



# 市政大数据、混合城市功能 及智慧电网

---

姚尧 博士, 副教授

地理与信息工程学院, 地图制图学与地理信息工程

阿里巴巴集团, 达摩院, 访问学者

Email: [yaoy@cug.edu.cn](mailto:yaoy@cug.edu.cn)

办公地点: 未来城校区地信楼522办公室



# CONTENT

- 01 市政大数据
- 02 决策树与森林算法
- 03 基于市政水耗剖析城市空间结构
- 04 智慧电网感知城市社会经济发展



High-performance Spatial Computational Intelligence Lab @ CUG

# 01

---

## 市政大数据

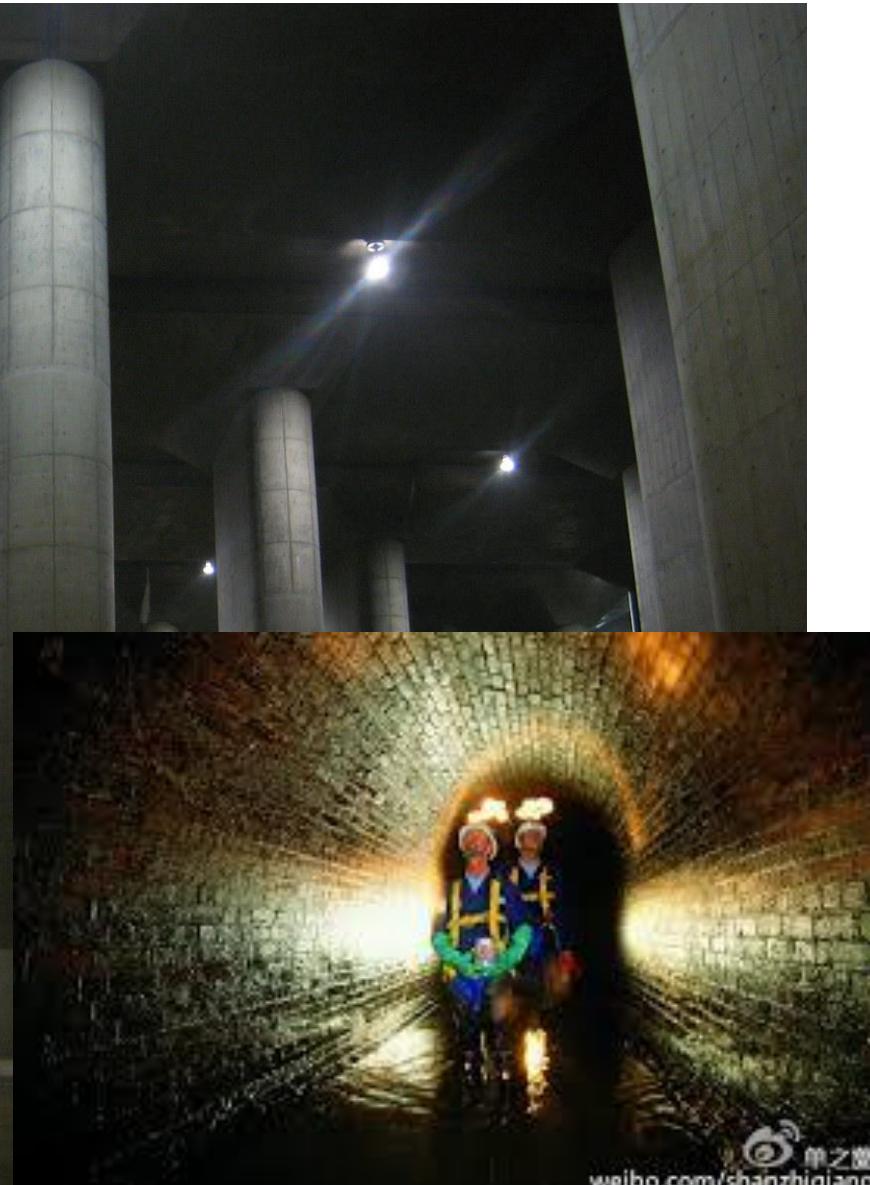


High-performance Spatial Computational Intelligence Lab @ CUG



市政基础设施：由国家城市建设行政主管部门分工进行行业管理、具体由城市政府组织实施管理的部分城市基础设施。其中包括：

- 城市公用事业：城市供水、供气、供热、公共交通等。
- 市政工程：城市道路、排水（包括污水处理）、防洪、照明等。
- 市容环境卫生事业：城市市容、公共场所保洁、垃圾和粪便清运处理、公共厕所等。
- 园林绿化业：城市园林、绿化等。



## ■ 城市市政基础设施建设“十二五”四大主要成就

- 市政设施能力普遍提高，支撑了城镇化快速发展  
投资95万亿，比“十一五”时期投资增长近90%  
  
• 低碳绿色智慧理念创新，引领市政基础设施转型发展  
海绵城市、地下综合管廊、智慧城市试点
  
- 基本公共服务水平稳步提高，惠及更多城乡居民  
公共供水普及率达93.1%（85.1%）  
污水处理率达91.9%（85.2%）  
燃气普及率达95.3%（75.9%）  
生活垃圾无害化处理率达94.1%（79.0%）  
  
• 创新建设和运营模式，激发市场主体活力  
出台政府和社会资本合作（PPP）的相关政策  
鼓励开展政府和社会资本合作

## ■ 城市市政基础设施建设面临**四大问题**

- 投入不够，总量不足
  - 历史欠账巨大
  - 城市市政基础设施投资占比持续下降
  - 市政基础设施服务需求持续扩大
- 设施水平偏低，“城市病”问题突出
  - 城市内涝
  - 水体黑臭
  - 交通拥堵
  - “马路拉链”
  - “垃圾围城”
  - 地下管线安全事故
  - .....
- 发展不均衡，服务水平差异较大
  - 中西部地区总体上仍落后于东部地区
  - 老城区明显低于城市新区
- 产业集中度低，服务效率和质量参差不齐
  - 市政公用企业“小、散、弱、差”



## ■ 城市市政基础设施建设“十三五”展望

加强市政公用设施和公共服务设施建设，增加基本公共服务供给，增强对人口集聚和服务的支撑能力。

市政基础设施是新型城镇化的物质基础，也是“实现1亿左右农业转移人口和其他常住人口在城镇落户，完成约1亿人居住的棚户区和城中村改造，引导约1亿人在中西部地区就近城镇化”（“三个1亿人”）城镇化目标的重要保障。

The screenshot shows the homepage of the Chinese Government website (www.gov.cn). The top navigation bar includes links for the State Council, Premier, News, Policies, Interaction, Services, Data, National Conditions, and the National Government Service Platform. A search bar is also present. Below the navigation, a banner for the 'National Plan for New Urbanization (2014-2020)' is displayed, featuring the title in red and the document number '中共中央 国务院印发《国家新型城镇化规划(2014—2020年)》'. The main content area contains the full text of the plan, which is partially visible.



## ■ 城市市政基础设施建设“十三五”展望

### 十二大规划任务

- 加强道路交通系统建设，提高交通综合承载能力
- 推进城市轨道交通建设，促进居民出行高效便捷
- 有序开展综合管廊建设，解决“马路拉链”问题
- 构建供水安全多级屏障，全流程保障饮用水安全
- 全面整治城市黑臭水体，强化水污染全过程控制
- 建立排水防涝工程体系，破解“城市看海”难题
- 加快推进海绵城市建设，实现城市建设模式转型
- 优化供气供热系统建设，提高设施安全保障水平
- 完善垃圾收运处理体系，提升垃圾资源利用水平
- 促进园林绿地增量提质，营造城乡绿色宜居空间
- 全面实施城市生态修复，重塑城市生态安全格局
- 推进市政设施智慧建设，提高安全运行管理水平

**中华人民共和国住房和城乡建设部**  
Ministry of Housing and Urban-Rural Development of the People's Republic of China (MOHURD)

2020年11月12日 星期四 检索 工作邮箱: 用户名 密码 登录 设为首页 收藏本站

您现在的位置: 首页>政策发布

索引号: 00001338/2017-00094	主题信息: 城市建设
发文单位: 中华人民共和国住房和城乡建设部 中华人民共和国国家发展和改革委员会	生成日期: 2017年05月17日
文件名称: 住房城乡建设部 国家发展改革委关于印发全国城市市政基础设施建设“十三五”规划的通知	有效期限:
文 号: 建城[2017]116号	主题词:
废改立情况:	

**住房城乡建设部 国家发展改革委关于印发全国城市市政基础设施建设“十三五”规划的通知**

各省、自治区、直辖市人民政府，新疆生产建设兵团：

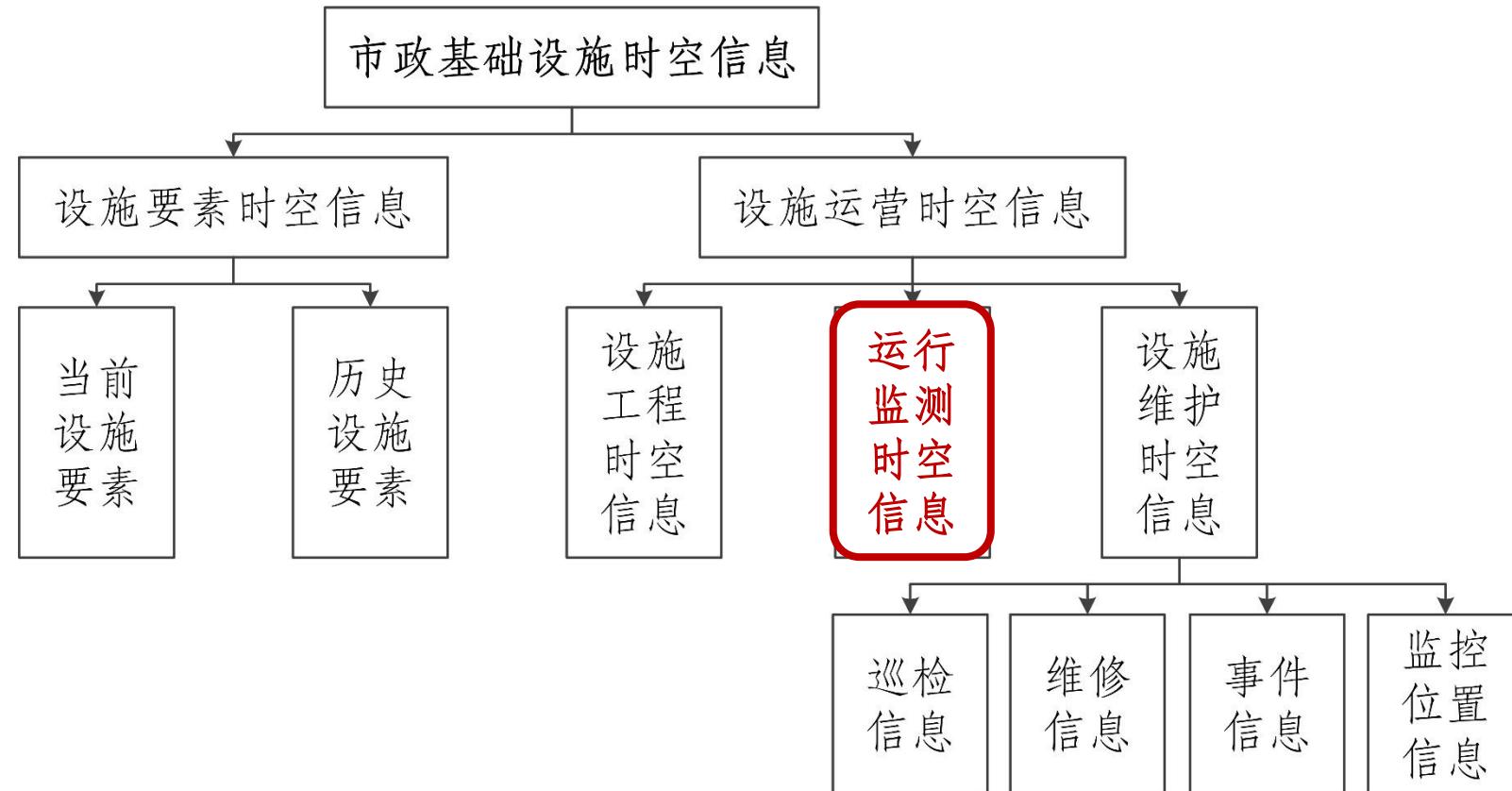
经国务院同意，现将《全国城市市政基础设施建设“十三五”规划》印发给你们，请认真组织实施。

