基于随机森林模型的珠江三角洲30 m格网人口空间化

谭 敏,刘 凯*,柳 林,朱远辉,王大山

(中山大学地理科学与规划学院 广东省城市化与地理环境空间模拟重点实验室
综合地理信息研究中心,广州 510275)

摘 要:人口空间化是实现人口统计数据与其他环境资源空间数据融合分析的有效途径。本文选取夜间灯光数据、道路网数据、水域分布数据、建成区数据、数字高程模型和地形坡度数据作为影响珠江三角洲人口分布的变量因子,利用随机森林模型对珠江三角洲2010年人口数据进行了30 m格网空间化,并将模拟结果与三个公开数据集作精度对比,最后基于随机森林模型的变量因子重要性分析珠江三角洲人口空间分布的影响因素。结果表明:本文模拟整体精度达到82.32%,均优于WorldPop数据集以及中国公里网格人口数据集,接近GPW数据集,而且在人口密度中等区域模拟精度最高;通过对变量因子重要性进行度量,发现夜间灯光强度是珠江三角洲人口分布的最重要指示性指标,到水域的距离、到建成区的距离和路网密度对珠江三角洲人口分布均具有重要作用。利用随机森林模型结合多源信息能够实现高空间分辨率的人口空间化,可为精细化城市管理提供重要数据源,也可为相关政策决策制定提供支持。

关 键 词:人口空间化;随机森林;人口分布;影响因素;珠江三角洲

1 引言

人口分布状况反映一个地区自然地理条件的差异和经济发展水平的高低,研究人口分布意义在于揭示人口分布的地域特点,进一步掌握人口空间分布的规律性(李玲等, 2008; 苏飞等, 2010)。准确掌握人口的空间分布信息,有助于科学制定区域发展规划、灾害风险防范与救助、经济建设、环境与生态保护等相关政策,有助于高效管理城市、改善居民生活环境,有助于提高人口、资源、环境综合管理的能力。人口密度网格化比人口密度行政单元更接近人口实际分布,而且是实现人口数据与其他社会统计数据、资源环境数据融合,提高人口、资源、环境综合管理能力的重要途径之一(周成虎等, 2009; 肖荣波等, 2011; 王鹤饶等, 2012; 卓莉等, 2014)。当前广泛使用的人口数据通常是以行政区划为单元,通过普查、抽样统计等方式逐级汇总获得的典型人口统计数据,在实际应用中存在时间分辨率低、空间分辨率低、直观性差、不支持空间运算和分析等不足(高义等, 2013)。遥感技术以其获取数据速度快、周期短、数据量大等优势在人口数据空间化中提供大量变量因子的数据来源,推动了基于多源数据融合的人口数据空间化。许多学者已经对不同数据源、不同尺度、不同模拟方法作了很多有益的探索(廖顺宝等, 2003; 康停军等, 2012; 王静等, 2012; 陈晴等, 2014; 张建辰等, 2014; Qi et al,

收稿日期:2017-03;**修订日期:**2017-08。

基金项目:国家自然科学基金重点项目(41531178);广州市科技计划项目(201510010081);国家自然科学基金项目(41001291)
[**Foundation:** Key Project of National Natural Science Foundation of China, No.41531178; Guangzhou Science and Technology Project, No.201510010081; National Natural Science Foundation of China, No.41001291]。

作者简介:谭敏(1993-),女,广东省广州市人,硕士研究生,主要从事资源环境遥感应用,E-mail: tanm3@mail2.sysu.edu.cn。

通讯作者:刘凯(1979-),男,黑龙江省伊春市人,博士,副教授,主要从事环境遥感与GIS应用、湿地遥感研究,
E-mail: liuk6@mail.sysu.edu.cn。

引用格式:谭敏, 刘凯, 柳林, 等. 2017. 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, 36(10): 1304-1312. [Tan M, Liu K, Liu L, et al. 2017. Spatialization of population in the Pearl River Delta in 30 m grids using random forest model[J]. Progress in Geography, 36(10): 1304-1312.]. DOI: 10.18306/dlkxjz.2017.10.012

2015; 柏中强, 王卷乐, 姜浩等, 2015; Gaughan et al, 2016; 王珂靖等, 2016)。但中尺度的研究较少，空间分辨率大部分为1 km，难以达到精细化城市管理的要求，而且对于变量因子的解释性较弱，较少利用空间变量分析影响人口分布的因素。

本文试图利用随机森林模型探索夜间灯光数据、道路网络数据、水域分布数据、建成区数据、数字高程模型和地形坡度数据等空间变量与珠江三角洲人口分布数据之间的关系，利用生成的随机森林模型实现珠江三角洲30 m网格人口空间化，并基于随机森林模型得到的变量因子重要性，分析珠江三角洲人口空间分布的影响因素。

2 研究区域与数据

2.1 研究区域

珠江三角洲位于中国广东省中部沿海，是西江、北江冲积形成的大三角洲和东江冲积形成的小三角洲的合称(图1)。三角洲属于亚热带气候，雨热同期，土壤肥沃，河道纵横，适宜农业种植(珠江三角洲城市群年鉴编纂委员会, 2015)。1995年，广东省政府在《珠江三角洲经济区经济社会发展规划(1996-2010年)》中，将“珠江三角洲经济区”范围调整为位于珠江沿岸的广州、深圳、佛山、珠海、东莞、中山、江门7个地级行政区及惠州、肇庆2个地级行政区的一部分。珠江三角洲是全国经济发展最迅速的地区之一。随着经济的快速发展，大量外来人口迁入，如今珠江三角洲是中国市场化程度最高、经济最发达、人口密度最高的地域之一(刘志佳等,

