第17届中美碳联盟年会

The 17th US-China Carbon Consortium Annual Meeting

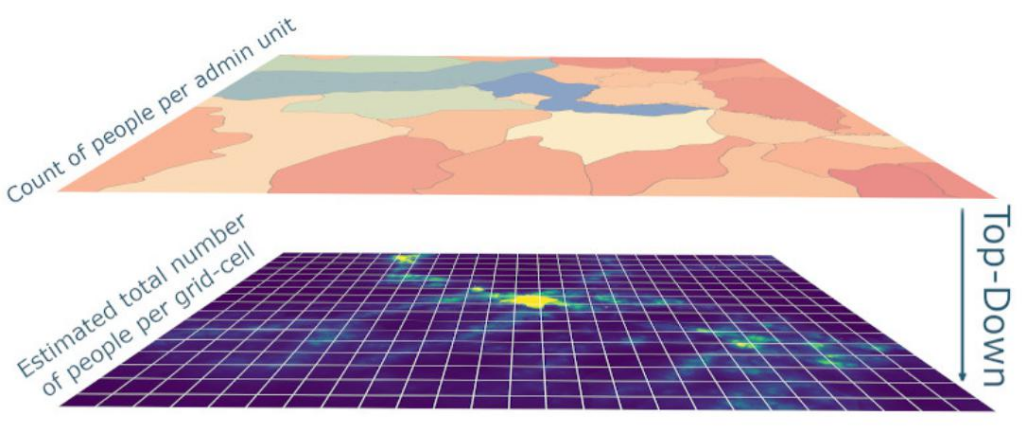
基于随机森林算法的城市人口多尺度空间化研究

汇报人：周云　指导老师：马明国教授

目录

一 研究绪论

基于随机森林算法的城市人口多尺度空间化研究



人口数据空间化：基于人口空间分布模型或采用一定的算法，对人口统计数据进行**离散化处理**，发掘并展现其中隐含的空间信息，以便模拟或再现客观世界的**人口地理分布**。

1.2 背景与应用

- 为什么研究人口？ 我国的基本国情使然
- 人口数据重要性？ 有助于城市精细化管理

人口是影响碳排放的关键因素之一

➢ 人口数据现状？ 现有人口数据获取难度大、精度低

人口普查现状

人力物力花费大，部分数据不对外公开。

人口数据产品质量不一

现有产品时空分辨率无法满足研究与应用需要。

1.4 理论现状

近二十年以来，地理信息技术的发展和移动通信网络技术的普及，使得人口空间化建模的数据来源更加丰富，方法更加智能先进。

分区密度制图

在保证每个多边形的输入要素**总量不变**时，再根据**空间权重**确定输入要素的**密度和分布**。

权重计算方法

一元线性回归　多元线性回归

地理加权回归　**随机森林回归**

由于**随机森林回归**在训练和解释模型方面的**突出优势**（王超等，2019），近年来该方法来被**广泛运用**于人口数据空间化研究（Sinha et al.，2019）。

人口空间化过程的尺度效应?

已有的研究表明人口分布具有显着的空间自相关性和尺度依赖性，因此在人口空间化过程中，尤其要注重空间分辨率的选择（王珂靖，2015）。合适的网格大小，既能体现出人口详细分布状况，又能反映人口分布的差异性（李双成和蔡运龙，2005）。

1. **全球尺度** 千米级别
2. **区域和国家尺度** 百米级别
3. **城市尺度** 十米级别
4. **街道尺度** 建筑物级别

现有研究中的尺度选择仍有较大主观性，针对人口密集的城市区域，从不同尺度上来研究人口分布格局，探求不同区域的尺度适宜性的研究尤为迫切。

1.5 研究内容

1. 如何选取及融合多源数据进行人口空间化?
2. 如何确定城市人口空间化的适宜尺度?
3. 如何用随机森林回归得到人口分布权重?
4. 如何定量分析特征变量的重要性程度?

二 数据与方法

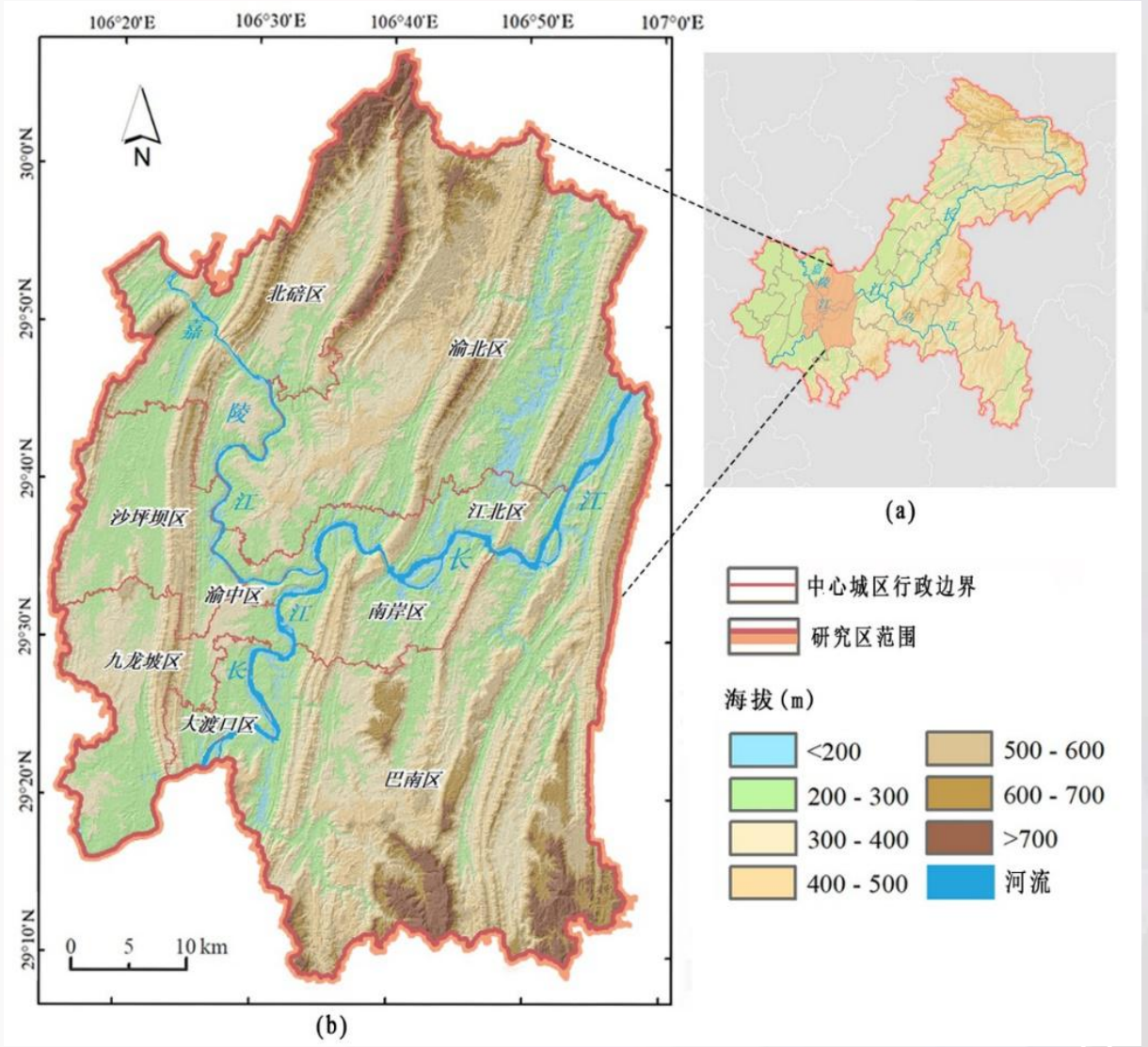
2.1 研究区域

为什么选择重庆市主城区？

1. 山地城市**自然条件复杂**，人口分布不均衡的现象较平原城市更加明显，是开展复杂城市人口网格化研究较为适宜区域。

2. 该地区日渐成为支撑**经济社会发展**的重要地区。人口分布的情况受到许多关注。

人口总数	全市占比	地区生产总值	全市占比	城镇化率
875万人	28.21%	8.21×10^7万元	40.31%	90.51%
平均海拔	最大高差	土地总面积	人口密度	
390m	1367m	5466.28km^2	1600人/km^2	



研究区数据简介及预处理

类型	格式	年份	预处理方法
13类地图兴趣点	矢量点	2018	核密度分析
住宅用地	矢量面	2018	新建渔网再栅格化
道路网	矢量线	2018	欧式距离分析
夜间灯光	栅格(130m)	2018	几何校正与辐射校正
数字高程模型	栅格(30m)	2010	坡度分析与阴影计算
人口统计	表格	2018	计算人口密度
行政区划边界	矢量面	2018	面链接到表格

2.2 多源数据

地图兴趣点数据

地图兴趣点数据（Points of interest，POI）通常是指存在于**电子地图**中有地理坐标的矢量点数据集（李泽宇和董春，2019）。

已有的研究表明：对于较小的研究区域而言，该数据能够提供更加**丰富的语义信息**，在人口密集的城市地区**提高**了人口**估算精度**（Langford，2013）。