中华人民共和国住房和城乡建设部  
中华人民共和国国家发展和改革委员会  
2017年5月17日

(此件主动公开)

关闭窗口 打印本页

附件下载: 1、全国城市市政基础设施建设“十三五”规划



## ■供水运营时空数据

包括应收帐流水号、用户编号、用户名称、用水类别、抄表日期、算费日期、上次读数、本次读数、本次水量等。

- 数据源：水务单位
- 数据获取：产学研结合，校企合作
- 时间分辨率：以单月/双月为主
- 空间分辨率：每个家庭/企事业




ATE	REC_LASTCODE	REC_CURRCODE
0	0	
0	0	
0	4	
4	8	
8	20	
20	34	
34	57	
57	77	
77	101	

RN	FIRSTNAME	MIDDLENAME	ADDRESS	APPOINTMENT	ACCOUNTNAME	NETTRENT	READ_DATE	REC_CREATE_DATE	REC_LASTCODE	REC_CURRCODE	WATERUSE
281							1 2006-01-21	2006-01-23	9	15	6
294							1 2006-03-21	2006-03-21	15	34	19
301							1 2006-05-22	2006-05-22	34	56	22
316							1 2006-07-18	2006-07-18	56	81	25
338							1 2006-09-21	2006-09-22	81	118	37
372							1 2006-11-20	2006-11-21	118	161	43
372							1 2007-01-20	2007-01-22	161	188	27
391							1 2007-03-22	2007-03-22	188	200	12
391							1 2007-05-18	2007-05-21	200	207	7
679							1 2007-07-18	2007-07-23	207	222	15

## ■ 变电站时空数据

- 数据源：国家电网的内网系统
- 数据获取：产学研结合，校企合作
- 时间分辨率：15min或30min
- 空间分辨率：每个变电站

表1.1 变电站站点信息数据样例

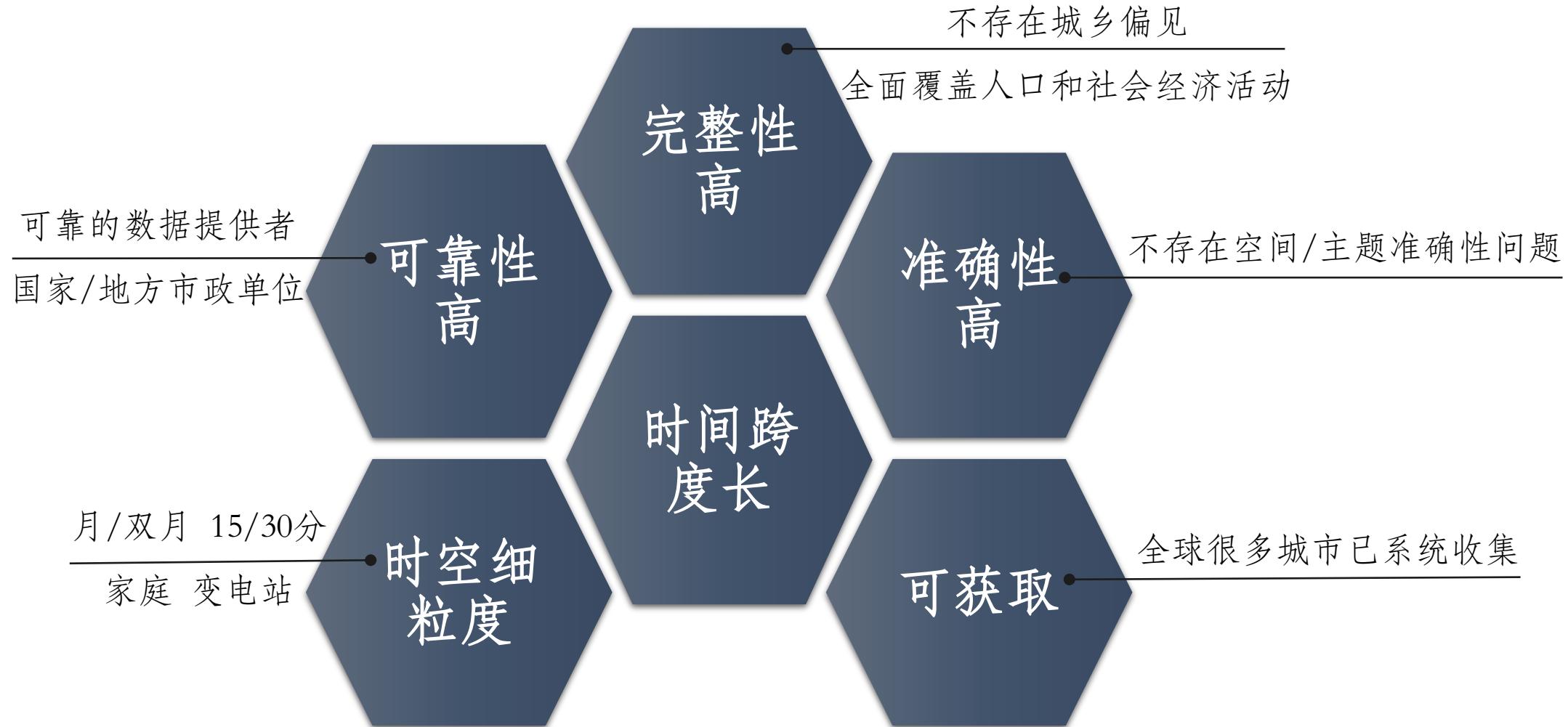
市公司	县公司	台区 编号	变电站 名称	线路名称	配变名称
国网南昌供 电公司	国网江西 南昌县	23**46	黄湖35kV变 电站		
国网南昌供 电公司	国网江西 南昌县	23**44	黄湖35kV变 电站		
国网南昌供 电公司	国网江西 南昌县	23**43	黄湖35kV变 电站		
:	:	:	:		

屏蔽

表1.2 变电站电力消耗数据样例

站点号	采集时间	结束采集时				
23**30	2018/1/1 0:00	2018/1/2 15:5				
13**58	2018/1/1 0:00	2018/1/2 15:5				
10**96	2018/1/1 0:00	2018/1/2 15:5				
:	:	:				

屏蔽



运行监测时空信息为**客观衡量和建模活动模式的性质和原因**提供了新机会  
*(Lansley, Li, & Longley, 2017; Lansley & Longley, 2016)*

## ■ “脏” 数据及数据清洗

		00000
CUSTID	CUSTNAME	00000
0000004273	黄伯铭	KN
0000004273	黄伯铭	134462
0000004273	黄伯铭	160729
0000004273	黄伯铭	179325
0000004273	黄伯铭	192084
0000004273	黄伯铭	211925
0000004273	黄伯铭	242294
0000004273	黄伯铭	262550
0000004273	黄伯铭	276063
0000004273	黄伯铭	7770
0000004273	黄伯铭	298193
0000004273	黄伯铭	304870
0000004273	黄伯铭	342947
0000004273	黄伯铭	363222
0000004273	黄伯铭	376223
0000004273	黄伯铭	428692
0000004273	黄伯铭	421130
0000004273	黄伯铭	467717
0000004273	黄伯铭	474717
0000004273	黄伯铭	515160
0000004273	黄伯铭	512667
		567135

ICMONTH	REC_LASTCODE
39912	6087
00001	6703
00001	6703
00002	7265
00003	7797
00004	8491
00005	9157
00006	9379
00008	9479
00008	9595
00009	9827
00011	9595
00101	500
00101	500
00101	670
00101	670
00102	756
00102	756

# 屏蔽

## 数据的完整性

## 数据的唯一性

## 数据的权威性

## 数据的合法性

## 数据的一致性<sup>14</sup>

■ 时间分辨率不一致及重采样

屏蔽

均等、插值生成固定时间间隔的数据

## ■ 地理坐标缺失及地理编码



屏蔽

# 1.4 | 运行监测时空信息—挑战及预处理



## (1) 文本相似性算法

- 最长公共子序列

(Longest Common Subsequence, LCS)

重建LCS    source: A B C B D A B  
              target: B D C A B A

target		B	D	C	A	B	A
source		0	0	0	0	0	0
A	0	0	0	0	1	1	1
B	0	1	1	1	1	2	2
C	0	1	1	2	2	2	2
B	0	1	1	2	2	3	3
D	0	1	2	2	2	3	3
A	0	1	2	2	3	3	4
B	0	1	2	2	3	4	4

最长公共子序列 LCS :

B != A 且 绿色 ≠ 黄色  
→ 删去绿色行

每次均从当前表格的右下角开始，  
比较该位置对应的行列元素是否相等

- 编辑距离

(Levenshtein Distance , LD)

		a	b	o	a	r	d
a							
b							
r							
o							
a							
d							

## (2) 拾取坐标系统

<http://api.map.baidu.com/lbsapi/getpoint/index.html>



## (3) 正/逆地理编码 API

<http://lbsyun.baidu.com/index.php?title=webapi/guide/webservice-geocoding>  
<https://lbs.amap.com/api/webservice/guide/api/georegeo/>

```
14→  
14→  
01→  
01→  
01→  
01→  
24→  
01→  
16→  
14→  
01→  
01→  
18→  
14→  
01→  
01→  
01→  
01→  
16→
```