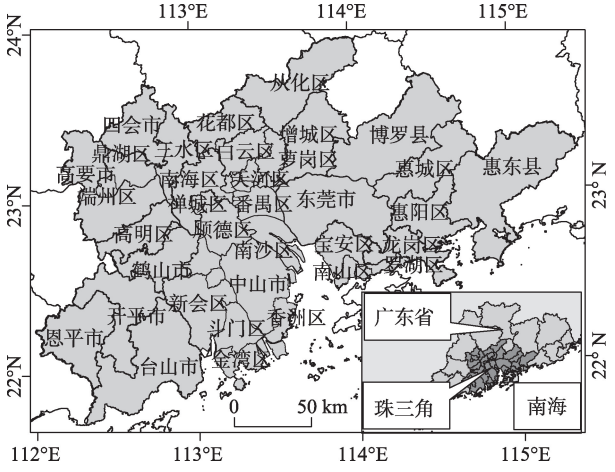

图1　珠江三角洲区位图

Fig.1　Location of the Pearl River Delta

2015; 周春山等, 2015)。

2.2 数据简介

本文使用的数据包括：①2010年的夜间灯光数据，来源于美国国家地理数据中心(http://ngdc.noaa.gov/eog/dmsp/downloadV4composites.html)，采用的是基于F18传感器的2010年夜间非辐射定标平均稳态数据，可见像素值范围为0~63；②珠江三角洲2010年30 m分辨率土地覆盖数据，包括建成区、道路、河流、水体数据，利用2010年Landsat5 TM卫星影像，通过基于面向对象的人机交互目视解译获得，经过与更高分辨率的影像对比及野外核查，分类总体精度达到90%以上；③珠江三角洲90 m分辨率的SRTM数字高程模型及计算的地形坡度数据；④2010年珠江三角洲区县级、广州市镇街级第六次人口普查数据；⑤珠江三角洲区县级行政区划矢量边界及广州市镇街级行政区划矢量边界数据；⑥用于精度检验的公开数据集，包括：a) 2010年的WorldPop数据集，来源于WorldPop项目官网(http://www.worldpop.org.uk)，空间分辨率为100 m；b) 2010年的GPW v4数据集，来源于NASA的社会经济数据和应用中心(http://sedac.ciesin.columbia.edu)，空间分辨率是30弧秒，赤道处约为1 km；c)2010年的中国公里网格人口分布数据集，数据来源于国家科技基础条件平台——国家地球系统科学数据共享平台(http://www.geodata.cn)，空间分辨率为1 km。

2.3 数据预处理

数据预处理主要包括投影转换和人口分布的影响因子的计算。首先将前述所有空间数据投影至统一坐标系统下，然后计算人口分布影响因子，具体步骤如下：

(1) 将河流、水体、道路网和建成区转换成30 m栅格数据，分别与行政区划边界叠加得到二值化栅格数据，即如果一个栅格的土地覆盖类型为河流、水体、道路网和建成区四种类型之一，则该栅格的值为1，否则为0。计算出珠江三角洲范围内每个30 m×30 m网格分别到河流、水体、道路网和建成区的欧氏距离后，再利用珠江三角洲区县级行政区划边界统计每个区县内分别到河流、水体、道路网和建成区的平均距离。

(2) 对于栅格格式的夜间灯光数据、数字高程模型和地形坡度数据，利用珠江三角洲区县级行政区划边界对其统计得到每个区县内的平均夜间灯光强度、平均高程和平均坡度。

(3) 对于道路网和水域，分别统计每个区县内的铁路长度、国道长度、省道长度和县道长度，同时引入路网密度(RSD)和河网密度(WSD)两个评价指标，其中路网密度计算公式为：

$$RSD=(3\times N_{r}+3\times N_{ne}+2\times N_{pe}+1\times N_{cr})/A \quad (1)$$

式中：RSD为路网密度；N_{r}、N_{ne}、N_{pe}、N_{cr}分别为各区县内铁路、国家主干道、省级公路、县道长度(km)；A为区县行政单元面积(km^2)。考虑到不同等级公路的运输容量和通行能力的差异，将各等级道路里程换算为标准县道长度，式中系数3、3、2、1分别为铁路、国家主干道、省级公路、县道的换算系数(柏中强, 王卷乐, 杨雅萍等, 2015)。

河网密度计算公式为

$$WSD=N_{w}/A \quad (2)$$

式中：N_{w}为各个区县内的总河流长度(km)；A为各个区县的面积(km^2)。按照上述方法，统计每个30 m×30 m网格内的铁路长度、国道长度、省道长度、县道长度以及路网密度和河网密度。