十三类点数据

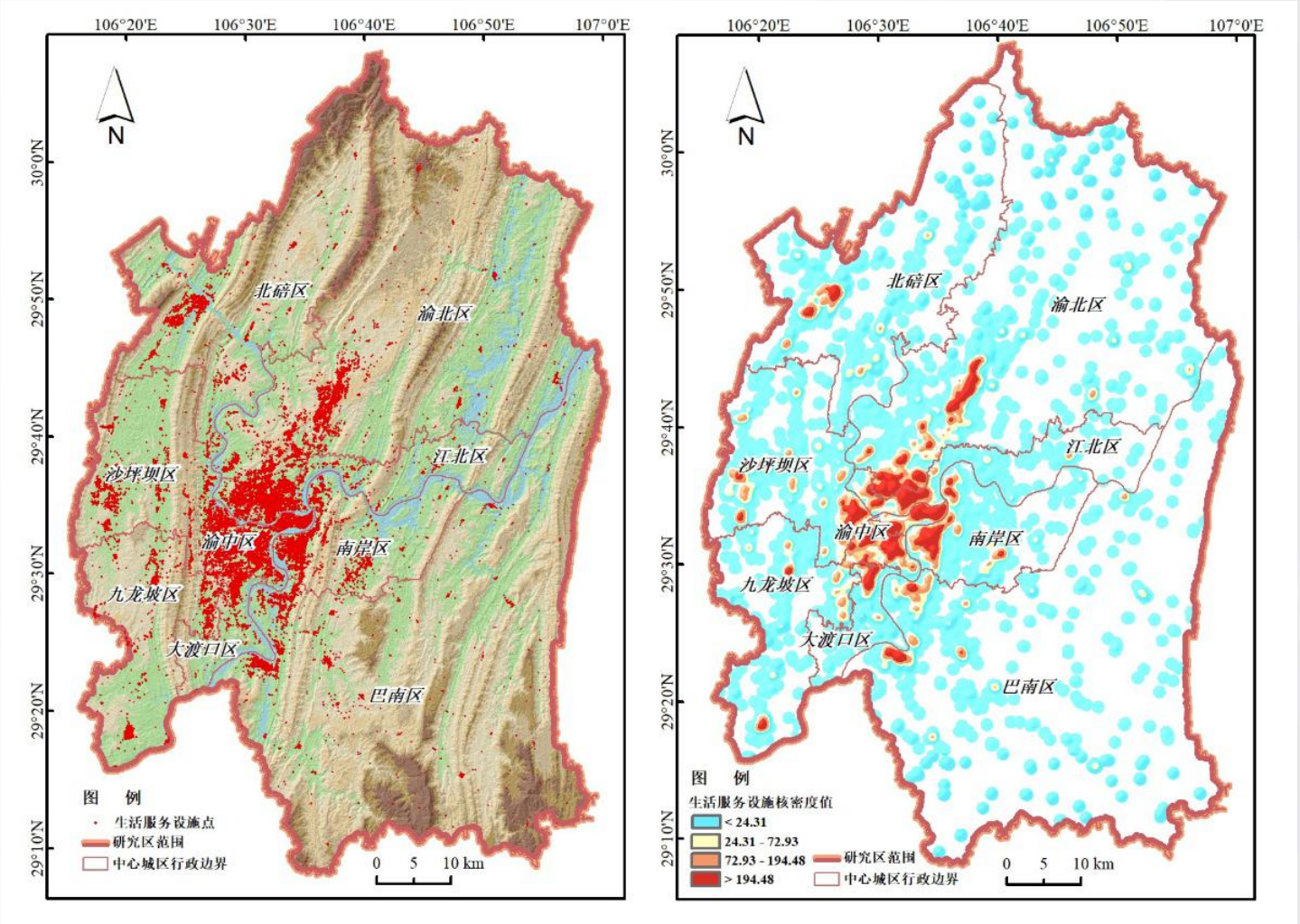
分类	举例	数量/个	百分比
购物服务	商场、超市、便利商店、特色商业街	170167	26.96%
餐饮服务	糕饼馆、咖啡厅、快餐厅、中餐厅	148559	23.53%
生活服务	快递、中介咨询、洗衣美容美发店	103171	16.34%
公司企业	工厂、公司、企业、农林牧渔基地	59352	9.40%
教育培训	学校、图书馆、博物馆、科研机构	23531	3.73%
医疗保健	诊所、专科综合医院、急救中心	22175	3.51%
休闲娱乐	影剧院、运动场馆、度假疗养场所	20911	3.31%
车辆服务	销售店、加油站、洗车场、维修店	20246	3.21%
商务住宅	办公楼宇、住宅区、产业园区	16413	2.60%
政府及社会团体	政府机关、公检法工商税务机构	15436	2.45%
住宿服务	宾馆、酒店、旅馆、招待所	13988	2.22%
金融保险	银行、自动提款机、保险证券公司	10794	1.71%
公共服务	报刊亭、公共厕所、紧急避难场所	6492	1.03%
总计		631235	100%