屏蔽

```
0.633951225CRLF  
3951225CRLF  
→120.750170507CRLF  
213→120.779977471CRLF  
0.774277999CRLF  
→120.769882993CRLF  
233CRLF  
.6348972092→120.781662747CRLF  
→120.633145208CRLF  
0.633951225CRLF  
1340325→120.757052855CRLF  
0.769140993CRLF  
822→120.633951225CRLF  
7292897CRLF  
1518745CRLF  
726→120.754411212CRLF  
→120.777599825CRLF  
世纪2,3幢之间)→31.6695959428→120.781627408  
→120.756299805CRLF  
7071865CRLF
```

# 02

---

## 决策树与森林算法



High-performance Spatial Computational Intelligence Lab @ CUG

# 2.1 | 决策树



## ■ 什么是决策树？

根据天气状态决定是否打球

同学A：天气如何？

同学B：晴天

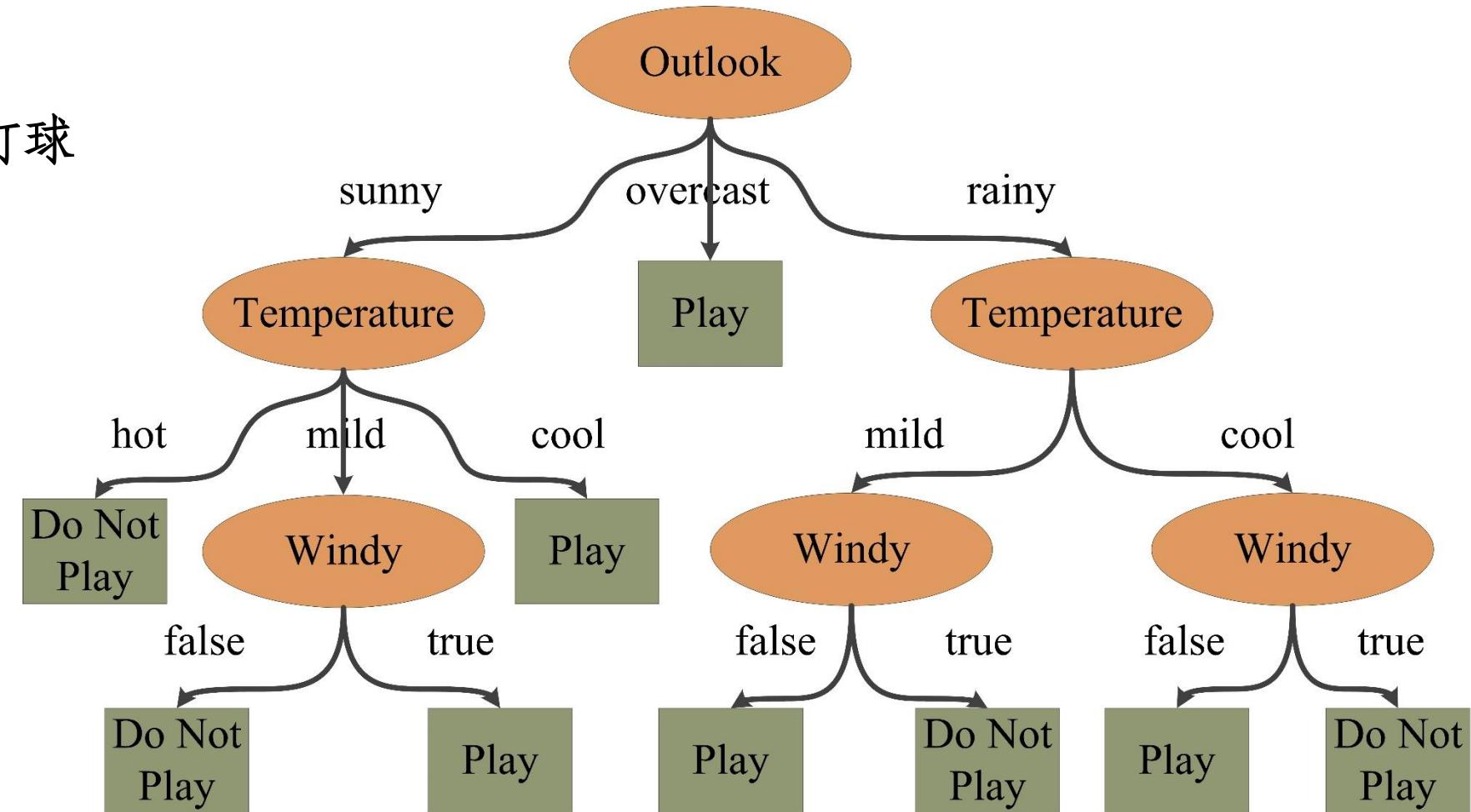
同学A：外面温度怎样？

同学B：一般吧

同学A：有风不？

同学B：有点风

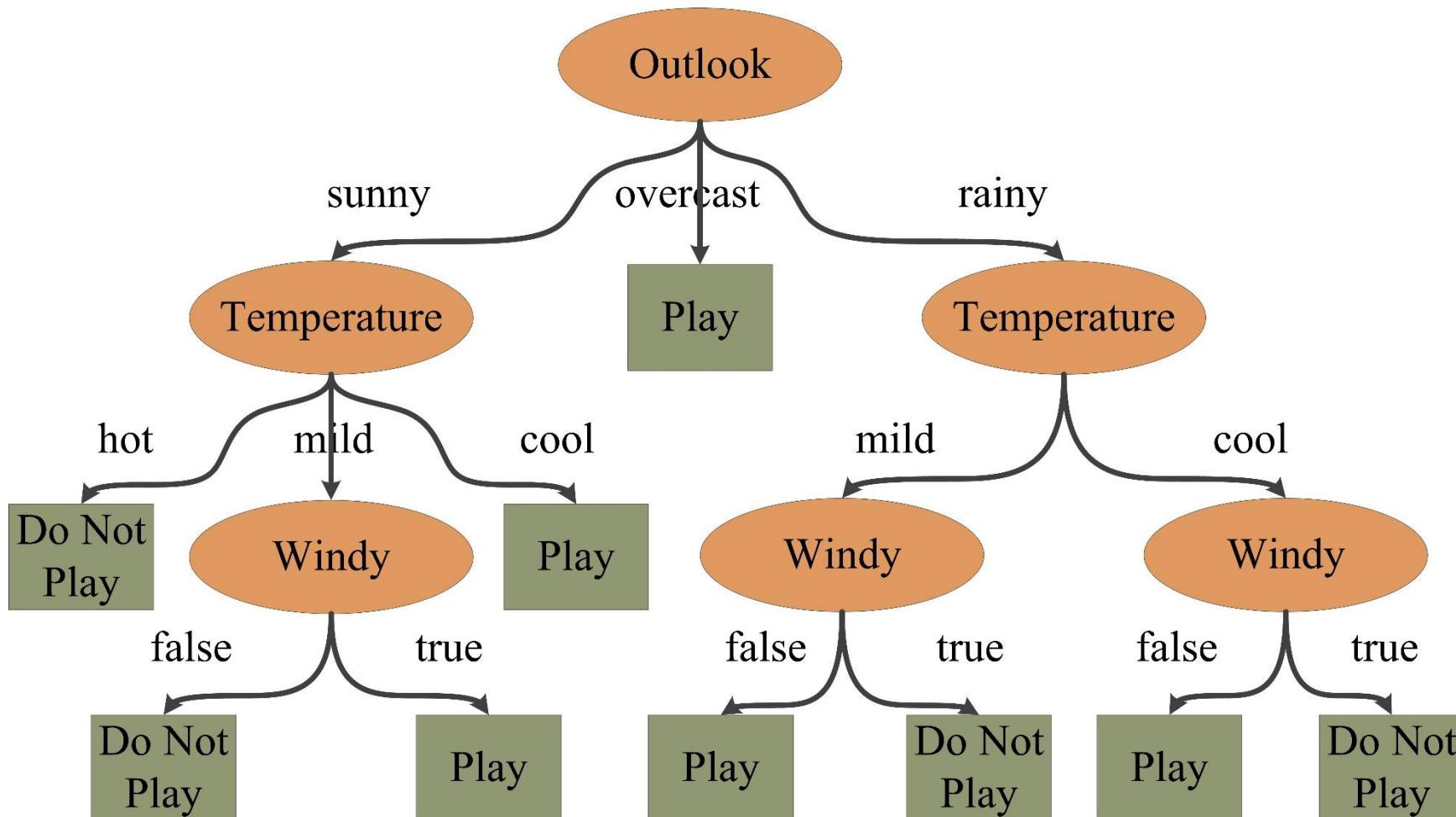
同学A：走，打球去！



# 2.1 | 决策树



## ■ 什么是决策树？



决策树：

基于**树结构**来进行决策。

决策树包含：

- **一个根结点**：对应属性测试。
- **若干个内部结点**：对应属性测试。
- **若干个叶结点**：对应决策结果。

## ■ 如何构造决策树？

如何选择根结点呢？

选择完根结点，又如何选择内部结点呢？

Outlook	Temperature	Humidity	Windy	Play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	23 no

## (1) 信息增益

信息熵 (**Information Entropy**) 是度量样本集合D纯度最常用的一种指标。

$$\text{Ent}(D) = - \sum_{k=1}^K p_k \log_2 p_k$$

$p_k$ 是D中第k类样本所占比例。Entropy越小，纯度越高。

信息增益 (**Information Gain**) 是衡量用属性a对样本集合D进行划分所得的“纯度提升”。可理解为原来信息需求与新信息需求之间的差。减号之前是未采用属性a划分的纯度，减号后是采用属性a划分后的纯度。

$$\text{Gain}(D, a) = \text{Ent}(D) - \sum_{v=1}^V \frac{|D^v|}{|D|} \text{Ent}(D^v)$$

Gain越大，纯度提升越大。选择属性  $a_* = \underset{a \in A}{\operatorname{argmax}} \text{Gain}(D, a)$