(4) 对于人口数据，统计每个区县的土地面积后计算每个区县的人口密度，对结果取对数。

3 研究方法

本文基于前人关于人口分布影响因素的研究(方瑜等, 2012; 柏中强, 王卷乐, 杨雅萍等, 2015; Gaughan et al, 2016)，并结合珠江三角洲的特点，选取夜间灯光强度、到水体的距离、到道路的距离、到建成区的距离、铁路长度、国道长度、省道长度、县道长度、路网密度、河网密度、行政区面积、高程和坡度作为人口分布的变量因子，运用随机森林模型建立人口密度与变量因子之间的关系，并利用生成的随机森林树对每个30 m×30 m栅格的人口密度进行估算，通过分区密度制图得到珠江三角洲的30 m×30 m网格的人口分布图并作精度验证，最后对变量因子进行重要性度量，分析影响珠江三角洲人口分布的因素。相关的技术路线如图2所示。

3.1 随机森林模型

随机森林是由Leo Breiman和Cutlery Adele在2001年提出的一种分类回归树的数据挖掘方法，是一种组合式的自学习技术(引自张雷等, 2014)。随机森林中每棵树的训练集是应用bootstrap方法从总训练集中有放回地随机抽样获得。假设有M个原始变量，对于采集的样本随机选择$mtry$($mtry$<<M)个特征向量作为决策树分裂的候选变量，从这$mtry$个候选变量中选择信息含量最丰富的变量进行节点分裂，而且在树的生长过程不进行修剪。按照这种方式生成$ntree$棵决策树，然后通过$ntree$棵树的反馈进行预测，如果是分类则由多数投票决定，如果是回归则计算平均值。每次未被抽到的样本组成了袋外数据(Out-Of-Bag, OOB)，这些袋外数据可用于度量变量因子的重要性，变量重要性的值越大说明该变量因子的重要性越高、越能解释因变量。变量因子重要性的度量方法常用的有两种，分别为平均精度减少法和平均基尼系数下降法。平均精度减少法通过计算袋外数据自变量值发生轻微扰动后的分类正确率与扰动前分类正确率的平均减少量来衡量变量的重要性；平均基尼系数下降法则是遍历所有树节点，统计每个特征变量对应的基尼系数下降总和作为该特征的贡献度。OOB数据还可以用于估计模型的性能，Breiman通过实验证明OOB估计是无偏估计(Breiman, 2001a)。

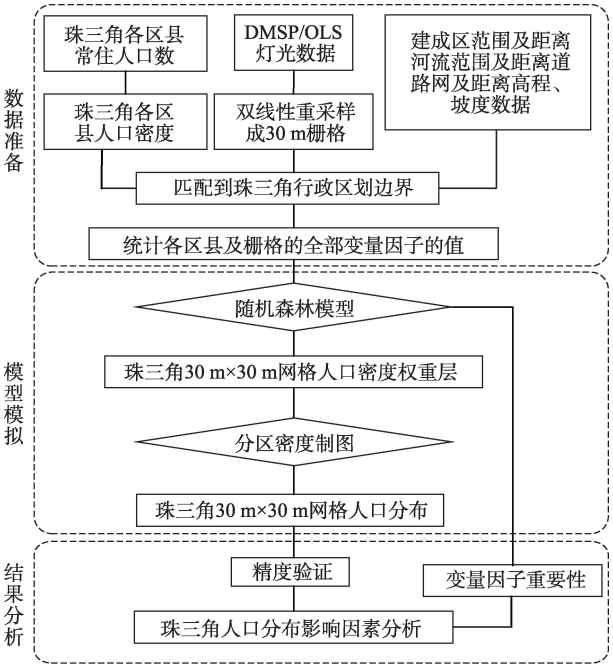

图2 技术路线图

Fig.2 Flowchart of spatializing population based on random forest model

随机森林模型的优点在于：第一，避免了过度拟合。因为在决策树生长过程中bootstrap的采样方法使得每棵决策树不是由全部样本生成，在生成决策树生长过程也不是利用全部变量进行分裂。第二，它对异常值和噪声具有很高的容忍度(Breiman, 2001b; 方匡南等, 2011)。第三，它能度量变量

的重要性，对于了解影响人口分布的机制有明显的积极作用。第四，随机森林能在运算量没有显著提高的前提下提高预测精度，为快速且准确地实现大范围精细栅格的人口空间化提供有力支撑。同时Stevens等(2015)指出，在人口空间化中使用随机森林模型，需要和GIS行政区域边界匹配得很好的人口普查数据。