本研究的POI数据采集于2018年百度地图公司

2.2 多源数据

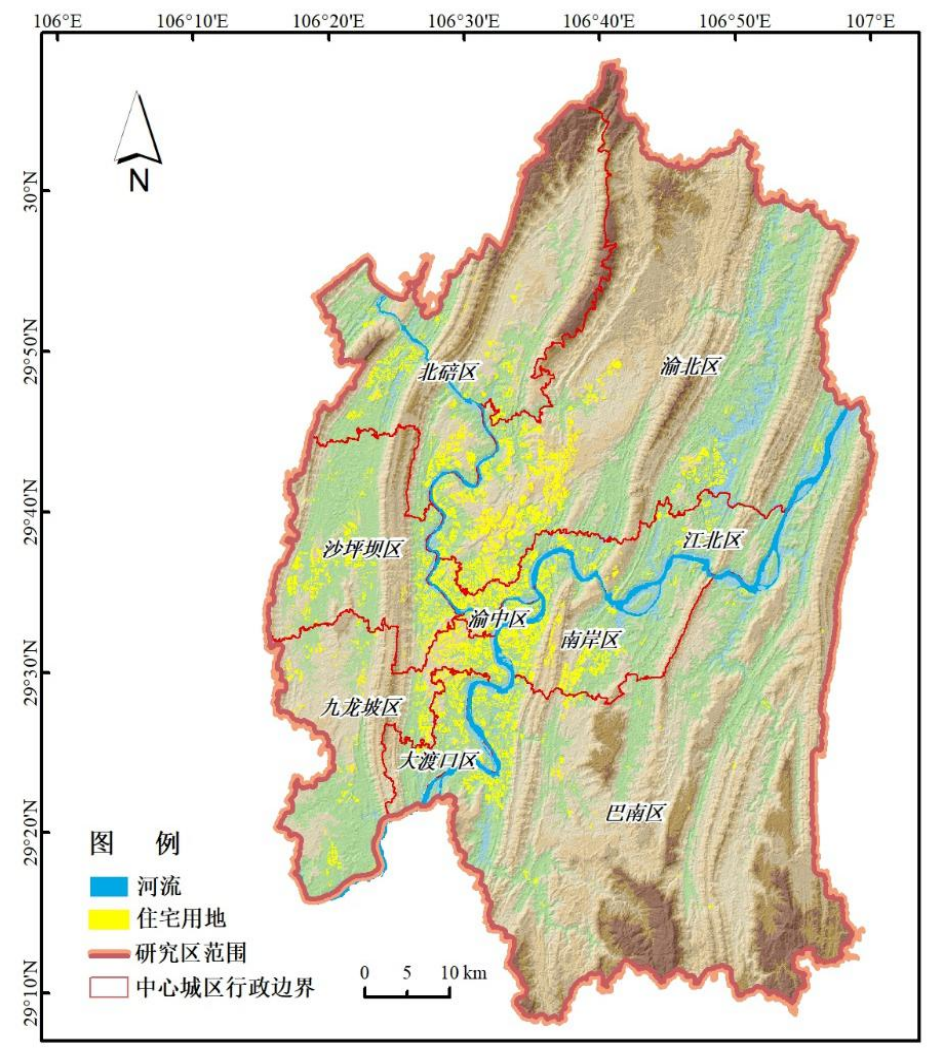
POI核密度分析

核密度优点在于引入了**核函数**，将分布密度随着**距离增加而衰减**的问题考虑其中，更加**符合客观世界中的物质分布**状态。

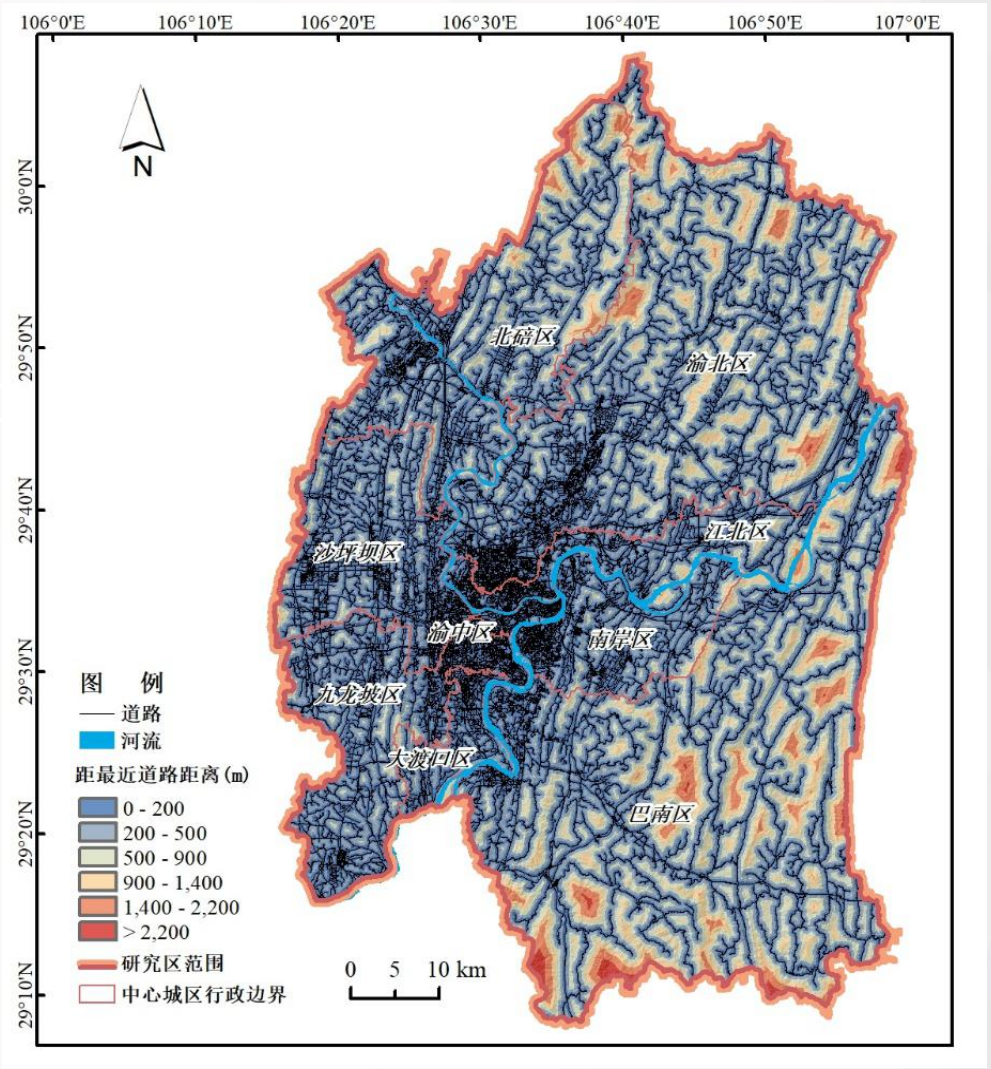


生活服务设施点分布及核密度分析图

2.2 多源数据



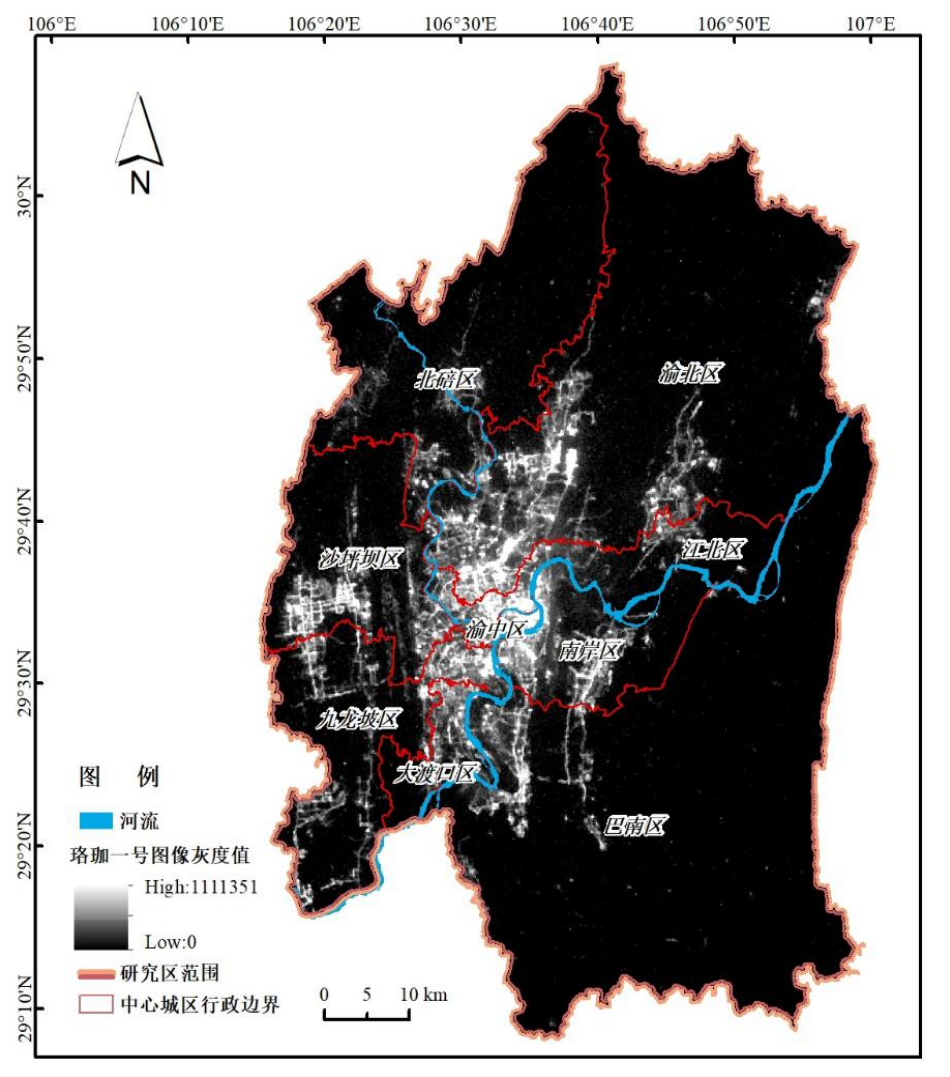
居住用地分布图



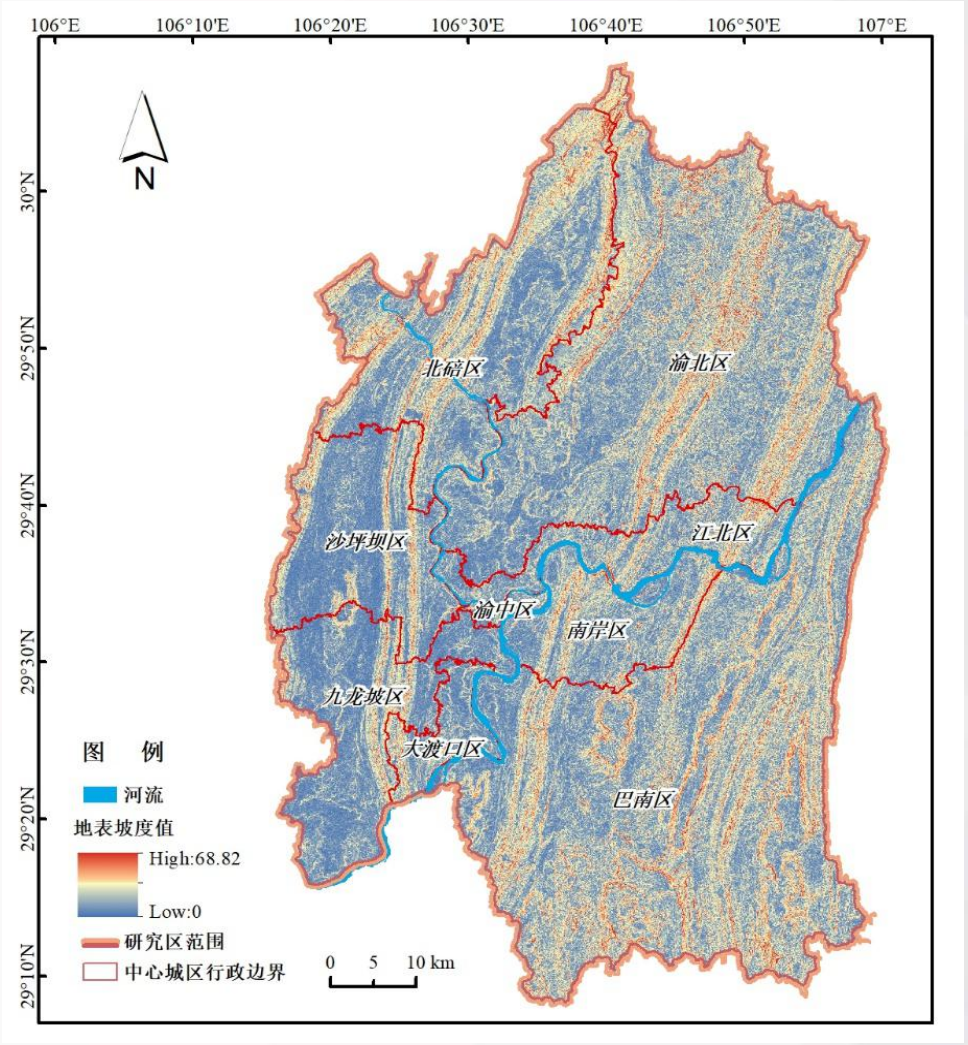