著名的典型的**ID3决策树**学习就是以信息增益为准则来选择划分属性。

构造树的基本想法就是随着树深度的增加，节点的熵迅速地降低。熵降低的速度越快，则更有可能得到一颗高度最矮的决策树。

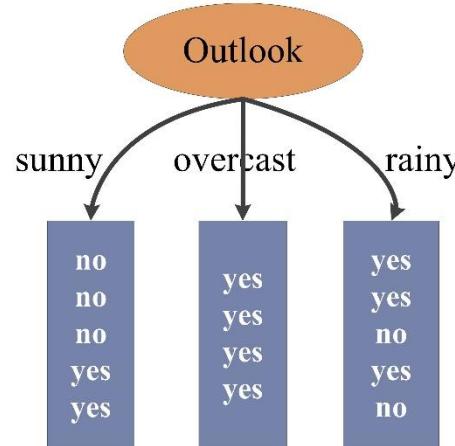
在没有给定任何天气信息时，根据历史数据，得某天打球概率是 $9/14$ ，不打球概率为 $5/14$ 。  
此时熵为：

$$-\frac{9}{14} \log_2 \frac{9}{14} - \frac{5}{14} \log_2 \frac{5}{14} = 0.940$$

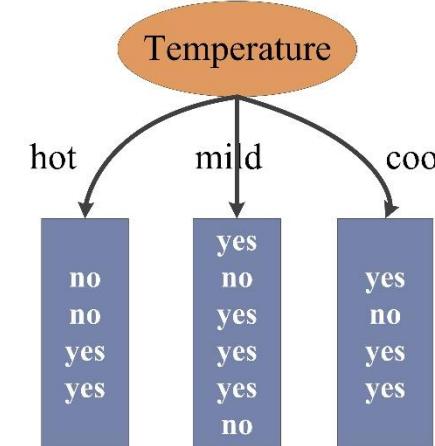
# 2.1 | 决策树



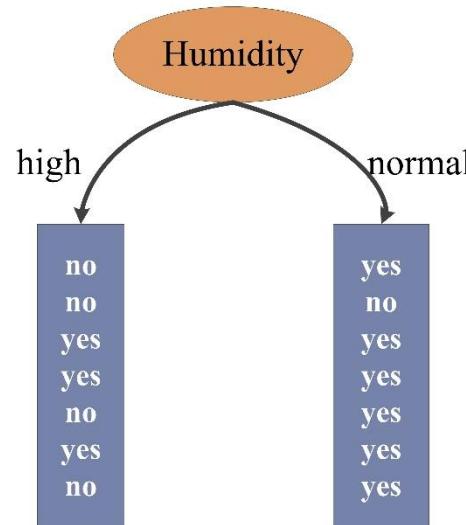
基于天气的划分



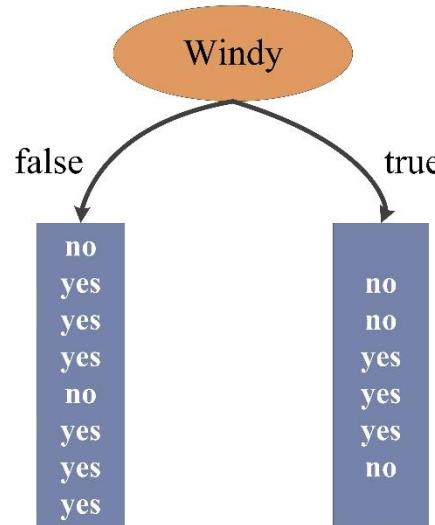
基于温度的划分



基于湿度的划分



基于有风的划分



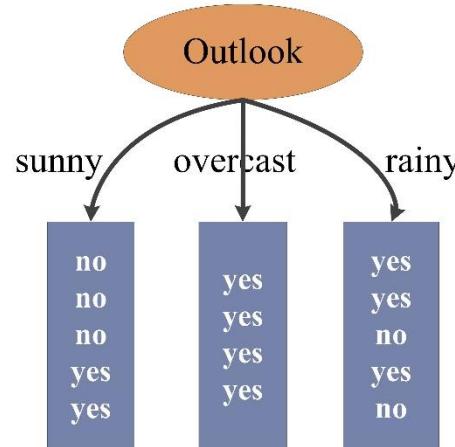
决定哪个属性作树的根节点呢？

以Outlook为例：

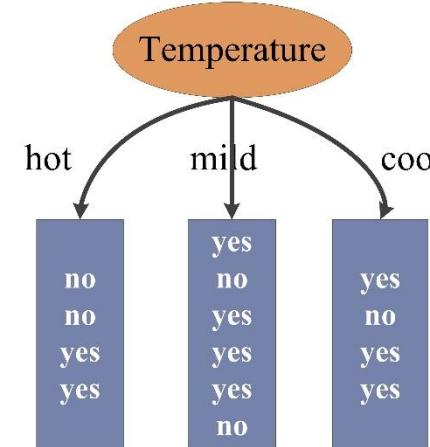
- Outlook = sunny时，打球概率为 $2/5$ ，不打球概率为 $3/5$ 。那么信息熵 $\text{entropy} = 0.971$
- Outlook = overcast时，熵 $\text{entropy} = 0$
- Outlook = rainy时，熵 $\text{entropy} = 0.971$

# 2.1 | 决策树

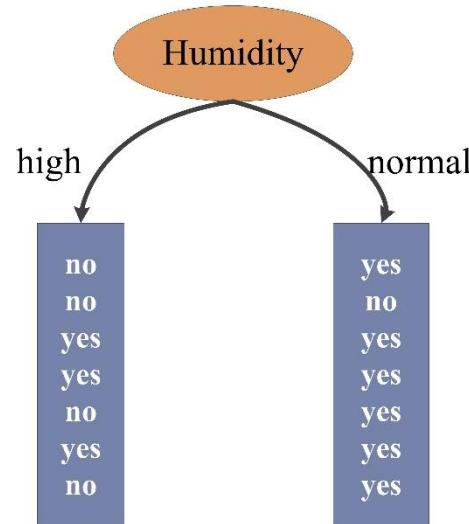
基于天气的划分



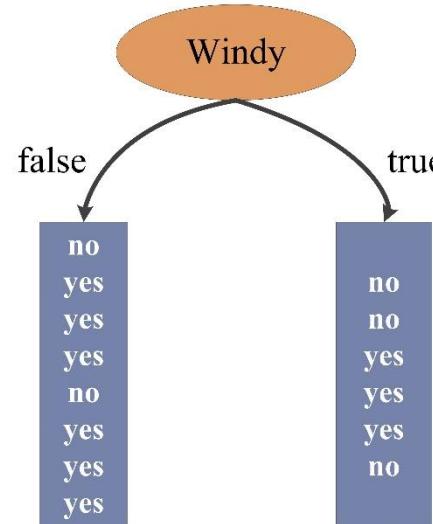
基于温度的划分



基于湿度的划分



基于有风的划分



决定哪个属性作树的根节点呢？

根据历史统计数据，Outlook为sunny、overcast、rainy的概率分别为 $5/14$ 、 $4/14$ 、 $5/14$ ，信息熵：

$$\frac{5}{14} \times 0.971 + \frac{4}{14} \times 0 + \frac{5}{14} \times 0.971 = 0.693$$

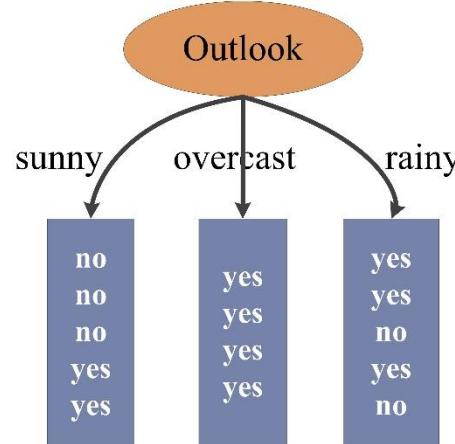
系统熵从 $0.940$ 下降为 $0.693$ ，信息增溢

$$\text{Gain(Outlook)} = 0.940 - 0.693 = 0.247$$

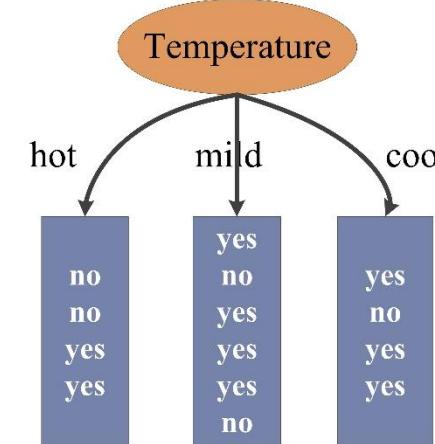
# 2.1 | 决策树



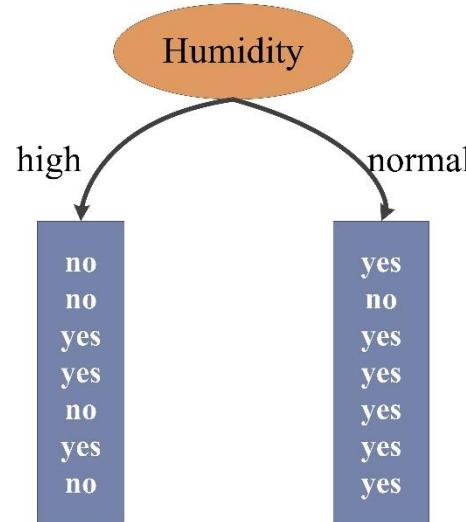
基于天气的划分



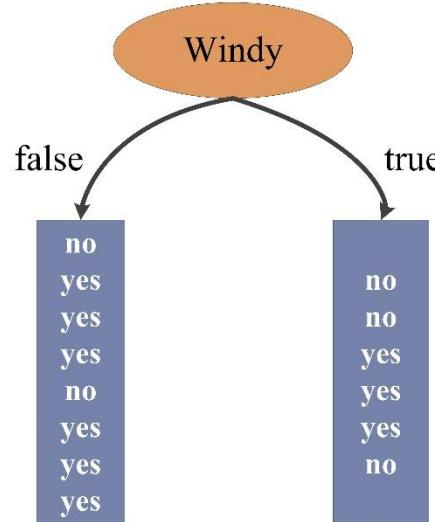
基于温度的划分



基于湿度的划分



基于有风的划分



决定哪个属性作树的根节点呢？

同样可得：

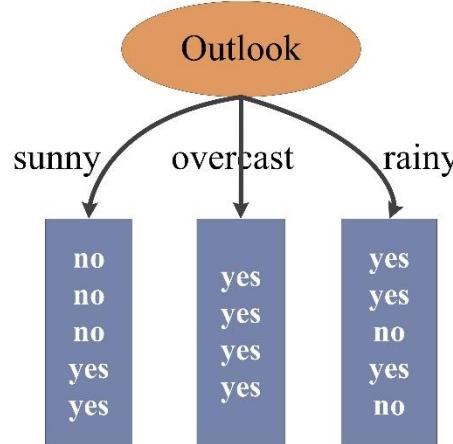
- 信息增溢  $\text{Gain}(\text{Temperature}) = 0.029$
- 信息增溢  $\text{Gain}(\text{Humidity}) = 0.152$
- 信息增溢  $\text{Gain}(\text{Windy}) = 0.048$