基于上述优点，本文在获得珠江三角洲2010年准确的行政边界及第六次人口普查数据的基础上，基于R语言的randomForest包实现利用随机森林模型进行珠江三角洲30 m网格的人口空间化。首先输入样本，以珠江三角洲43个区县的人口密度作为因变量，13个影响因子作为自变量，包括夜间灯光强度、到水体的距离、到道路的距离、到建成区的距离、铁路长度、国道长度、省道长度、县道长度、路网密度、河网密度、行政区面积、高程和坡度。然后对随机森林模型进行训练，训练时有两个重要的参数：*ntree*和*mtry*。*ntree*表示决策树的数量，*mtry*表示决策树分裂时候选变量的个数。由于采样时使用的是bootstrap有放回的采样方法，因此原始训练集中约63.2%的样本被采集，剩余的36.8%样本组成袋外数据，对样本进行交叉验证(方匡南等, 2011)。所以本文利用OOB无偏估计得到不同参数设置下随机森林模型的精度，进行参数设置(陈凯等, 2015)。首先确定参数*mtry*，在决策树的棵数较大的前提下(*ntree*=500)，测试*mtry*不同取值时随机森林模型的精度。如图3所示，通过OOB无偏估计得到的模型精度随着*mtry*的增加先大幅提高后缓慢降低，在*mtry*=4处取得最大值89.83%，所以本文中*mtry*参数设置为4。当*mtry*=4时，*ntree*的增加使得模型精度不断提高，在*ntree*=500时模型精度达到89.93%，之后精度都接近90%并有小幅度波动(图4)。综合考虑模型精度与计算机运行性能，本文中*ntree*参数设置为500。采用设置好的参数，在珠江三角洲区县级进行随机森林模型的训练，然后将生成的随机森林应用到每个30 m×30 m的网格中，预测每个30 m×30 m网格的人口密度，初步实现珠江三角洲30 m×30 m格网的人口空间化。基于OOB数据，采用平均基尼系数下降法对变量因子进行重要性度量。

3.2 分区密度制图

由于用随机森林模型估计得到的每个网格的人口数是根据区县级的人口变量因子与人口数生成的随机森林估计，所以实际的人口密度分布需要每个区县的人口总数进行总量控制，按照随机森林得到的每个网格的人口占一个区县所有网格总人口的比例重新分配每个网格的人口数，公式如下：

$$P_i = S_j \times D_i / D_j \tag{3}$$

式中：i为每个网格，j为每个行政区，P_i为每个网格内的人口数；S_j为该网格所在的行政区的人口总数；D_i为该网格根据随机森林模型估计得到的人口数；D_j为该网格所在行政区的所有网格根据随机森林模型估计得到的人口总数。

3.3 精度检验

本文模拟30 m网格人口分布是根据区县级人口普查数据进行的人口数据空间化，为反映人口空间化模型的精度水平，选取广州市166个乡镇街道的人口普查数据进行精度检验。将166个乡镇街道按照第六次人口普查数据计算得到的人口密度分成三组：人口密度小于0.5万人/km²的人口密度低

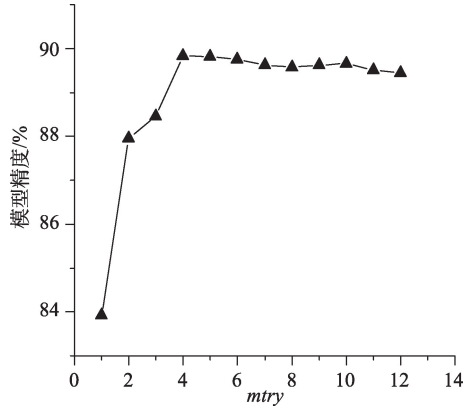

图3　模型精度与*mtry*之间的关系(*ntree*=500)

Fig.3　Relationship between model accuracy and predictive variables with 500 trees

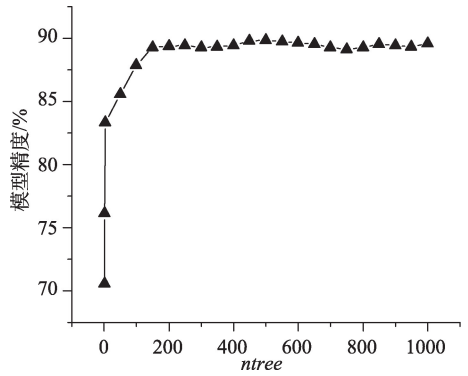

图4　模型精度与*ntree*之间的关系(*mtry*=4)

Fig.4　Relationship between model accuracy and ntree with 4 predictive variables

组(72个镇街),介于0.5万人/km²和5.6万人/km²之间的人口密度中等组(80个镇街),以及大于5.6万人/km²的人口密度高组(14个镇街);同时采用平均绝对误差(*MAE*)、均方根误差(*RMSE*)和相对均方根误差(%*RMSE*)来衡量对比每组及全部的人口普查数据与WorldPop数据集、GPW v4数据集、中国公里网格人口分布数据集的精度,并进一步采用Taylor图(Taylor, 2001)对比本文模拟结果和三种公开数据集。其中*MAE*是相对误差取绝对值再算平均,避免了正负相抵消的情况;*%RMSE*是通过均方根误差除以人口普查数的平均值得到,可以反映模型模拟的精度高低(Stevens et al, 2015)。

$$MAE=\frac{1}{N}\sum|f_i-r_i| \tag{4}$$

$$RMSE=\sqrt{\frac{1}{N}\sum(f_i-r_i)^2} \tag{5}$$

$$\%RMSE=\frac{RMSE}{\frac{1}{N}\sum r_i} \tag{6}$$

式中:f_i是第i组数据的估算值,即本文进行人口空间化后得到的人口密度估算值;r_i是第i组数据的参考值,即由人口普查数据得到的人口密度值;N代表组数数据。

4 结果分析

经过随机森林学习及分区密度制图后得到2010年珠江三角洲30 m网格空间分辨率的人口密度分布图(图5)。2010年珠江三角洲人口分布总体呈现出较显著的“双核心—边缘—外围”结构,人口大量集中分布在广州越秀区、荔湾区、海珠区和天河区以及深圳福田区和罗湖区,最高人口密度出现在广州市越秀区。