道路及其最短距离分析图

2.2 多源数据

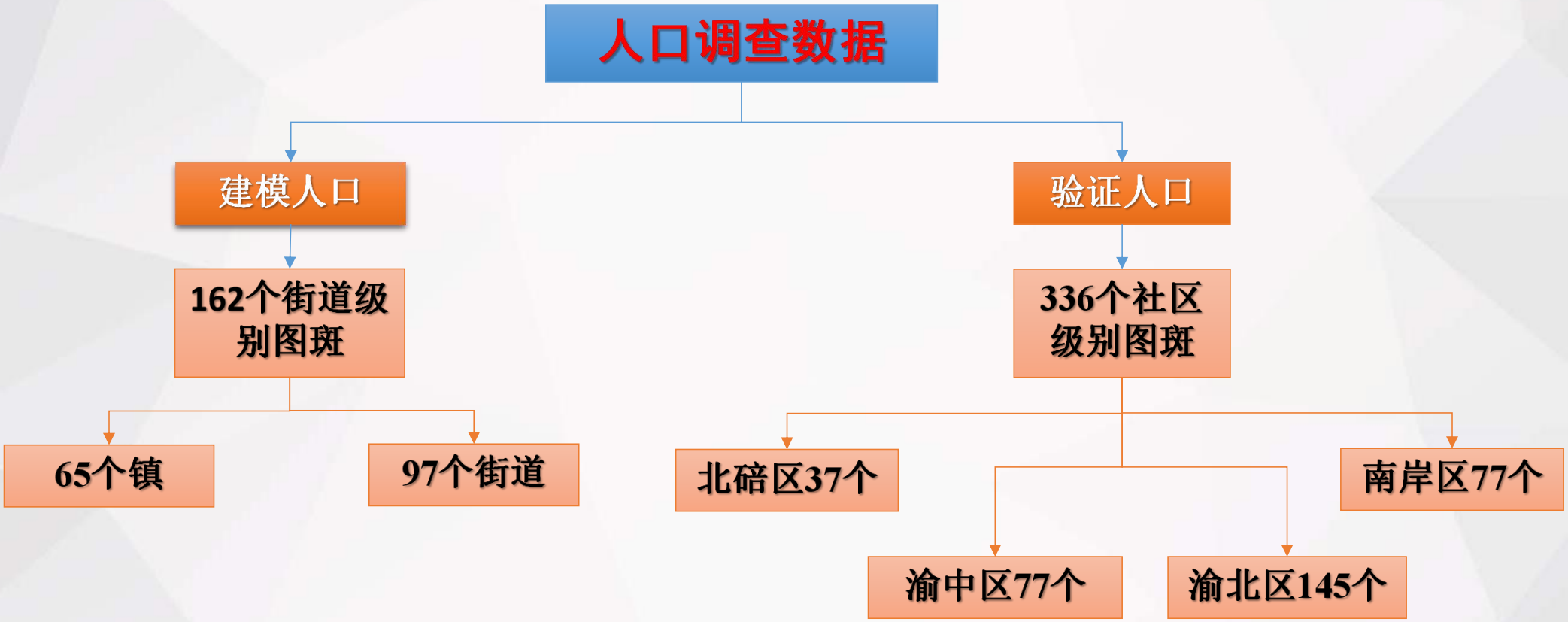
夜间灯光图



地表坡度图

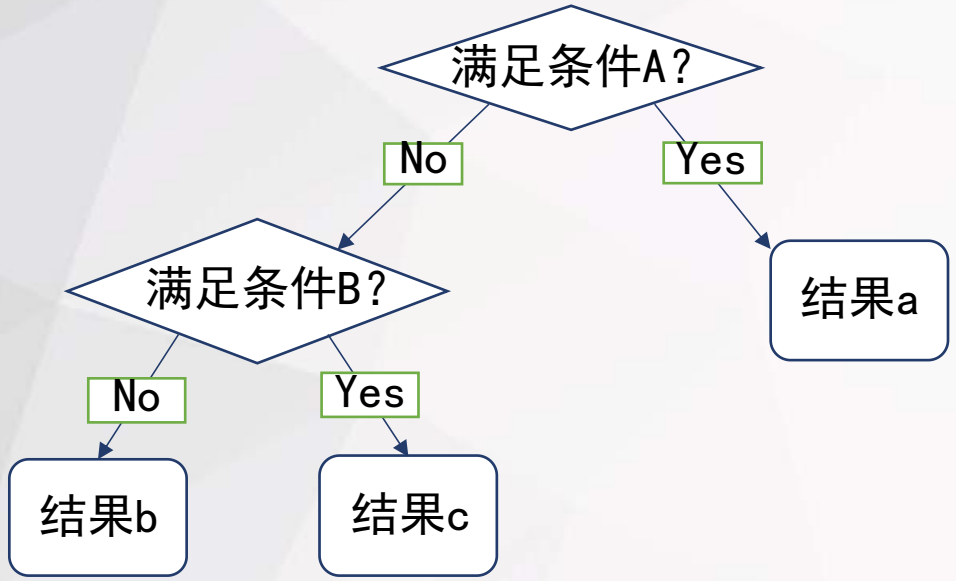


2.2 多源数据



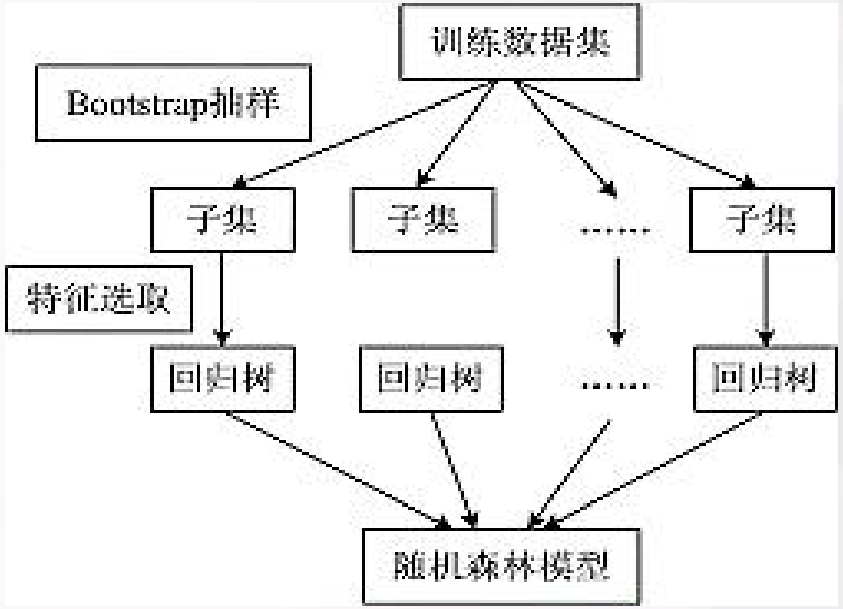
数据来自重庆市公安局和重庆市规划设计研究院

2.3 研究方法



基础决策树

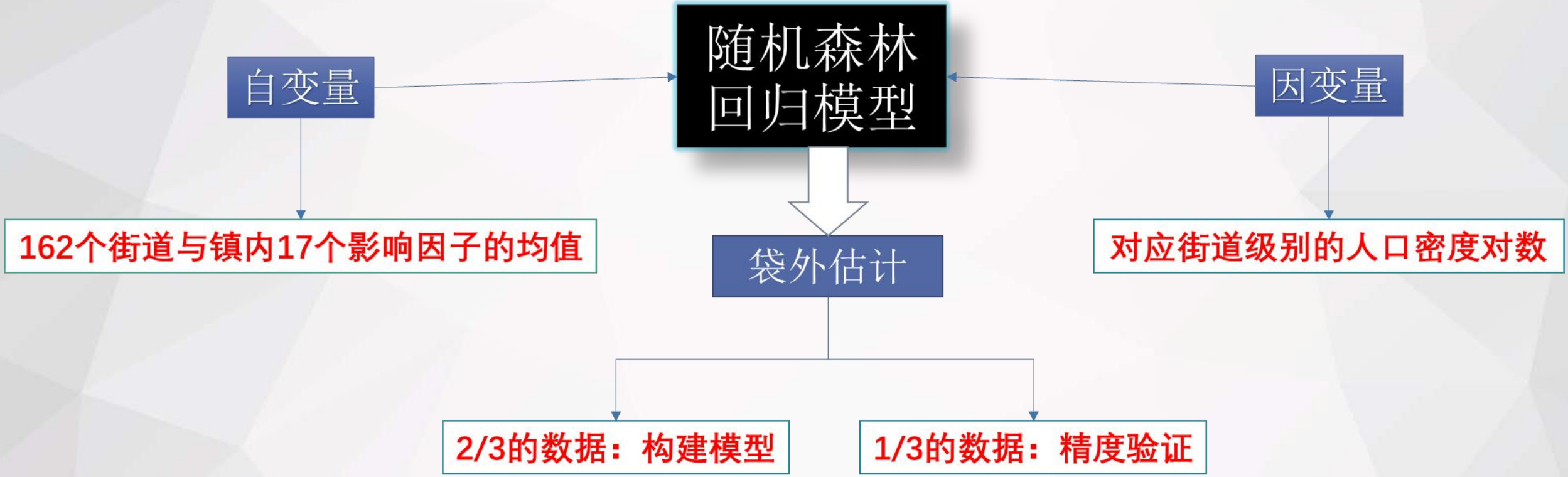
该方法通过询问一系列的**判断问题**来做出决定，是一种通过设定判断条件将**数据集细分**为更小的数据子集来**预测目标值**的方法。