Gain(Outlook)最大，所以根节点选择Outlook。

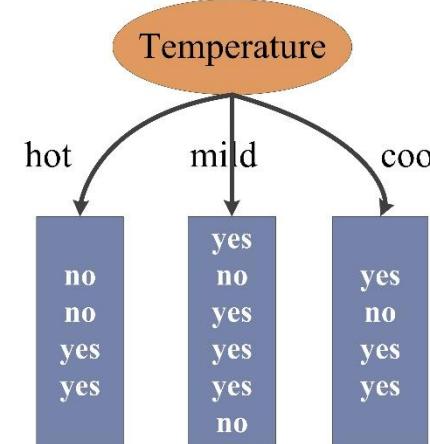
# 2.1 | 决策树



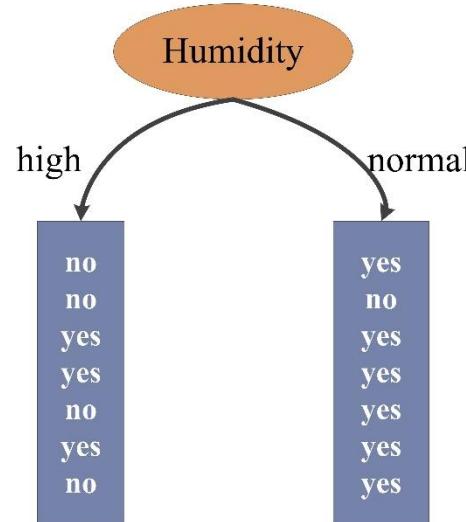
基于天气的划分



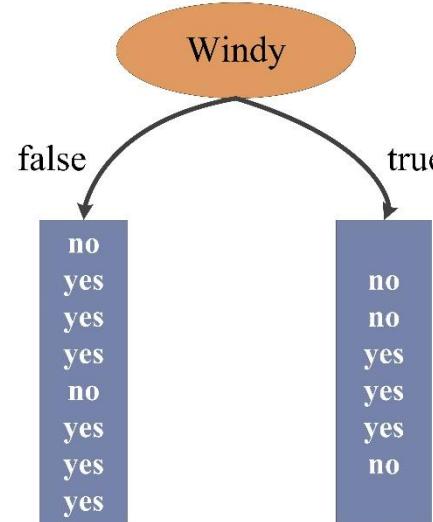
基于温度的划分



基于湿度的划分



基于有风的划分



决定哪个属性作树的内部节点呢？

依此类推，构造决策树。

当系统的信息熵将为0时，则没有必要继续构造决策树了，此时叶子节点都是纯的，这是理想情况。

最坏的情况，决策树的高度为属性个数，叶子节点不纯（意味着我们要以一定的概率作出决策）。

(2) 增益率

ID	Outlook	Temperature	Humidity	Windy	Play
1	sunny	hot	high	false	no
2	sunny	hot	high	true	no
3	overcast	hot	high	false	yes
4	rainy	mild	high	false	yes
5	rainy	cool	normal	false	yes
6	rainy	cool	normal	true	no
7	overcast	cool	normal	true	yes
8	sunny	mild	high	false	no
9	sunny	cool	normal	false	yes
10	rainy	mild	normal	false	yes
11	sunny	mild	normal	true	yes
12	overcast	mild	high	true	yes
13	overcast	hot	normal	false	yes
14	rainy	mild	high	true	30 no

## (2) 增益率

$$\text{Gain\_ratio}(D, a) = \frac{\text{Gain}(D, a)}{\text{IV}(a)}$$

$$\text{IV}(a) = - \sum_{v=1}^V \frac{|D^v|}{|D|} \log_2 \frac{|D^v|}{|D|}$$

属性a的可能取值数目越多，则 $\text{IV}(a)$ 的值通常越大，则增益率越小。可见增益率对于可取值数目较少的属性有所偏好。

著名/典型的**C4.5决策树**学习就是以增益率为准则来选择划分属性。不过并非直接选择增益率最大的候选划分属性，而是使用了一个启发式：先从候选划分属性中找到信息增益高于平均水平的属性，再从中选择增益率最高的划分属性。

## (3) Gini指数

以基尼系数替代熵，最小化不纯度，而不是最大化信息增益。

**CART决策树**使用基尼指数（Gini Index）来衡量由样本集合D的纯度。

$$\text{Gini}(D) = - \sum_{k=1}^K \sum_{k' \neq k} p_k p_{k'} = 1 - \sum_{k=1}^K p_k^2$$

Gini(D)反映了从数据集D中随机抽取两个样本，其类别标记不一致的概率。Gini(D)越小，D纯度越高。

属性a的基尼指数：

$$\text{Gini\_index}(D, a) = \sum_{v=1}^V \frac{|D^v|}{|D|} \text{Gini}(D^v)$$

选择属性  $a_* = \operatorname{argmin}_{a \in A} \text{Gain\_index}(D, a)$ 。即选择使得划分后基尼指数最小的属性。

## ■ 剪枝处理

是决策树学习算法**对付“过拟合”的主要手段。**

在决策树学习中，为了尽可能正确分类样本训练样本，结点划分过程将不断重复，有时会造成决策树分支过多，此时就可能因训练样本学得不太好，以至于把训练集自身的一些特点或者噪声当作所有数据都具有的一般性质而导致过拟合。因此，需要**去掉一些分支来降低过拟合风险**。