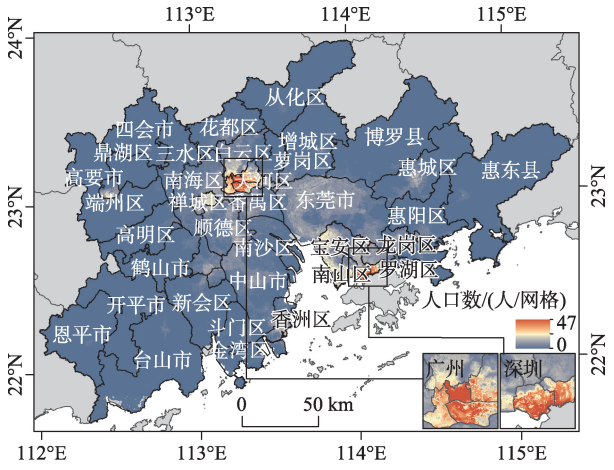

图5 2010年珠江三角洲30 m网格空间分辨率的人口分布图

Fig.5 Spatial distribution of population of the Pearl River Delta in 30 m × 30 m grids, 2010

4.1 精度检验

将广州市166个街道进行人口密度空间化结果与WorldPop、GPW、中国公里网格人口数据集三个公开数据集进行对比分析,误差计算结果如表1所示。由表1可知,在验证区广州市,本文模拟整体精度达到82.32%,均优于WorldPop数据集以及中国公里网格人口数据集,略低于GPW数据集。在人口密度低的区域,本文的模拟精度低于另外三个数据集;在人口密度中等的区域,本文的模拟精度明显优于另外三个数据集;在人口密度较高的区域,本文的模拟精度明显优于WorldPop数据集以及中国公里网格人口数据集,略低于GPW数据集。

采用Taylor图将本文结果与三个数据集进行对比。图6中的虚线同心圆弧表示标准化的标准差,实线同心圆弧代表了标准化的中心均方根误差,数据点与圆心的连线与90°的半径所围成扇形的角度代表了相关系数。也就是说,数据点到半径为1的虚线圆弧的距离越近,则它的标准差越小,数

表1 精度检验指标计算结果

Tab.1 Accuracy assessment result

	数据集	*MAE*	*RMSE*	%*RMSE*/%
广州区域	本文	7146.73	13530.23	17.68
	WorldPop	9599.64	17597.13	23.00
	GPW	7272.93	13198.29	17.25
	中国公里网格人口数据集	13396.08	22551.34	29.73
人口密度低	本文	1374.80	2815.25	151.98
	WorldPop	1300.97	2020.59	107.27
	GPW	753.38	1170.74	63.09
	中国公里网格人口数据集	767.42	1101.52	58.02
人口密度中等	本文	7221.09	10087.84	45.93
	WorldPop	10068.19	13889.55	63.26
	GPW	10663.92	15526.47	70.69
	中国公里网格人口数据集	18472.54	22538.66	104.95
人口密度高	本文	36823.59	39434.55	58.09
	WorldPop	49972.75	50613.87	74.56
	GPW	24609.52	28022.69	41.46
	中国公里网格人口数据集	61669.22	62285.80	93.20

据点离实线圆弧的圆心的距离越小，则它的中心均方根误差就越小，数据点与圆心的连线与90°的半径所围成的扇形的角度越大则相关系数越高。总而言之，数据点与*REF*点的距离越小则精度越高，距离相等时越接近值为1的虚线圆弧的精度越高。

将四个数据集在人口密度较低、中等、较高以及全广州区域的模拟结果分别以点表示在Taylor图上，如图6所示。可以看出，*D*1、*D*2、*D*3和*D*4四个点中，*D*1与*REF*的距离最近，所以本文的模拟结果是四个数据集中的最优。而且在人口密度中等的区域，即*B*1、*B*2、*B*3和*B*4四个点中，*B*1与*REF*的距离最近，所以*B*1的精度优于*B*2、*B*3、*B*4，亦即是本文的模拟结果在人口密度中等的区域优于其他三个数据集。同理，在人口密度较低的区域，本文模拟结果低于其他三个数据集；在人口密度较高的区域，本文的模拟结果均优于WorldPop与中国公里网格人口数据集，略差于GPW数据集。

从数据精度分析来看，本文方法的空间化人口数据在广州市乡镇级尺度精度评价结果相比其他三种数据源的提升不明显，但是本文数据空间分辨率的优势是数据能够展示更多空间细节和格局差异。选取人口密度高的深圳福田、人口密度中等的佛山禅城、人口密度低的江门台山进行人口空间化比较结果展示(图7)。四个数据集的人口密度分布趋势大致相同，但本文30 m格网化结果不管在人口密度高、中或低的区域人口分布异质性更大，而且在同一比例尺下，边缘锯齿状不明显，过渡更加自然符合实际，可以反映更加丰富的人口密度信息。