随机森林回归

该方法基于集成的回归决策树，通过计算随机产生的众多的**决策树的平均值**作为结果，是一种可对模型进行训练和预测的**机器学习方法**。

2.3 研究方法

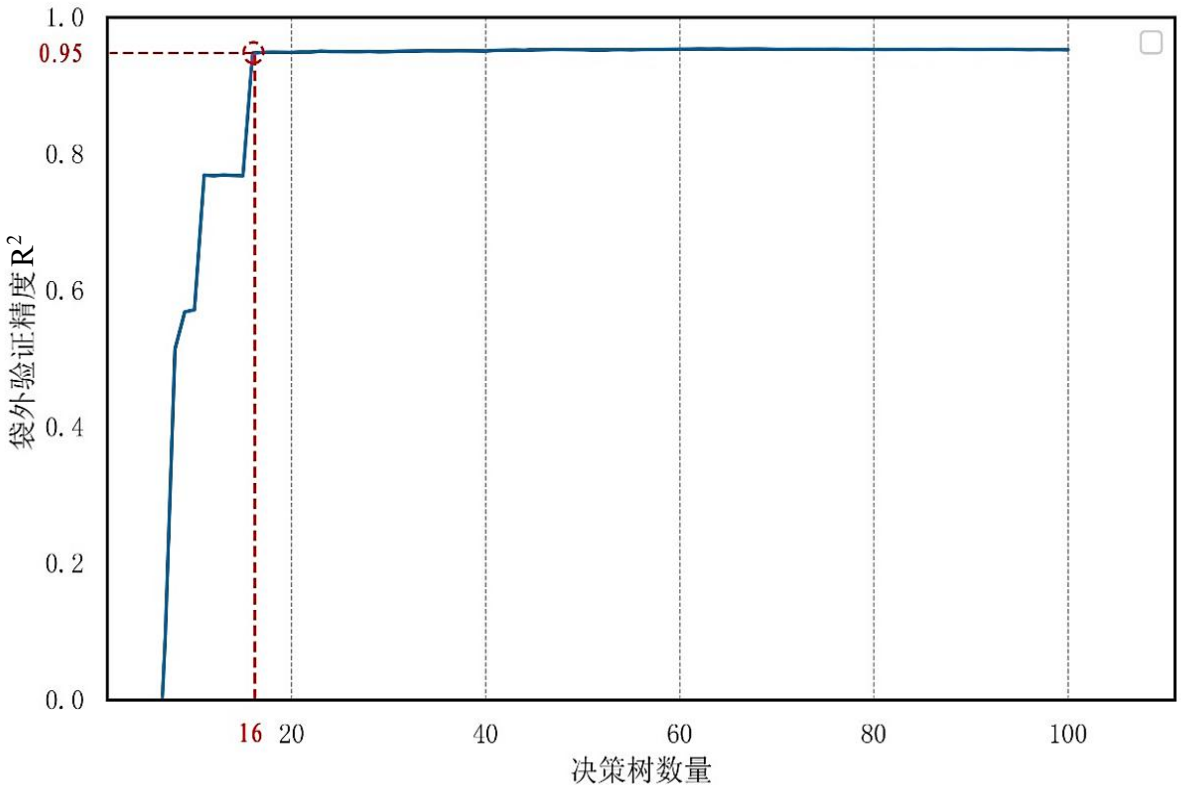


本研究的建模编程语言采用**Python3.7.6**版本，核心算法来自开源机器学习库**scikit-learn**库。

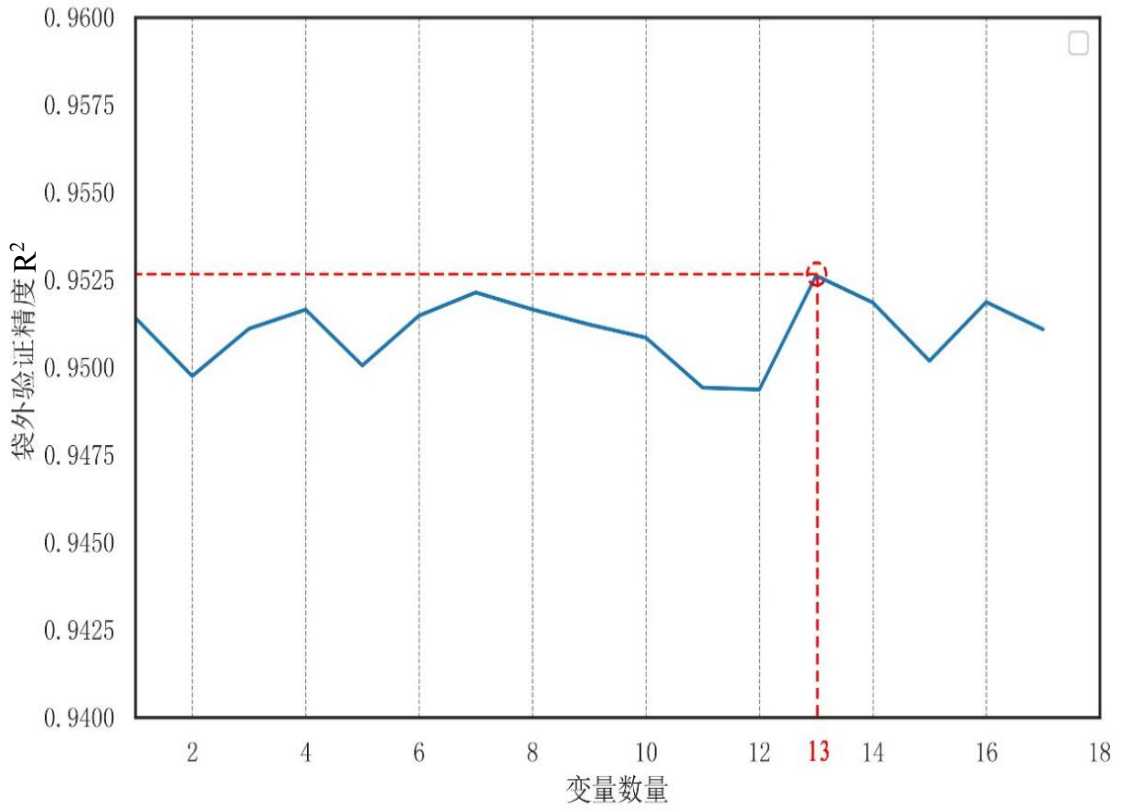
三

建模与结果

3.1 参数调整



以30m格网为例，决策树数量与模型精度关系图



变量数量与模型精度关系图

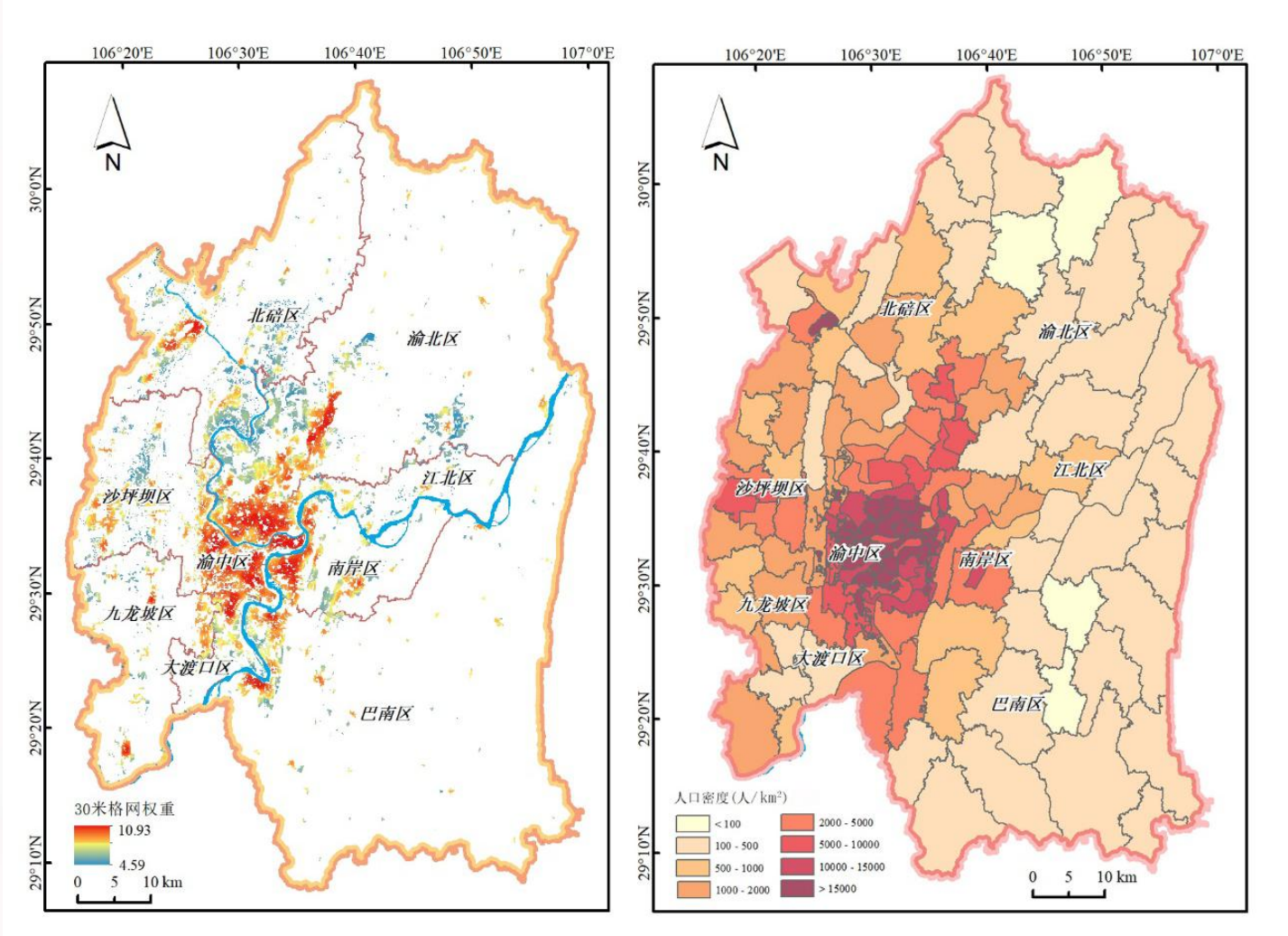
3.2 人口空间化原理

街道或者镇的**总人口数量**

格网在随机森林回归模型中的**预测权重**

$$POP_{格网} = \frac{POP_{街道总人口} \times W_{格网}}{W_{街道总和}}$$

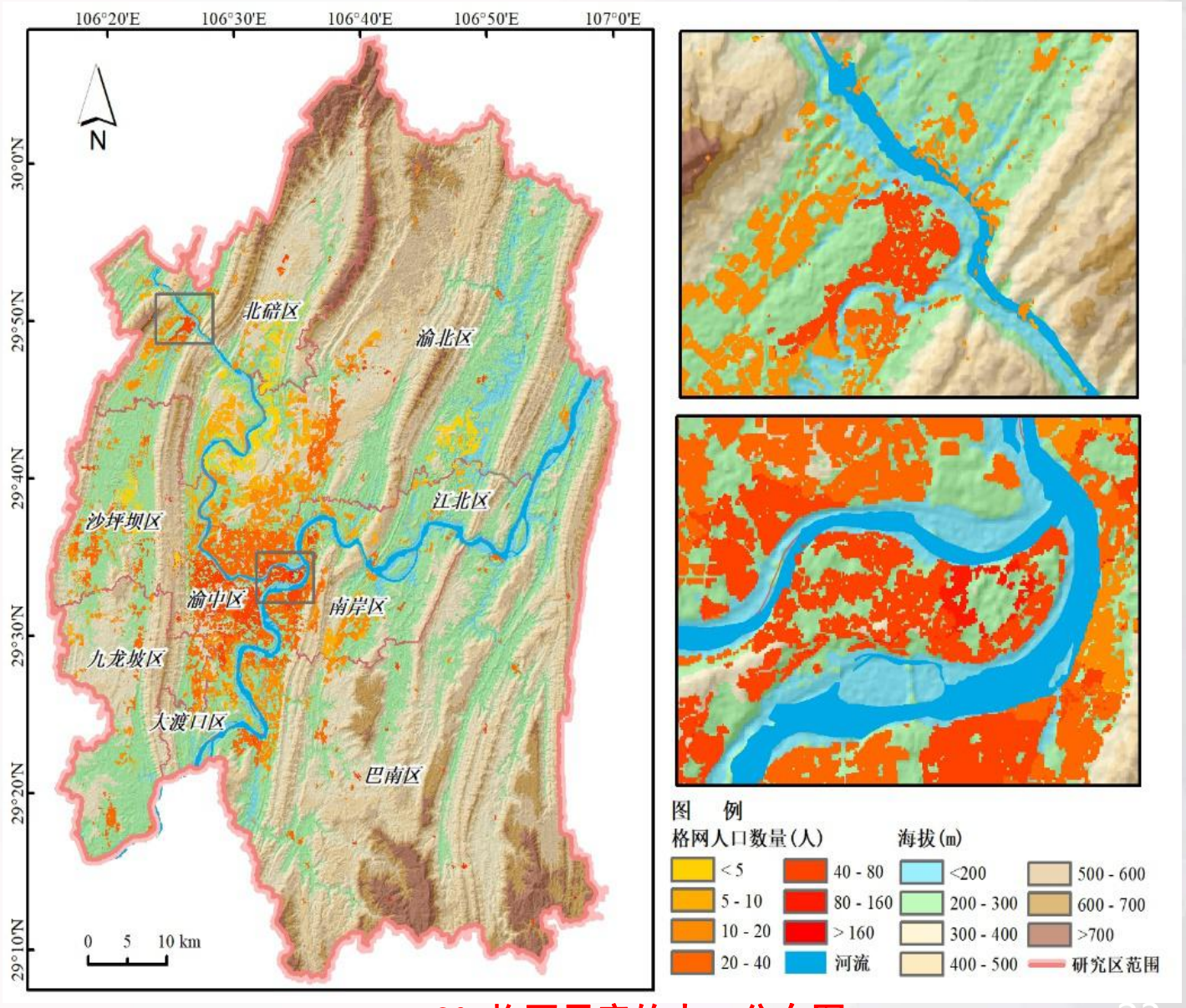
整个街道或镇的所有格网的预测**权重之和**



30m格网尺度的人口分布权重图　　街道级别的人口密度图

3.3 人口空间化结果

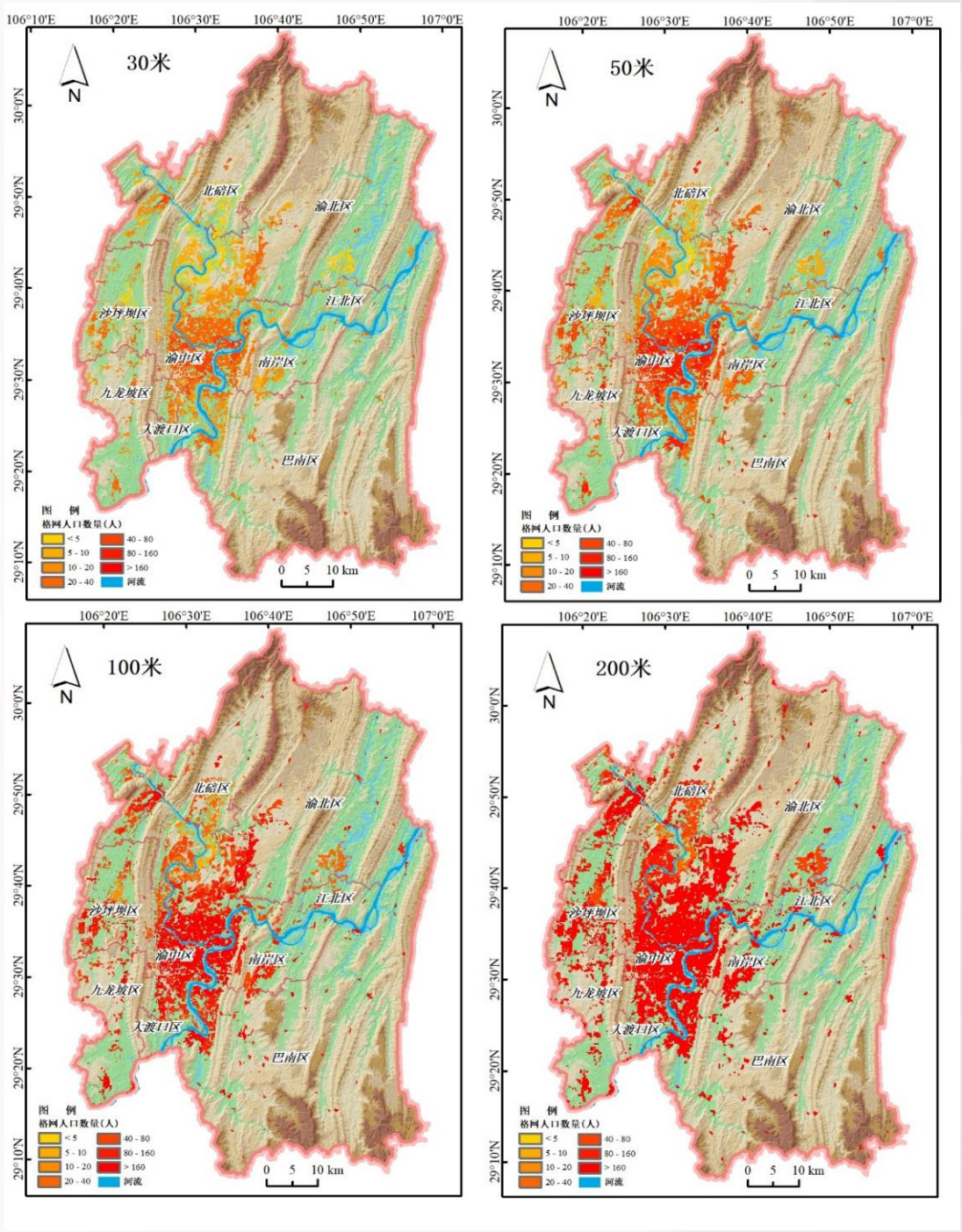
1. 人口主要集中在山脉之间的**槽谷地带**和长江和嘉陵江的**沿江地带**；
2. **人口密度较高**的区域集中在渝中区、渝北区西南部，沙坪坝区东南部，江北区、南岸区和北碚区的西部，巴南区西北部，大渡口区北部；
3. **人口密度较低**的区域分布在渝北区东北部、巴南区东南部和北碚区北部地区；
4. 人口分布特点与主城区“**多中心、组团式**”的城市空间结构相吻合。