# 2.1 | 决策树



## ■ 剪枝处理

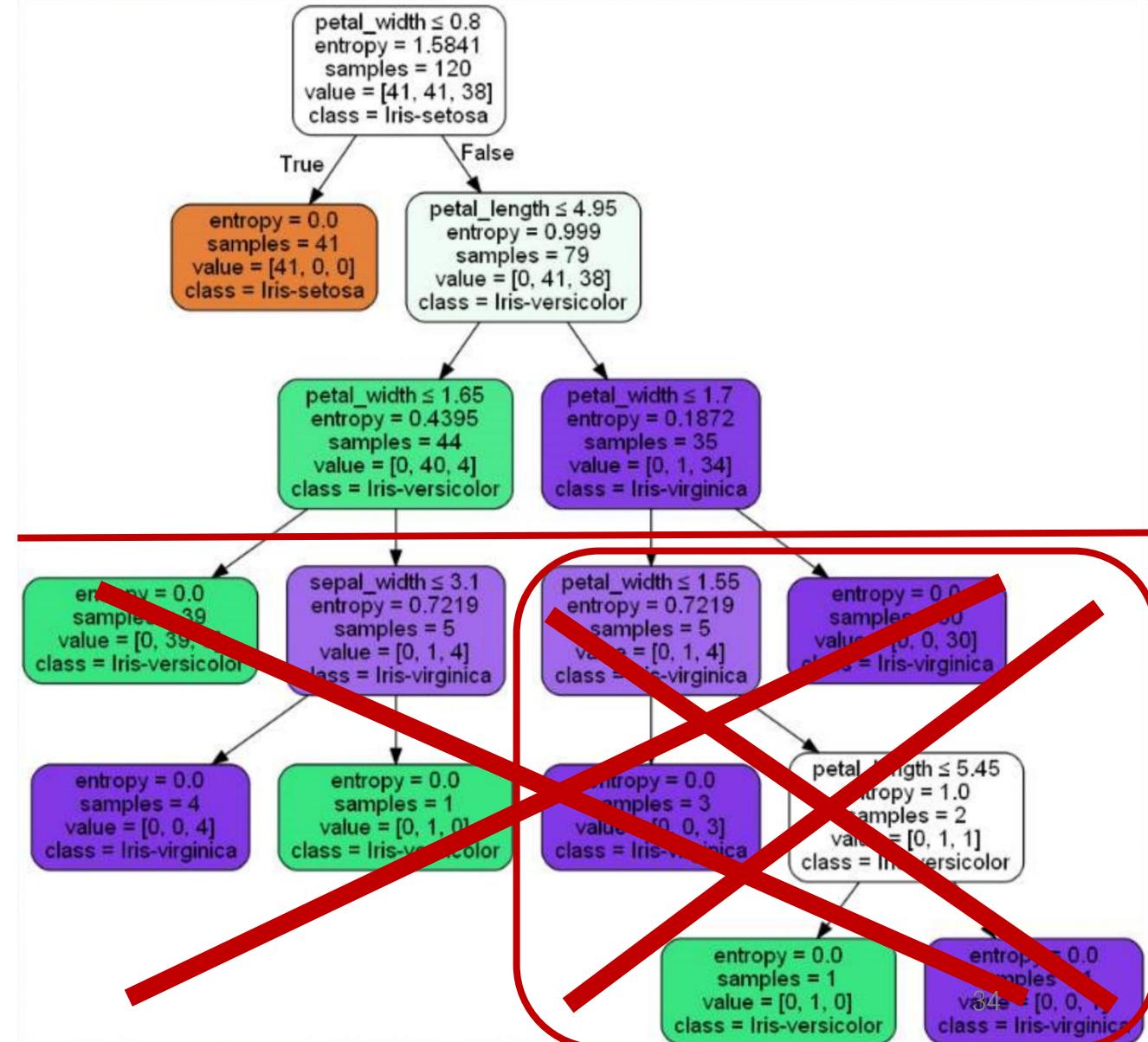
决策树剪枝策略：预剪枝、后剪枝。

预剪枝：

在构建决策树的过程时，提前停止。

比如设置：

- 决策树的深度
- 节点中最少样本个数



# 2.1 | 决策树

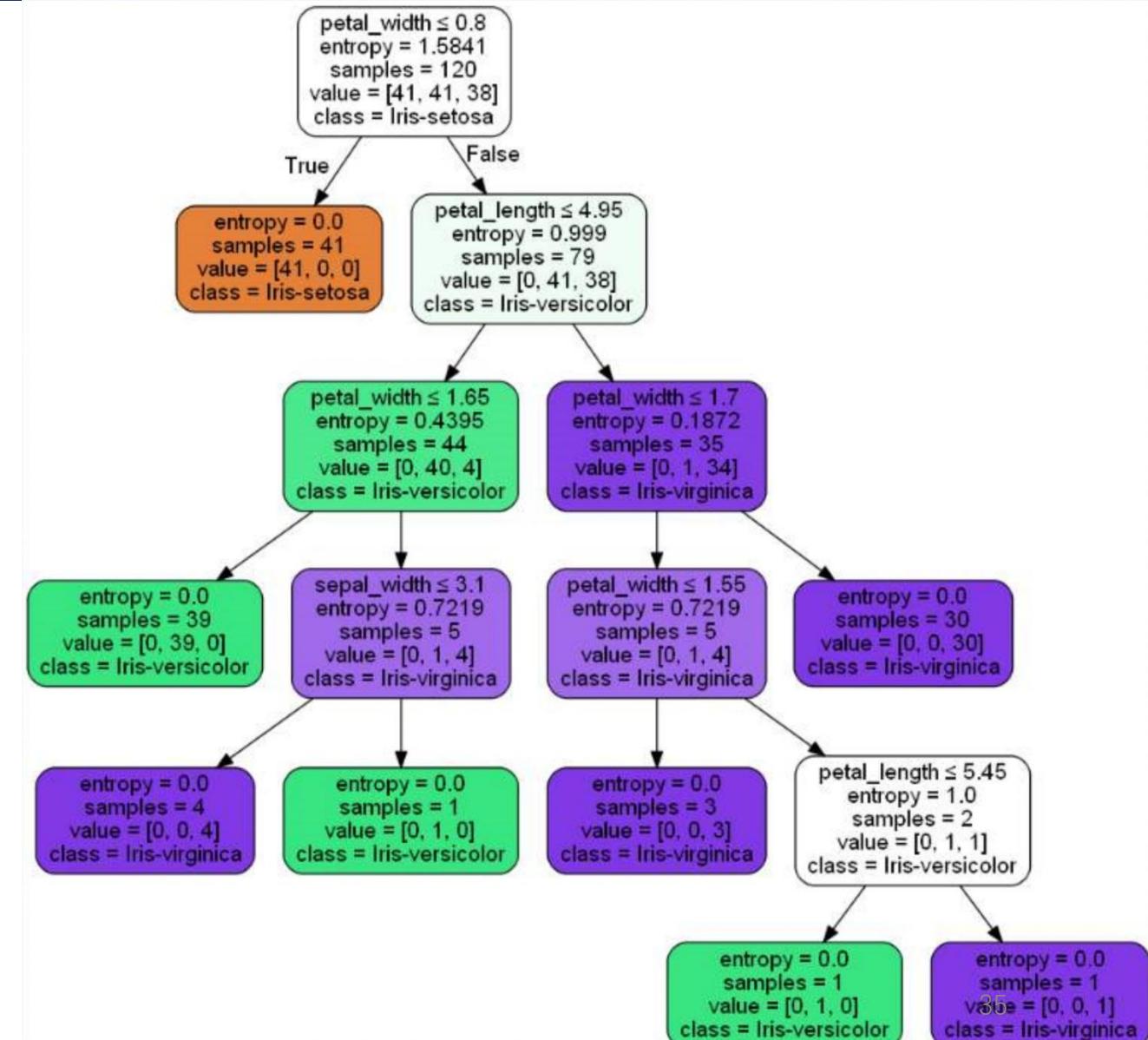


## ■ 剪枝处理

决策树剪枝策略：预剪枝、后剪枝。

后剪枝：决策树构建好后，然后才开始剪枝。

若将结点对应的子树替换为叶结点能带来决策树泛化能力的提升，则将该子树替换为叶结点。



## ■ 连续值与缺失值处理

连续值：

取值不再有限，不能直接根据连续属性的可取值对结点进行划分。

需连续属性离散化。

可用二分法处理。从哪里二分？

对于 $n$ 个数，有 $n-1$ 种分法。若采用信息增益，则可取使Gain最大化的划分点。

## ■ 连续值与缺失值处理

缺失值：

- (1) 如何在属性值缺失的情况下进行划分属性选择？

仅可以根据没有缺失值的样本子集  $\tilde{D}$  来判断属性  $a$  的优劣。

$$\text{Gain}(D, a) = \rho \times \text{Gain}(\tilde{D}, a)$$

其中  $\rho$  是无缺失值样本所占比例。

- (2) 给定划分属性，若样本在该属性上的值缺失，如何对样本进行划分？

取值可知，划分到对应的结点。

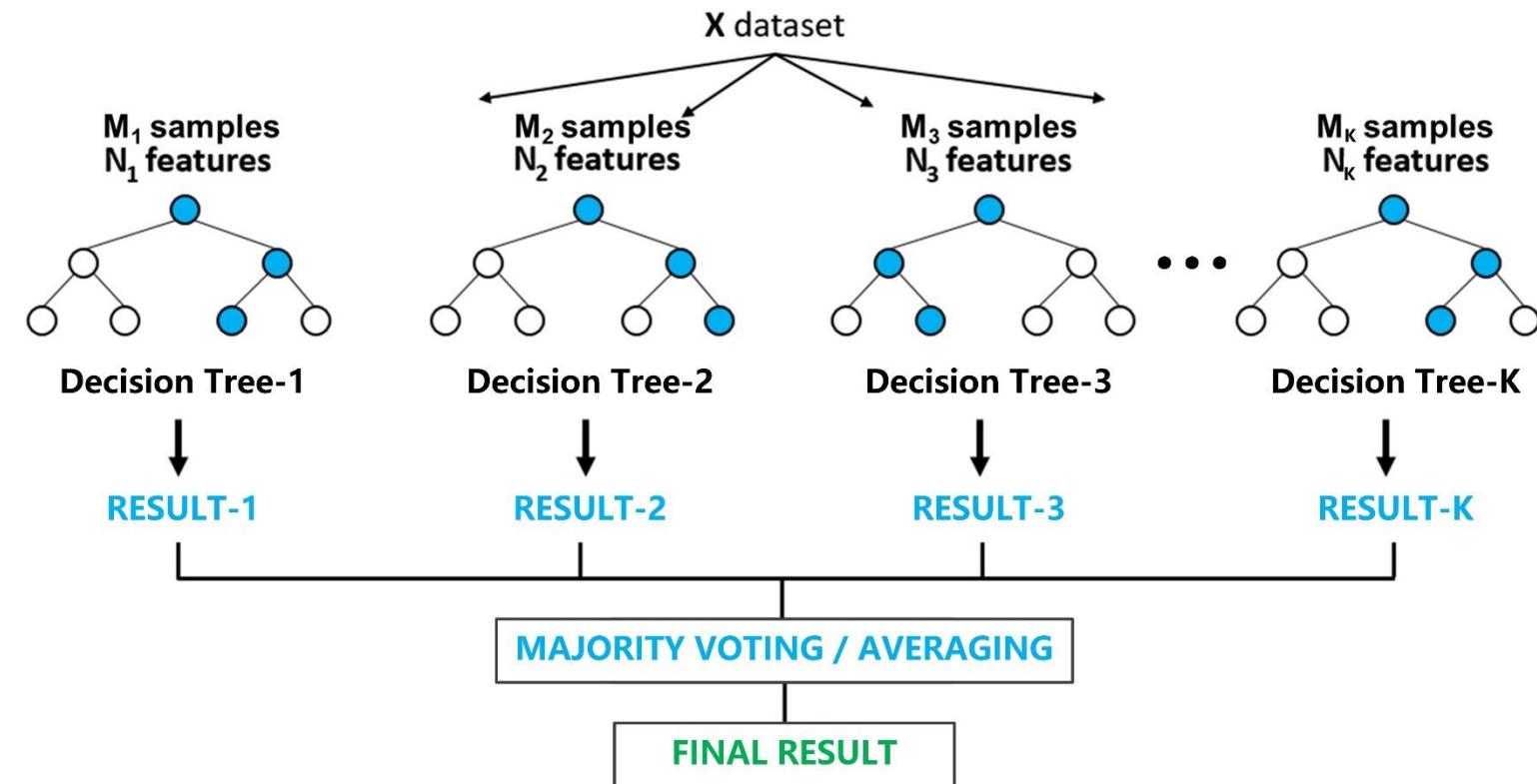
取值不可知，划入所有子结点。

## 2.2 | 随机森林



随机森林是一种“**集成**”算法，通过组合多个弱分类器，最终结果通过“**投票**”或“**取均值**”，使得整体模型的结果具有较高的精确度和泛化性能。

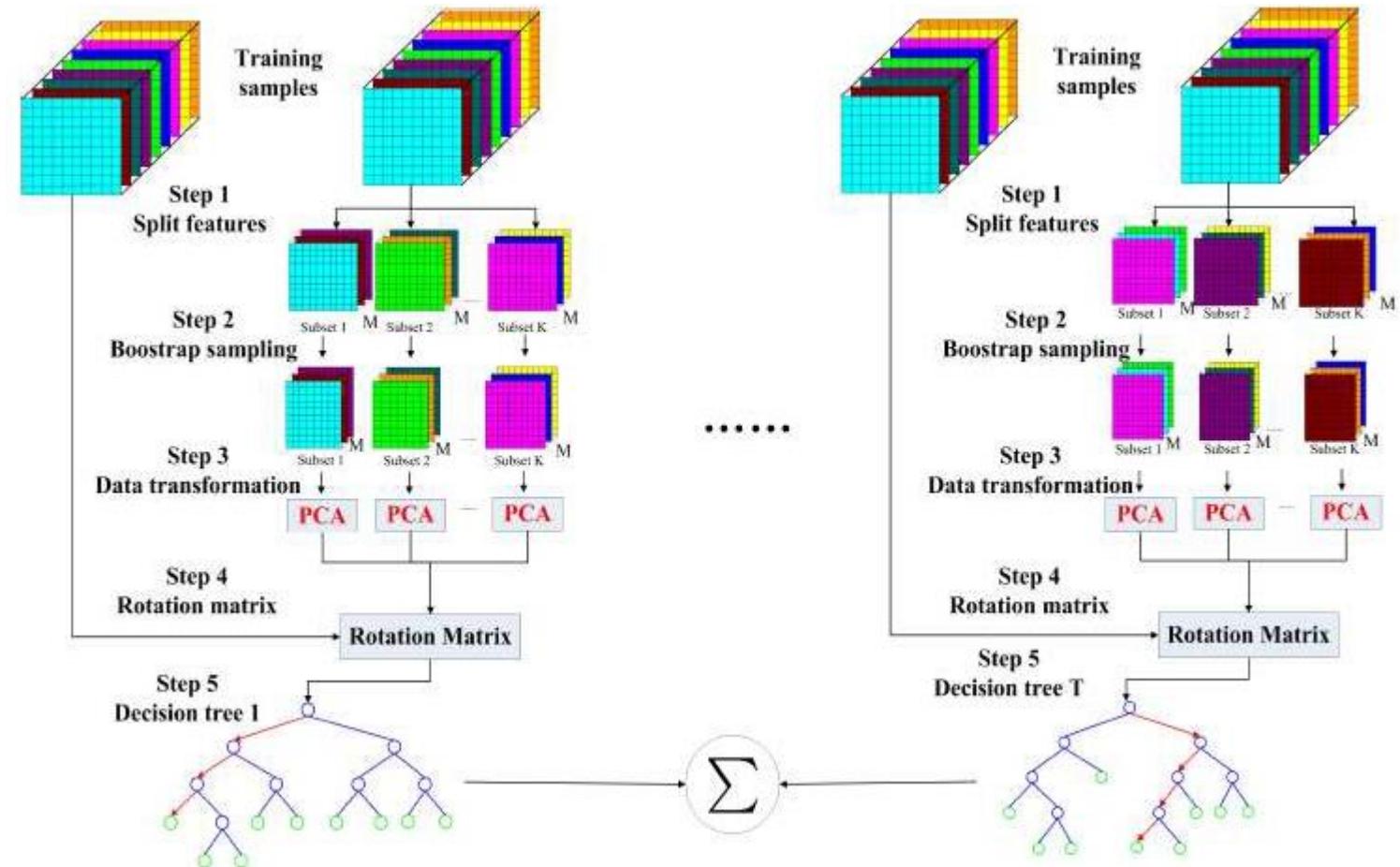
随机森林优异主要归功于“**随机**”和“**森林**”，一个使它具有抗过拟合能力，一个使它更加精准。



