 2013, 68(1): 95-107. [姚永慧, 张百平. 基于MODIS数据的青藏高原气温与增温效应估算. 地理学报, 2013, 68(1): 95-107.]

[59] Zhou Qianqian, Ding Jianli, Tang Mengying, et al. Inversion of soil organic matter content in oasis typical of arid area and its influencing factors. Acta Pedologica Sinica, 2018, 55(2): 313-324. [周倩倩, 丁建丽, 唐梦迎, 等. 干旱区典型绿洲土壤有机质的反演及影响因素研究. 土壤学报, 2018, 55(2): 313-324.]

[60] Liu Li, Han Mei, Liu Yubin, et al. Spatial distribution of wetland vegetation biomass and its influencing factors in the Yellow River Delta Nature Reserve. Acta Ecologica Sinica, 2017, 37(13): 4346-4355. [刘莉, 韩美, 刘玉斌, 等. 黄河三角洲自然保护区湿地植被生物量空间分布及其影响因素. 生态学报, 2017, 37(13): 4346-4355.]

Population distribution pattern and influencing factors in Tibet based on random forest model

WANG Chao[1], KAN Aike[2], ZENG Yelong[3], LI Guoqing[4], WANG Min[1], CI Ren[5]
(1. School of Geography, Faculty of Geographical Science, Beijing Normal University, Beijing 100875, China; 2. College of Geophysics, Chengdu University of Technology, Chengdu 610059, China; 3. Institute of Remote Sensing and Digital Earth, CAS, Beijing 100101, China; 4. School of Resources and Environmental Engineering, Ludong University, Yantai 264025, Shandong, China; 5. Institute of Science & Technology Information of Tibet Autonomous Region, Lhasa 85000, China)

Abstract: Clarifying the spatial pattern of population distribution, its influencing factors and regional differences at the township level is of great guiding significance for formulating sustainable development policies in ecologically fragile areas. Based on the population census data of Tibet at the township level in 2010, the population density and spatial factors were extracted. The density and clustering characteristics of the population distribution were analyzed by spatial statistical method. The multiple linear regression method and the random forest regression method were used to explore the population influencing factors and their regional differences of population distribution. The results showed that: (1) The population density of Tibet at the township level showed a strong spatial non-equilibrium. The general trend was high in the southeast and low in the northwest, and there was a strong spatial coupling between the main rivers and the main traffic trunks in high density area. (2) The "core-edge" characteristic of population clustering was obvious, and roughly to the wave of Borong (Nyalam County)-Gangni (Anduo County) as the demarcation line. (3) In the multiple linear regression method, the artificial surface index had the greatest influence on the population distribution, followed by the nighttime light index and road network density. (4) Random forest method was more accurate than multiple linear regression method to predict the population density, which can be used to sort the importance of the influencing factors. The influencing factors of the first six factors were the night light index, artificial surface index, road network density, industrial output value, GDP and multi-year average temperature, and these factors were positively correlated with population density. Among topographic factors, the contribution rate of elevation and slope was the largest, which was negatively correlated with population density. (5) The influencing factors and their interactions of population distribution in Tibet showed obvious regional differences. The valley was a gathering area for population in the study region, mainly in Lhasa River Valley, Nianchu River Valley and Sanjiang River Valley. (6) Through the analysis of random forest regression, the conceptual model can be used to express the influencing factors of population distribution, and the dominant factors were summarized as land use structure, road accessibility and urbanization level.

Keywords: population distribution; influencing factor; township scale; random forest; conceptual model