30m格网尺度的人口分布图

3.4 人口多尺度空间化

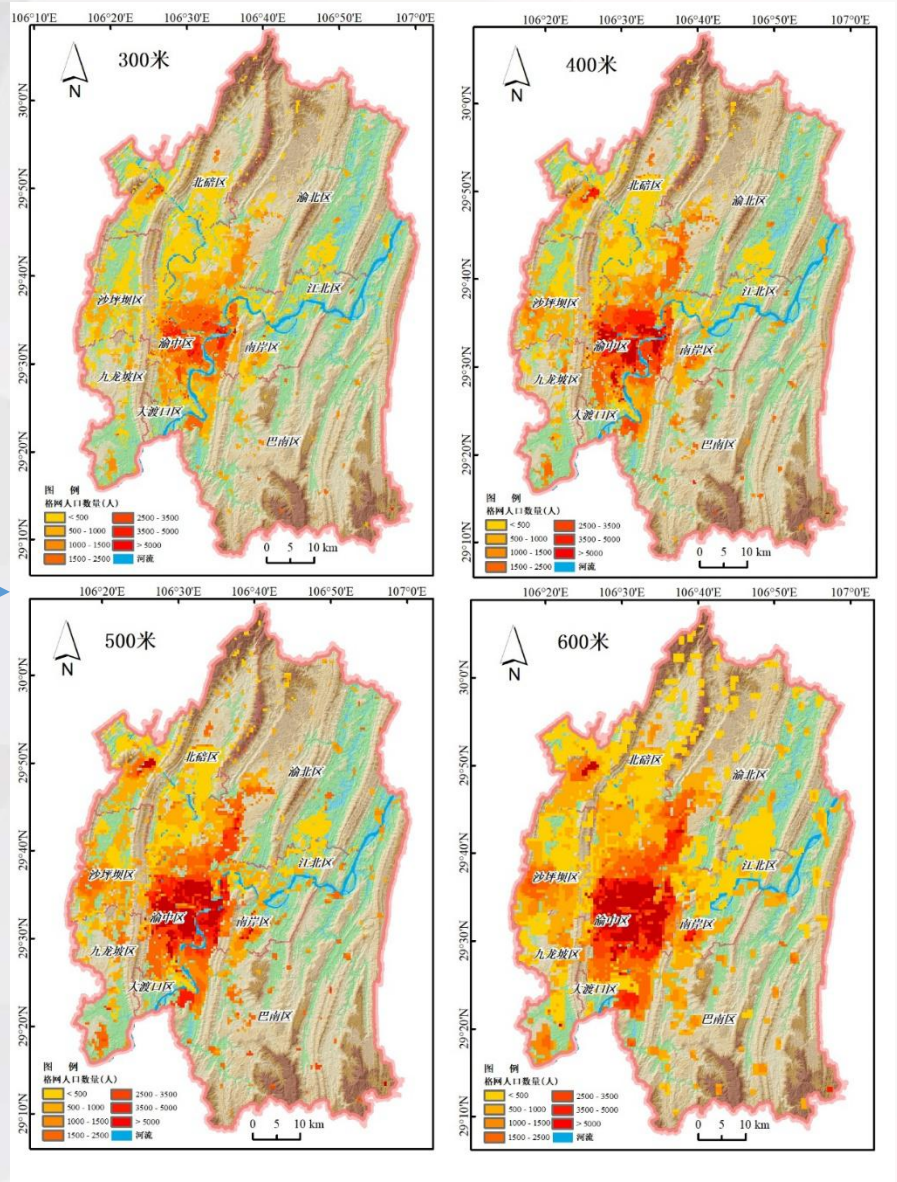
按照上述的30m人口空间化模型构建方法，我们分别对50m、100m、200m、300m、400m、500m、600m、700m、800m、900m、1000m，共12个格网尺度进行了人口空间化建模。