旋转森林是一种多分类器集成策略。

**“集成”**结合不同的学习模块，对分类或者预测结果进行**“投票”**，加强模型的稳定性和预测能力。

旋转森林利用了随机森林的基础，实现了改进。在每次抽取子样本前，对样本属性集进行随机分组，采用PCA对各组子属性集之间的数据进行特征变换，使得各子样本有所区别，进而提高各基分类器的准确性和差异性。



03

---

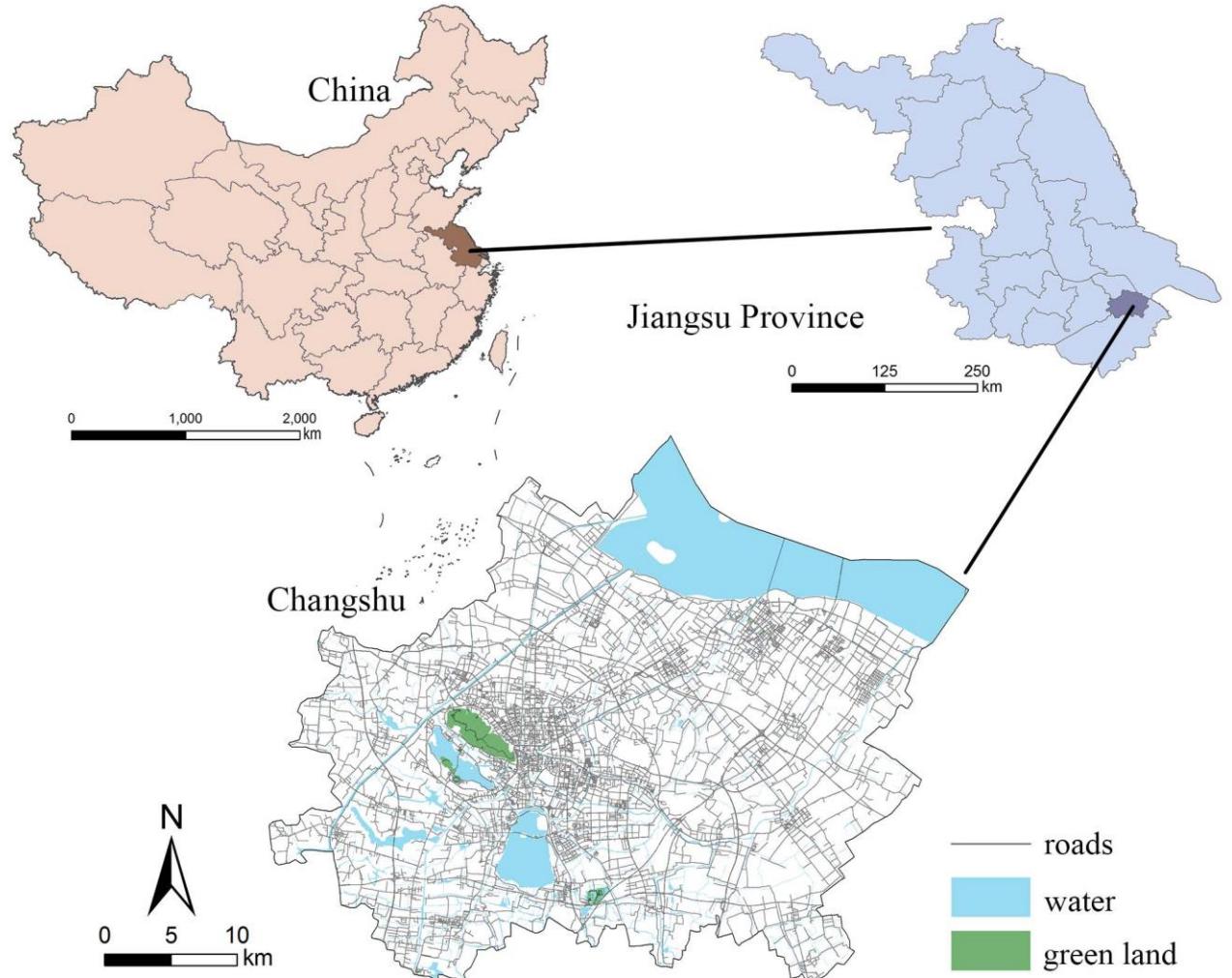
## 基于市政水耗剖析城市空间结构



High-performance Spatial Computational Intelligence Lab @ CUG

## 研究区：江苏省常熟市

- 面积 $1264 \text{ km}^2$
- $31^\circ 31' \sim 31^\circ 50' \text{N}, 120^\circ 33' \sim 121^\circ 03' \text{E}$
- 人口1,068,700
- 中国经济发达县之一，纺织业发达

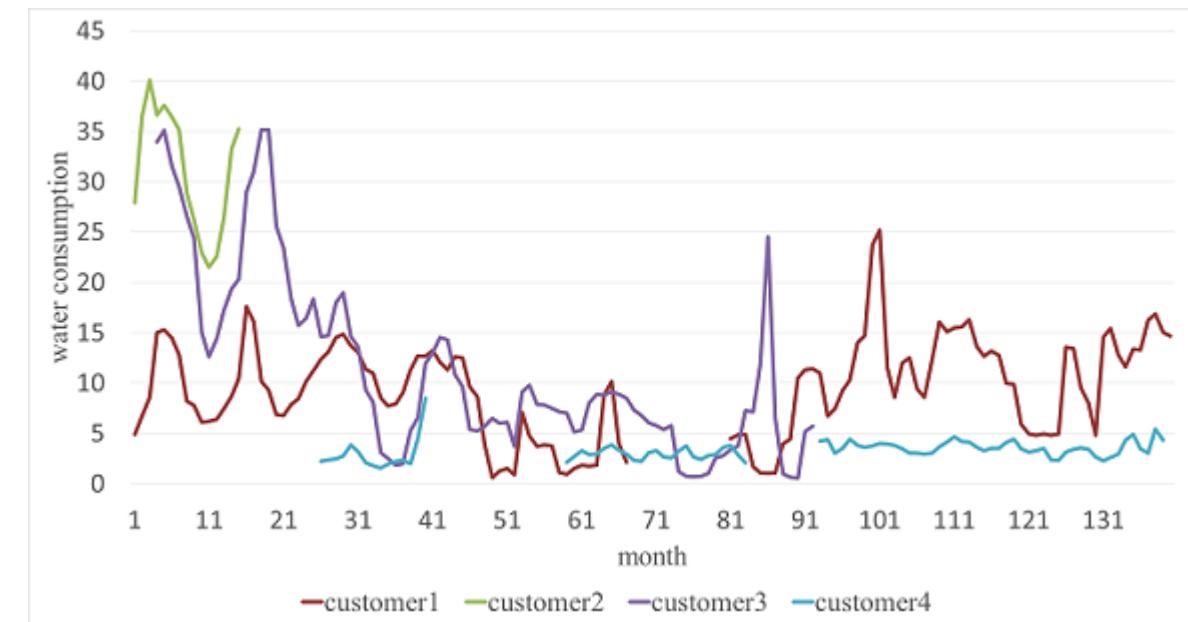


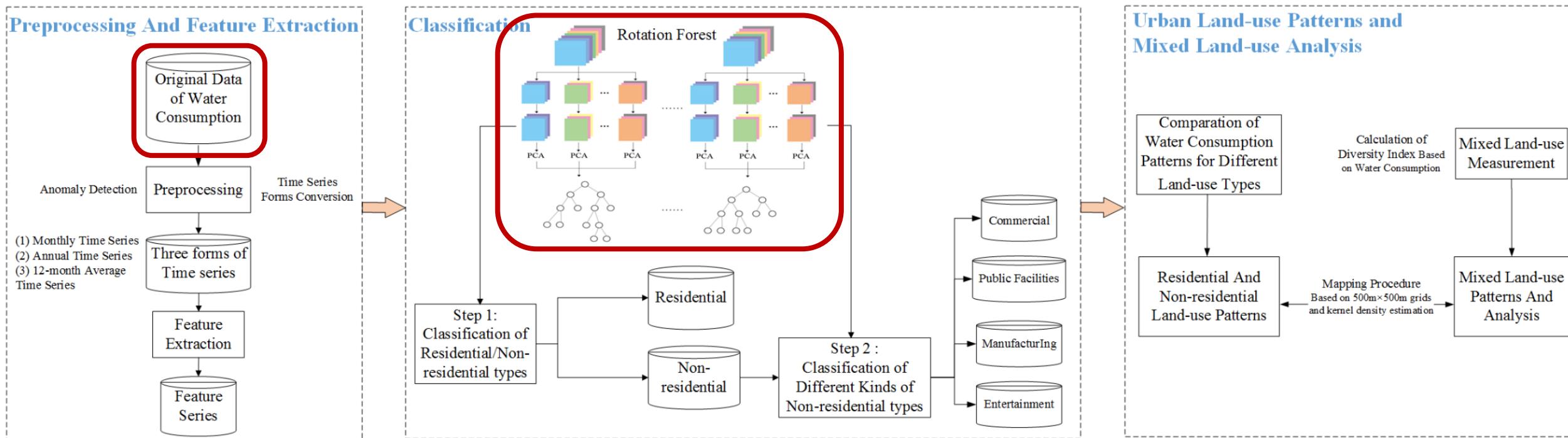
研究区：中国江苏省常熟市， $1,264 \text{ km}^2$ ，包括 10 区

## 供水运营数据

- 包含水表ID，用户地址，水表读表记录，读表时间，收费等
- 时间跨度为1997年1月起至2013年10月
- 超过 400,000个 水表(户)

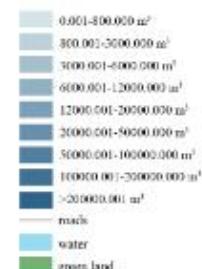
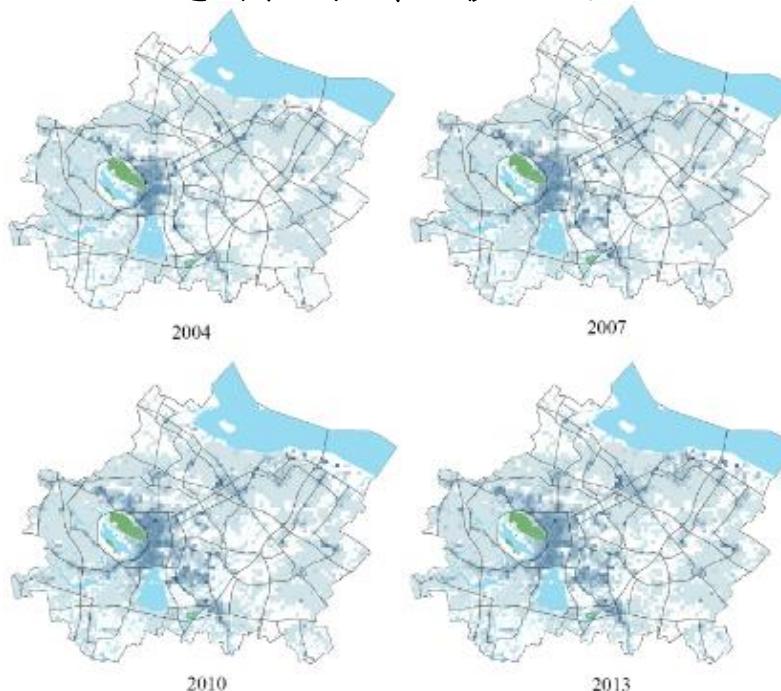
不同用户的时序示例：