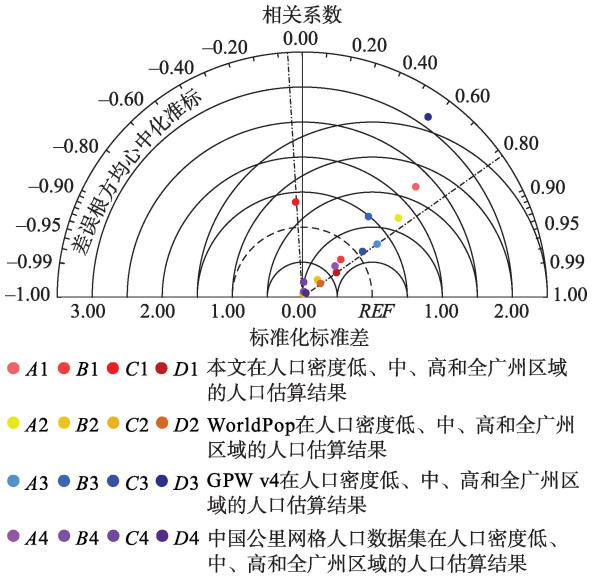

图6　四个数据集的Taylor图

Fig.6　Taylor diagram of the four datasets

4.2　变量因子重要性

基于随机森林模型的OOB数据得到变量因子重要性值(图8)，值越大，表明该变量因子作用越大。

从上述分析可见，夜间灯光强度是珠江三角洲人口分布最重要的指示性指标，到水域的距离、到建成区的距离和路网密度对珠江三角洲人口分布均具有重要作用。这个结果也和前人的研究基本一致(廖顺宝等, 2003; 柏中强, 王卷乐, 杨雅萍等, 2015; Stevens et al, 2015; Gaughan et al, 2016)。改革开放以来，以广州市、深圳市为中心的珠江三角洲地区的第二、三产业快速发展，城市也开始快速扩展，成为经济增长的热点区域，吸引了大量的外来人口迁入，为城市管理与正常运行需要，基础设施逐步完善，例如路灯设施，所以夜间灯光强度与人口分布具有显著的相关性，灯光强的区域人口分布密集。珠江三角洲地区地处珠江流域下游，气候

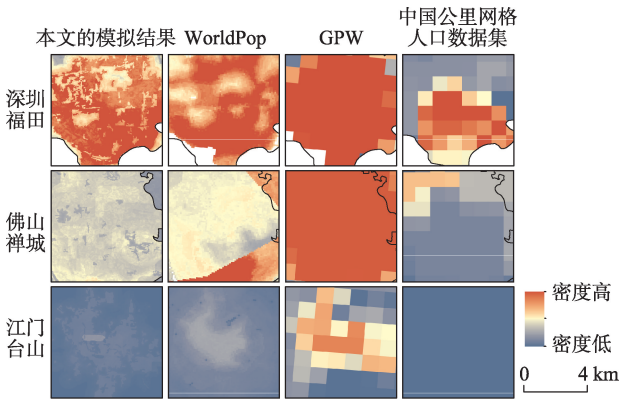

图7　四个数据集在不同人口密度区域的空间显示效果

Fig.7　Four datasets displayed in space in different density area

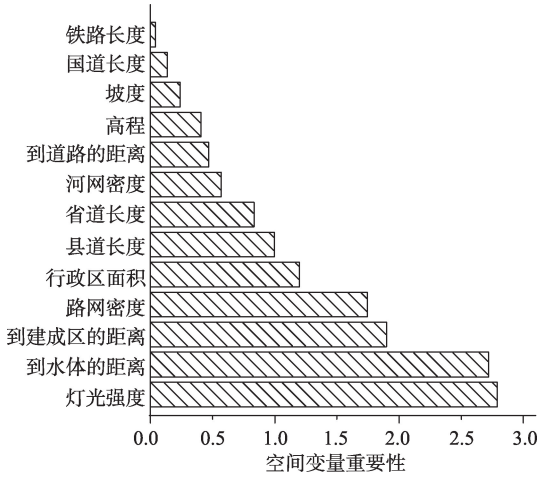

图8　随机森林模型生成的变量因子重要性图

Fig.8　Variable importance for random forest regression

温热多雨，有着数目繁多的池塘与沼泽湿地，给原始居民创造了得天独厚的区位条件，于是流传下来的依山傍水而居思想影响了后代人对居住地的选择，所以到水域的距离也成为影响珠江三角洲人口分布的一个重要因素。城市发展具有集聚效应，建设用地连片形成建成区，建成区内市政设施完善，政府机关较多，卫生教育机构齐全，商业发达，吸引住户商户，所以建成区内的人口密度最高，按照距离衰减学说(Martin, 1989)，距离建成区越远则人口分布越稀疏，珠江三角洲的人口分布也符合此模型。路网密度也是影响珠江三角洲人口分布的重要因素之一，因为在城市中物质与居民的交流运输都要依靠道路，路网密度越高，则道路的通达性越好，就越吸引居民居住；同时越多的居民迁入后，为了居民的出行也会更加完善道路交通，路网密度进一步提高，所以路网密度也是影响人口分布的重要因素。

通过与其他三个人口空间化数据集对比，发现四个数据集的人口分布趋势是相同的，本文的估算值也接近人口普查的人口数，说明本文结果科学合理。30 m空间分辨率的人口分布数据是实现人口、资源、环境和社会经济综合有效管理的基础，为精细化城市管理提供了重要参考，具备实用意义。