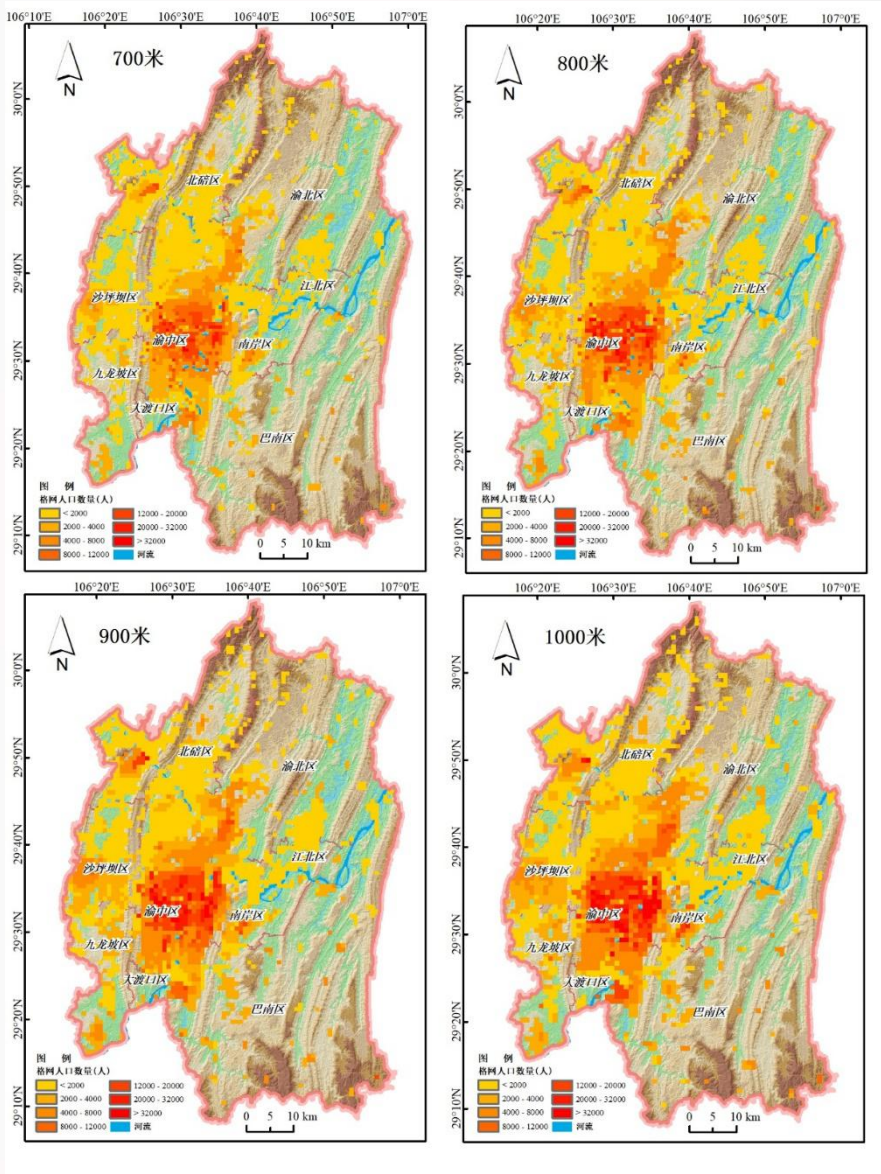
30m、50m、100m和200m格网尺度的人口分布图

3.4 人口多尺度空间化

300m、400m、500m和600m格网尺度的人口分布图



700m、800m、900m和1000m格网尺度的人口分布图



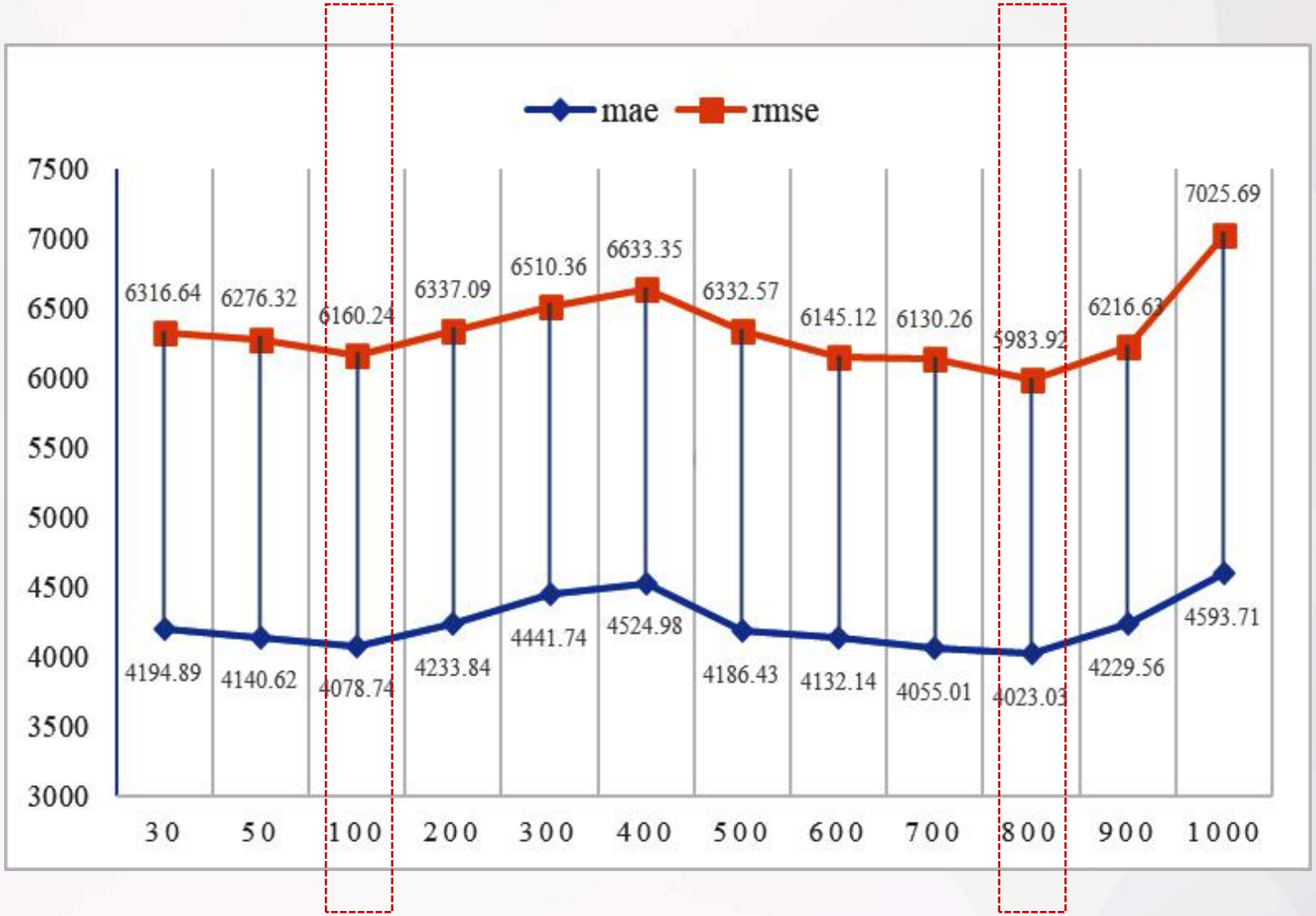
3.5 精度评价

平均绝对误差

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum (R_i - P_i)^2}$$

均方根误差

$$\mathrm{MAE} = \frac{1}{n}\sum |R_i - P_i|$$

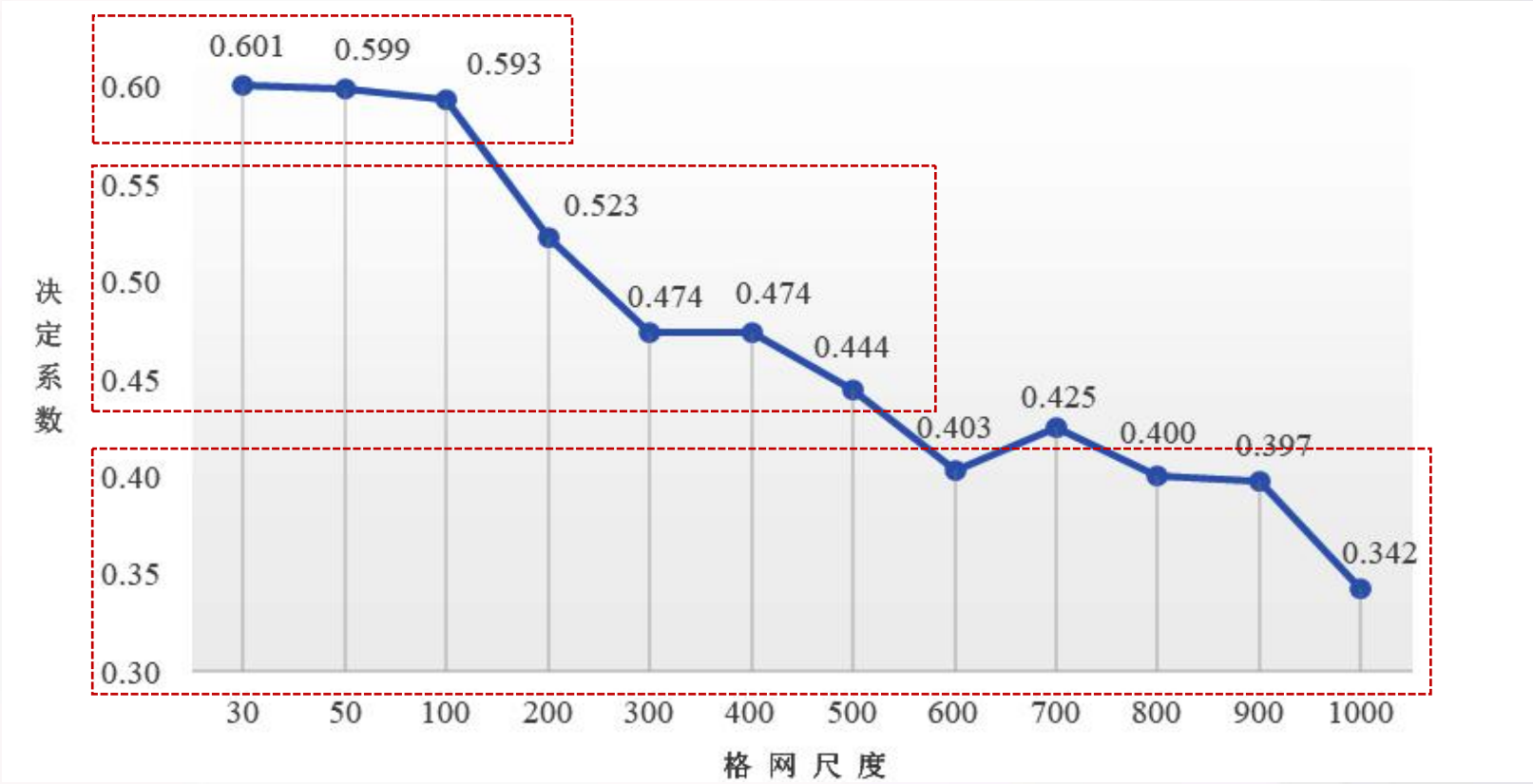


3.5 精度评价

决定系数 R^2

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

R^2越大，表示模型的拟合度越好，可解释程度越高。



取**精度误差较小**且**决定系数较大**的格网作为最佳人口空间化建模格网。本研究**100m格网**是较为适宜重庆市中心城区人口空间化建模的格网单元。

四

因子定量分析

4.1 特征重要性分析

为了解释随机森林黑箱模型，特征重要性（Feature importance）分析常被引入，用于定量解释每个变量对于人口分布的重要程度，该方法可对高精度的人口空间化模型的变量选择提供借鉴（Robinson *et al.*，2017）。

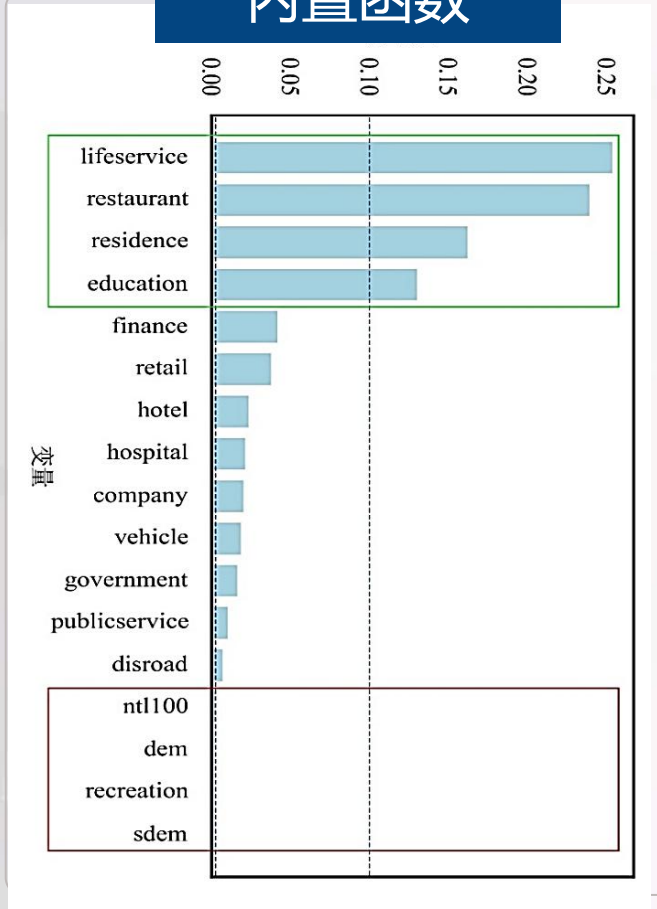
（1）随机森林内置函数计算的模型重要性程度；

（2）采用换位方式计算得到的模型重要性程度；

（3）基于SHAP（SHapley Additive exPlanation）模型的重要性程度。

4.1 特征重要性分析

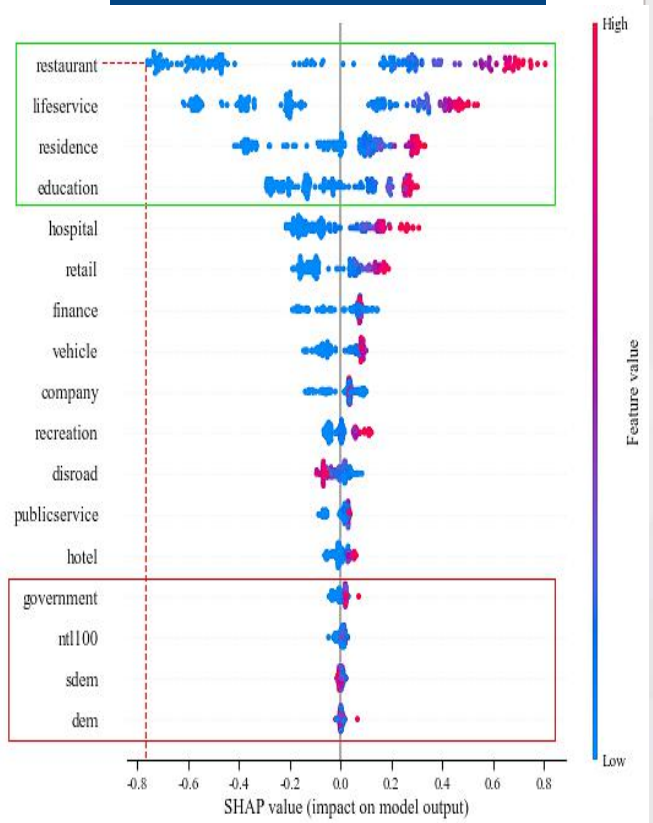
1 内置函数



2 换位方式

Weight	Feature
0.1376 ± 0.0348	restaurant
0.0766 ± 0.0141	lifeservice
0.0288 ± 0.0085	residence
0.0234 ± 0.0049	education
0.0188 ± 0.0046	hospital
0.0091 ± 0.0023	finance
0.0086 ± 0.0016	retail
0.0043 ± 0.0011	company
0.0042 ± 0.0011	vehicle
0.0026 ± 0.0005	disroad
0.0025 ± 0.0005	hotel
0.0023 ± 0.0012	recreation
0.0017 ± 0.0006	ntl100
0.0017 ± 0.0008	publicservice
0.0014 ± 0.0002	dem
0.0011 ± 0.0001	government
0.0008 ± 0.0003	sdem

3 SHAP模型



4.2 部分依赖图分析

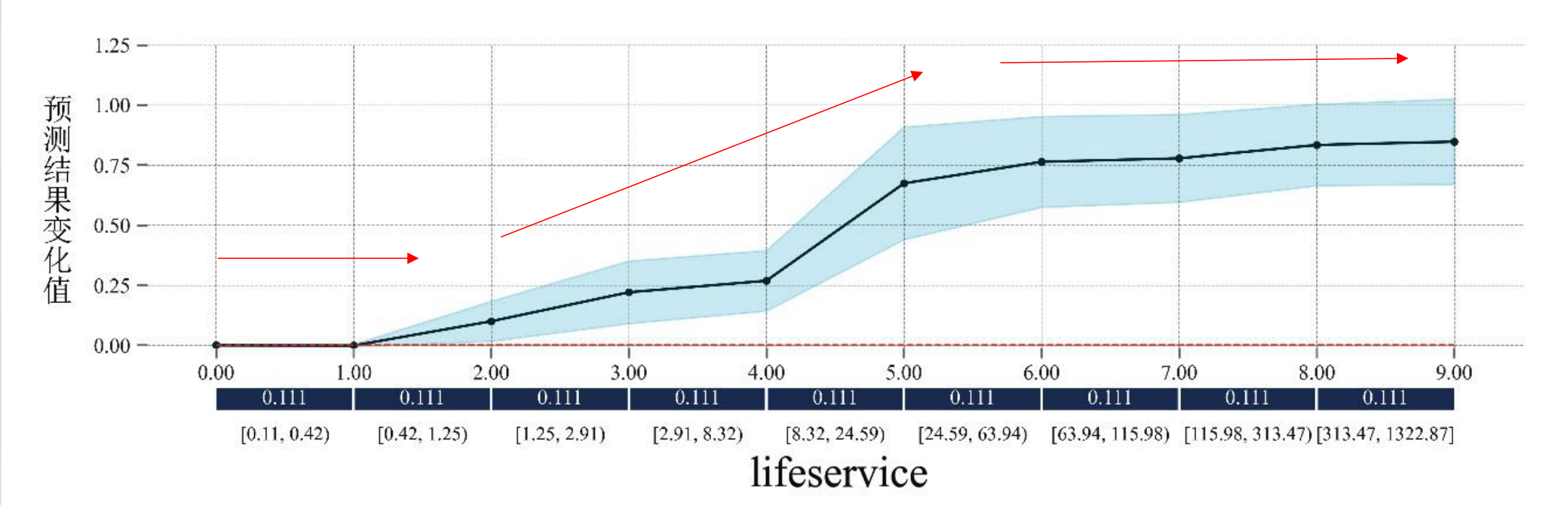
部分依赖图（Partial Dependence Plot, PDP）显示了特征变量在机器学习过程中对于预测结果的边际影响（Friedman，2001）。

$$\hat{f}_{xS}(xS) = E_{xC}\left[\hat{f}(xS,\ xC)\right] = \int \hat{f}(xS,\ xC)dP(xC)$$