5 结论与讨论

本文利用夜间灯光遥感数据、土地覆盖数据、数字高程模型及地形坡度数据，基于随机森林模型，对珠江三角洲区域2010年的人口统计数据进行30 m格网的空间化，利用平均绝对误差、均方根误差和相对均方根误差以及Taylor图分析对模拟结果作精度对比评价，并利用随机森林模型的变量因子重要性分析珠江三角洲人口分布的影响因素。主要结论如下：

(1) 本文模拟整体精度达到82.32%，均优于WorldPop数据集以及中国公里网格人口数据集，接近GPW数据集，而且在人口密度中等区域的模拟精度最高，在人口密度较低或较高区域模拟精度均有所下降。同时本文的最终模拟结果的空间分辨率为30 m，较另外三种公开人口数据能细致地描述珠江三角洲人口分布的空间异质性，同时能满足精细化城市管理的需求以及多种尺度需求的应用。

(2) 夜间灯光强度是珠江三角洲人口分布的最重要指示性指标，到水域的距离、到建成区的距离和路网密度对珠江三角洲人口分布均具有重要作用。

本文结果与方瑜等(2012)的研究结果，即地形因子是影响人口分布的重要因素不一致，原因是该研究为全国尺度的，全国地形分为三级阶梯，地形起伏大，而珠江三角洲是冲积平原，大部分区域是平原或丘陵，整体地形比较平缓，高程、坡度略高的区域的气候环境变化不明显，因此地形因子对珠江三角洲区域的人口分布约束较小。

虽然本文整体的模拟精度较高，但在人口密度较低或较高的区域模拟精度不够理想，这可能由于随机森林模型建立时选取的因素不能完全反映人口密度较低或较高的区域的人口分布的特征，因此后续研究可考虑对研究区进行分级建模，同时进一步研究人口分布的机制，更加准确合理地选择变量因子。

参考文献(References)

柏中强, 王卷乐, 姜浩, 等. 2015. 基于多源信息的人口分布格网化方法研究[J]. 地球信息科学学报, 17(6): 653-660. [Bai Z Q, Wang J L, Jiang H, et al. 2015. The gridding approach to redistribute population based on multi-source data[J]. Journal of Geo-Information Science, 17(6): 653-660.]

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

陈凯, 刘凯, 柳林, 等. 2015. 基于随机森林的元胞自动机城市扩展模拟: 以佛山市为例[J]. 地理科学进展, 34(8): 937-946. [Chen K, Liu K, Liu L, et al. 2015. Urban expansion simulation by random-forest-based cellular automata: A case study of Foshan City[J]. Progress in Geography, 34(8): 937-946.]

陈晴, 侯西勇, 吴莉. 2014. 基于土地利用数据和夜间灯光数据的人口空间化模型对比分析: 以黄河三角洲高效生态经济区为例[J]. 人文地理, 29(5): 94-100. [Chen Q, Hou X Y, Wu L. 2014. Comparing of population spatialization models based on land use data and DMSP/OLS data respectively: A case study in the efficient ecological economic zone of the Yellow River Delta[J]. Human Geography, 29(5): 94-100.]

方匡南, 吴见彬, 朱建平, 等. 2011. 随机森林方法研究综述[J]. 统计与信息论坛, 26(3): 32-38. [Fang K N, Wu J B,

Zhu J P, et al. 2011. A review of technologies on random forests[J]. Statistics & Information Forum, 26(3): 32-38.]

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

高义, 王辉, 王培涛, 等. 2013. 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, 35(12): 2517-2523. [Gao Y, Wang H, Wang P T, et al. 2013. Population spatial processing for Chinese coastal zones based on census and multiple night light data[J]. Resources Science, 35(12): 2517-2523.]

康停军, 张新长, 赵元, 等. 2012. 基于多智能体的城市人口分布模型[J]. 地理科学, 32(7): 790-797. [Kang T J, Zhang X C, Zhao Y, et al. 2012. Agent-based urban population distribution model[J]. Scientia Geographica Sinica, 32(7): 790-797.]

李玲, 沈静, 袁媛. 2008. 人口发展与区域规划[M]. 北京: 科学出版社. [Li L, Shen J, Yuan Y. 2008. Renkou fazhan yu quyu guihua[M]. Beijing, China: Science Press.]

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

刘志佳, 黄河清. 2015. 珠三角地区建设用地扩张与经济、人口变化之间相互作用的时空演变特征分析[J]. 资源科学, 37(7): 1394-1402. [Liu Z J, Huang H Q. 2015. Tempo-spatial characteristics of interactions among changes in built-up land, GDP and demography in the Pearl River Delta[J]. Resources Science, 37(7): 1394-1402.]

苏飞, 张平宇. 2010. 辽中南城市群人口分布的时空演变特征[J]. 地理科学进展, 29(1): 96-102. [Su F, Zhang P Y. 2010. Spatio-temporal dynamics of population distribution in the middle and southern Liaoning Urban Agglomeration [J]. Progress in Geography, 29(1): 96-102.]

王鹤饶, 郑新奇, 袁涛. 2012. DMSP/OLS数据应用研究综述[J]. 地理科学进展, 31(1): 11-19. [Wang H R, Zheng X Q, Yuan T. 2012. Overview of researches based on DMSP/OLS nighttime light data[J]. Progress in Geography, 31(1): 11-19.]