特征向量xS和xC组成整个向量空间x。 xS代表了偏相关图中被呈现变量，xC代表了其他在机器学习模型$\hat{f}$中用到的变量。

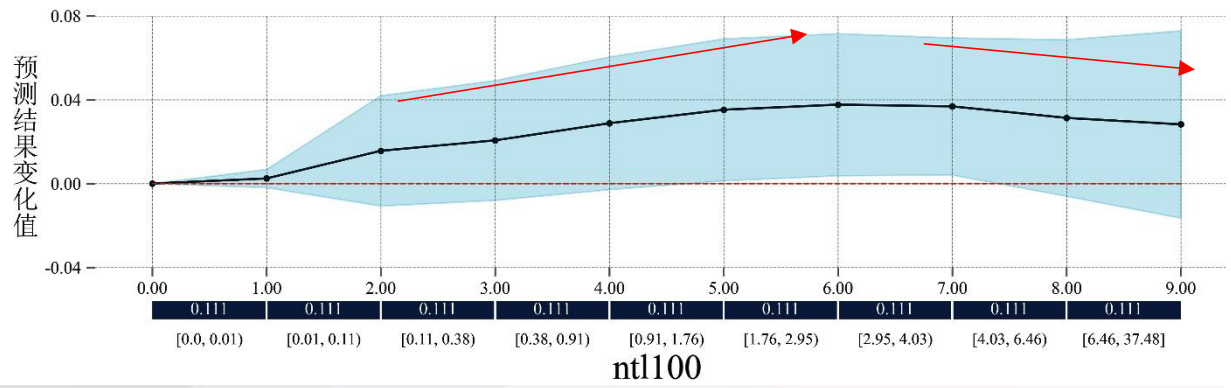
部分依赖图的计算机制 ： 在**集合C**的特征分布上**边缘化**机器学习模型输出，以便该让**函数显示**我们感兴趣的**集合S**中的特征与预测结果之间的关系。

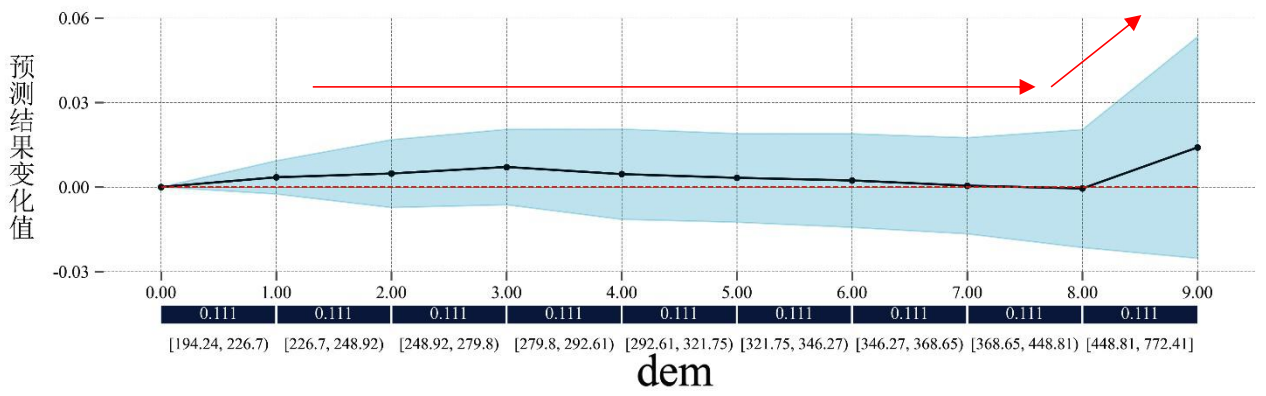
4.2 部分依赖图分析

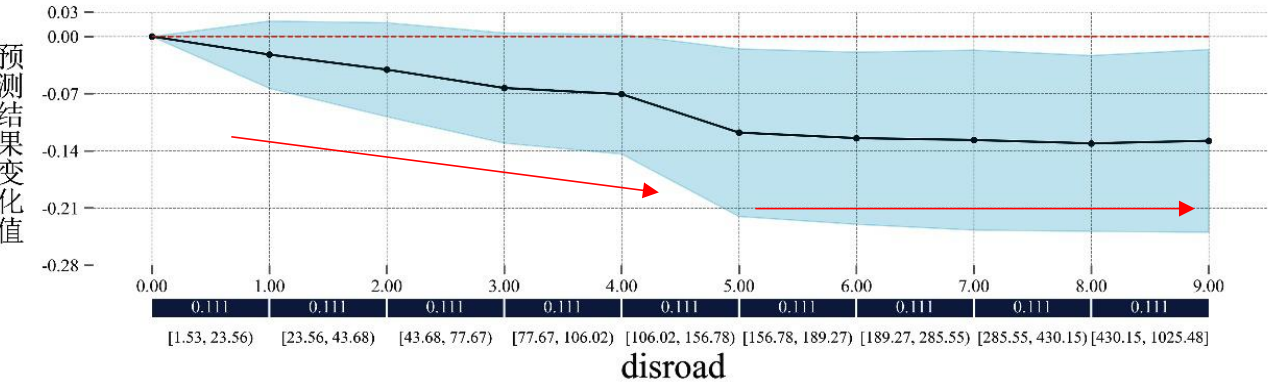


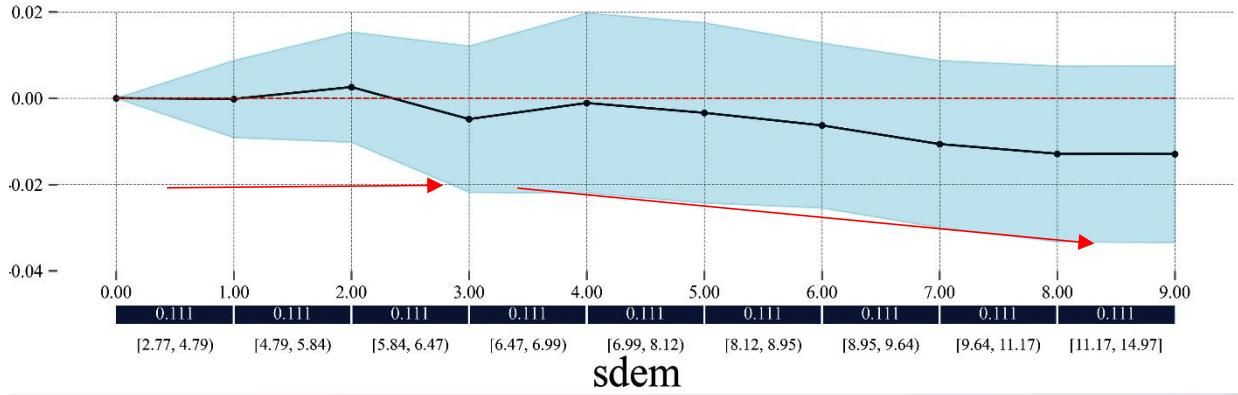
- **横坐标**表示生活服务设施的**取值范围**，带有0.111的深蓝色底的**9个矩形对应这9等份数据**；
- **纵坐标**为预测结果的**变化值**，即偏相关函数计算结果；
- 深蓝色的曲线为**变化曲线图**，在其周围的浅蓝色的区域为变化值可能**波动的范围**。

4.2 部分依赖图分析


预测结果变化值
0.08
0.04
0.00
-0.04
0.00 1.00 2.00 3.00 4.00 5.00 6.00 7.00 8.00 9.00
0.111 0.111 0.111 0.111 0.111 0.111 0.111 0.111 0.111
[0.0, 0.01) [0.01, 0.11) [0.11, 0.38) [0.38, 0.91) [0.91, 1.76) [1.76, 2.95) [2.95, 4.03) [4.03, 6.46) [6.46, 37.48]
ntl100


预测结果变化值
0.06
0.03
0.00
-0.03
0.00 1.00 2.00 3.00 4.00 5.00 6.00 7.00 8.00 9.00
0.111 0.111 0.111 0.111 0.111 0.111 0.111 0.111 0.111
[194.24, 226.7) [226.7, 248.92) [248.92, 279.8) [279.8, 292.61) [292.61, 321.75) [321.75, 346.27) [346.27, 368.65) [368.65, 448.81) [448.81, 772.41]
dem


预测结果变化值
0.03
0.00
-0.07
-0.14
-0.21
-0.28
0.00 1.00 2.00 3.00 4.00 5.00 6.00 7.00 8.00 9.00
0.111 0.111 0.111 0.111 0.111 0.111 0.111 0.111 0.111
[1.53, 23.56) [23.56, 43.68) [43.68, 77.67) [77.67, 106.02) [106.02, 156.78) [156.78, 189.27) [189.27, 285.55) [285.55, 430.15) [430.15, 1025.48]
disroad


0.02
0.00
-0.02
-0.04
0.00 1.00 2.00 3.00 4.00 5.00 6.00 7.00 8.00 9.00
0.111 0.111 0.111 0.111 0.111 0.111 0.111 0.111 0.111
[2.77, 4.79) [4.79, 5.84) [5.84, 6.47) [6.47, 6.99) [6.99, 8.12) [8.12, 8.95) [8.95, 9.64) [9.64, 11.17) [11.17, 14.97]
sdem

五 结论与展望

5.1 四点结论

主要结论

（1）多源数据的选取融合

通过对**13个多源数据**进行**栅格化和重采样**处理，加入到随机森林回归模型中，以此融合多源数据对人口普查数据进行降尺度。

（2）人口分布的适宜尺度

通过对比**12个空间分辨率**，100m格网下精度误差较小且拟合精度较高（R^2=0.59，$p < 0.01$），本研究发现重庆市主城区的**适宜网格**是100m。

（3）特征重要性分析

起**主要影响**作用的为生活设施、餐饮设施、居住点和教育**设施点核密度**，而**自然地理因素**高程和坡度、道路距离及夜间灯光亮度值对于人口分布结果**影响较小**。

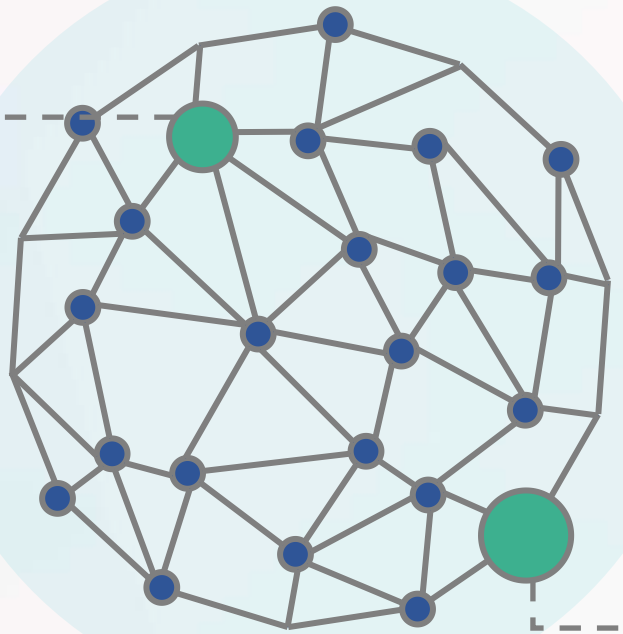
（4）部分依赖图分析

POI核密度对人口分布起**正向**影响，而高程、坡度因子与道路距离与人口分布呈现**负相关**关系。夜间灯光主要为正向影响，也有小部分有出现负向的趋势。

5.2 未来展望

（1）人口数据的时空动态建模

由于人口及建模因子数据源的限制，本文主要使用2018年的人口调查数据，因此该方法主要适用于**静态人口**分布。借助于更高时空分辨率的多源地理位置大数据，人口数据的**时空动态建模**有望取得突破性进展。



（2）人口数据建模的方法对比与推广

随着2021年我国**第七次人口普查**结果的发布，在后续研究中可利用本文的研究方法对最新的**全国人口分布结果**进行更深入的探索。未来可以考虑将本研究的方法和其他机器学习的**方法比较**，得出精度更高、更加实用可靠的人口空间化方法。

基于随机森林算法的城市人口多尺度空间化研究

汇报人：周云　指导老师：马明国教授

报告完毕　|　敬请指教