王静, 杨小唤, 石瑞香. 2012. 山东省人口空间分布格局的多尺度分析[J]. 地理科学进展, 31(2): 176-182. [Wang J, Yang X H, Shi R X. 2012. Spatial distribution of the population in Shandong Province at multi-scales[J]. Progress in Geography, 31(2): 176-182.]

王珂靖, 蔡红艳, 杨小唤. 2016. 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[J]. 地理科学进展, 35(12): 1494-1505. [Wang K J, Cai H Y, Yang X H. 2016. Multiple scale spatialization of demographic data with multi- factor linear regression and geographically weighted regression models[J]. Progress in Geography, 35 (12): 1494-1505.]

肖荣波, 丁琛. 2011. 城市规划中人口空间分布模拟方法研究[J]. 中国人口·资源与环境, 21(6): 13-18. [Xiao R B, Ding C. 2011. Modeling spatial distribution of population density for urban planning[J]. China Population Resources and Environment, 21(6): 13-18.]

张建辰, 王艳慧. 2014. 基于土地利用类型的村级人口空间分布模拟: 以湖北鹤峰县为例[J]. 地球信息科学学报, 16 (3): 435-442. [Zhang J C, Wang Y H. 2014. Simulation of village-level population distribution based on land use: A case study of Hefeng County in Hubei Province[J]. Journal of Geo-Information Science, 16(3): 435-442.]

张雷, 王琳琳, 张旭东, 等. 2014. 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[J]. 生态学报, 34(3): 650-659. [Zhang L, Wang L L, Zhang X D, et al. 2014. The basic principle of random forest and its applications in ecology: A case study of *Pinus yunnanensis*[J]. Acta Ecologica Sinica, 34(3): 650-659.]

周成虎, 欧阳, 马廷. 2009. 地理格网模型研究进展[J]. 地理科学进展, 28(5): 657-662. [Zhou C H, Ou Y, Ma T. 2009. Progresses of geographical grid systems researches[J]. Progress in Geography, 28(5): 657-662.]

周春山, 金万富, 史晨怡. 2015. 新时期珠江三角洲城市群发展战略的思考[J]. 地理科学进展, 34(3): 302-312. [Zhou C S, Jin W F, Shi C Y. 2015. Development strategy of the Pearl River Delta Urban Agglomeration under the current socioeconomic situation[J]. Progress in Geography, 34(3): 302-312.]

珠江三角洲城市群年鉴编纂委员会. 2015. 珠江三角洲城市群年鉴2015[M]. 广州: 广东人民出版社. [Urban Agglomeration in the Pearl River Delta Yearbook Committee Compilation. 2015. Urban Agglomeration in the Pearl River Delta yearbook 2015[M]. Guangzhou, China: Guangdong People's Publishing House.]

卓莉, 黄信锐, 陶海燕, 等. 2014. 基于多智能体模型与建筑物信息的高空间分辨率人口分布模拟[J]. 地理研究, 33 (3): 520-531. [Zhuo L, Huang X R, Tao H Y, et al. 2014. High spatial resolution population distribution simulation based on building information and multi-agent[J]. Geo-

graphical Research, 33(3): 520-531.]

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

Breiman L. 2001b. Statistical modeling: The two cultures[J]. Statistical Science, 16(3): 199-231.

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

Martin D. 1989. Mapping population data from zone centroid locations[J]. Transactions of the Institute of British Geographers, 14(1): 90-97.

Qi W, Liu S H, Gao X L, et al. 2015. Modeling the spatial distribution of urban population during the daytime and at night based on land use: A case study in Beijing, China[J]. Journal of Geographical Sciences, 25(6): 756-768.

Stevens F R, Gaughan A E, Linard C, et al. 2015. Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data[J]. PLoS One, 10(2): e0107042.

Taylor K E. 2001. Summarizing multiple aspects of model performance in a single diagram[J]. Journal of Geophysical Research, 106(D7): 7183-7192.

Spatialization of population in the Pearl River Delta in 30 m grids using random forest model

TAN Min, LIU Kai*, LIU Lin, ZHU Yuanhui, WANG Dashan

(Center of Integrated Geographic Information Analysis, Guangdong Key Laboratory for Urbanization and Geo-simulation, School of Geography and Planning, Sun Yat-sen University, Guangzhou 510275, China)

Abstract: Grid population data can enable integrated analysis of population statistics with other spatial data on resources and the environment. Based on a Random Forest model and using nighttime lights, road network, surface water network, built-up area, slope, and DEM as control variables, the 2010 population data of the Pearl River Delta were distributed into 30 m grids. The estimation results were compared with three other public datasets. The importance of input variables was analyzed based on the results. The result shows that the accuracy of this simulation reached 83.32%, which is better than the WorldPop and the Population Grids of China datasets, and more close to the GPW dataset. Moreover, the 30 m resolution of our result furnishes detailed information of population density of the Pearl River Delta. According to the importance of covariates from the Random Forest model, strength of nighttime lights, distance to water, distance to built-up area, and density of roads are important factors in population distribution modeling in the Pearl River Delta. With the Random Forest model and multi-source data, high resolution population spatialization can be achieved. High spatial resolution grid data can provide important data source for high precision city management and policy making.

Key words: population spatialization; random forest; population distribution; impact factors; the Pearl River Delta