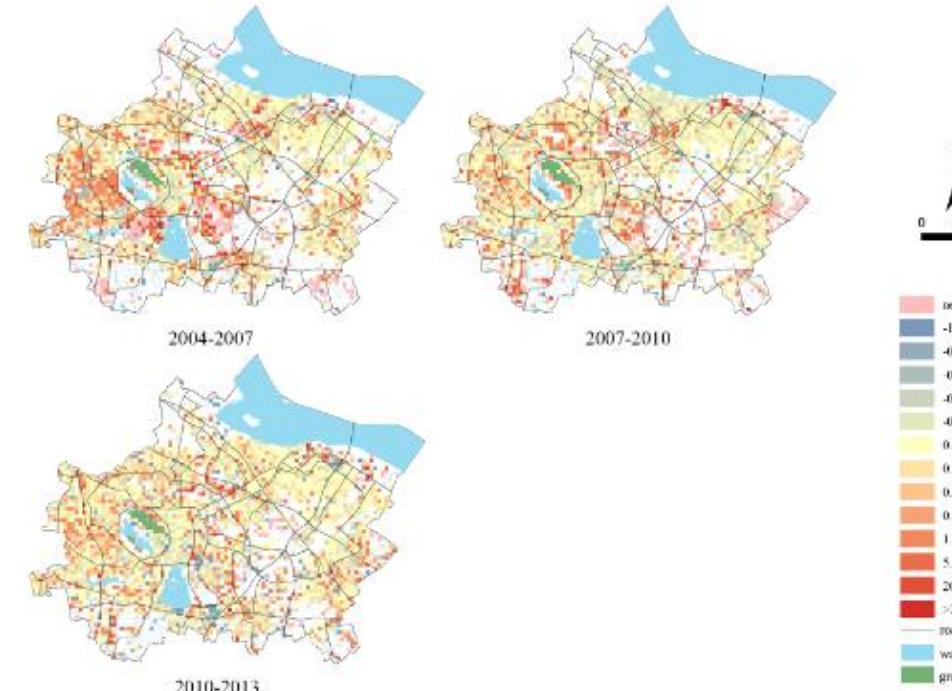
Guan, Q.\*, Cheng, S.\*, Pan, Y.\*, Yao, Y.\* & Zeng, W.\* (2020). Sensing mixed urban land-use patterns using municipal water consumption time series. *Annals of the American Association of Geographers*. DOI:10.1080/24694452.2020.1769463

## ■ 总体水耗模式



(A) Spatial distribution of water consumption.

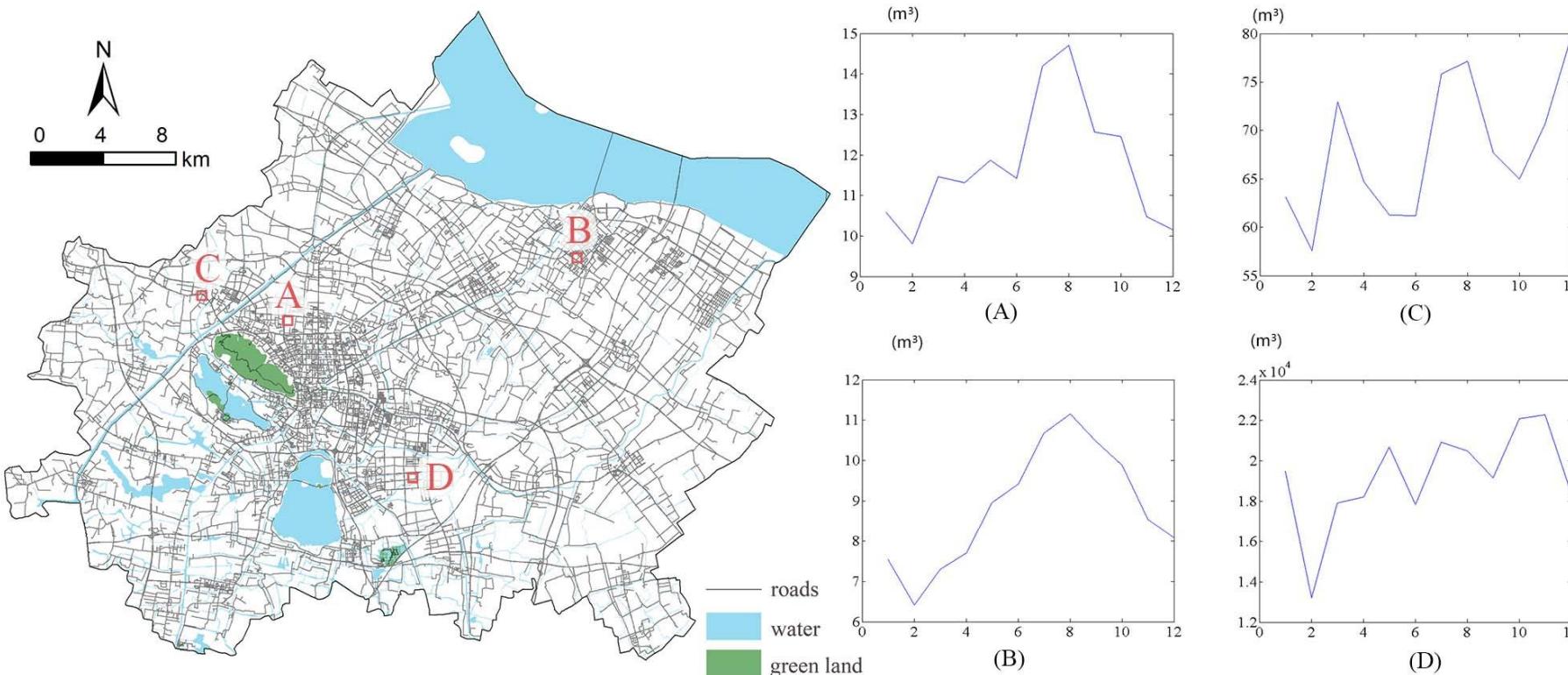
2004年至2013年水耗的空间分布和增长率



(B) Growth rate of water consumption from 2004 to 2013.

- 非零水耗的格网从46.73%增加到55.31%，平均水耗从1,897.36m<sup>3</sup>增加到2,674.3m<sup>3</sup>
- 水耗较高的地区位于**市中心**
- 水耗快速增长的地区**集中在市中心周围，并逐年向外扩张**

## ■ 典型用地用水模式分析

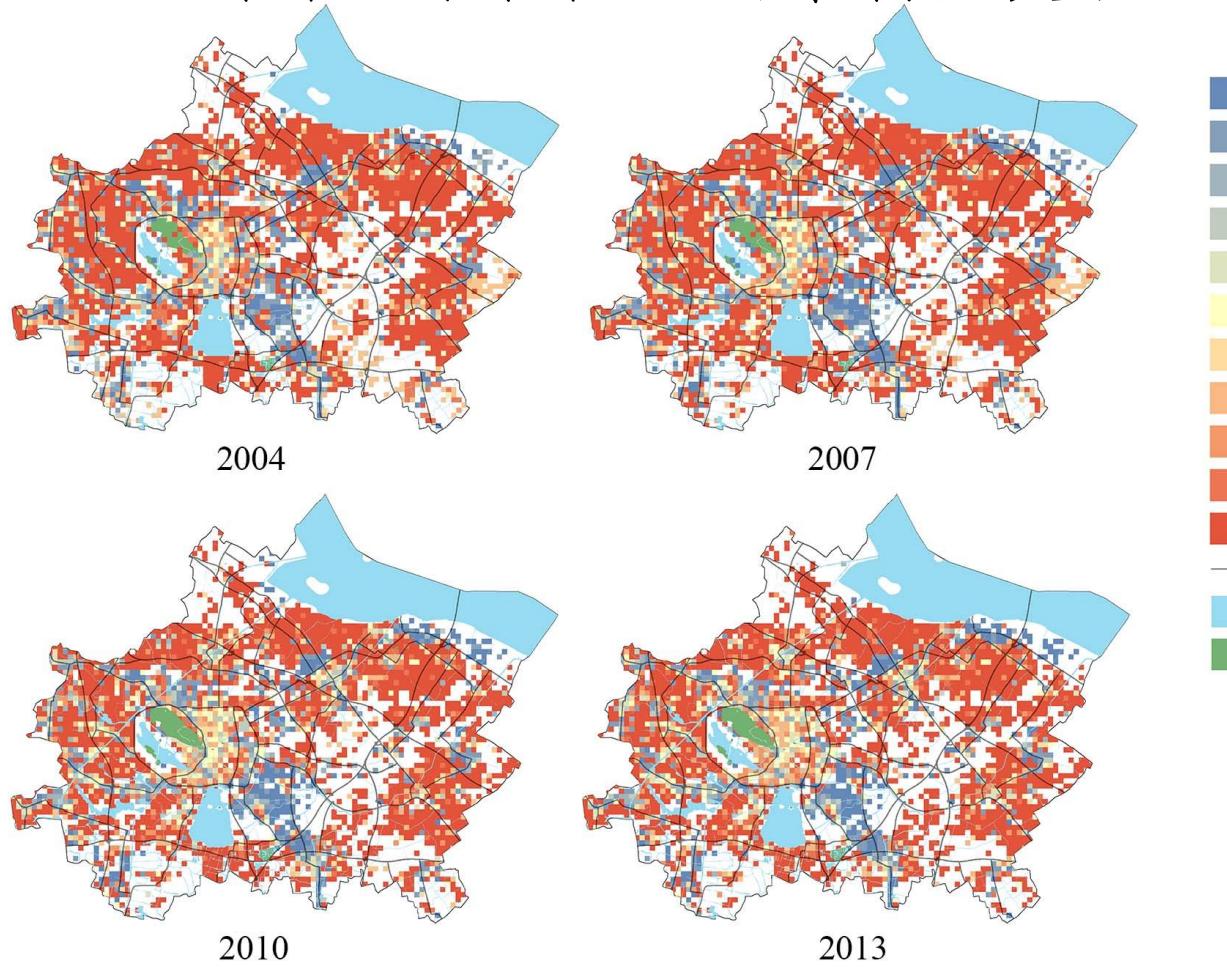


典型居民区、非居民区的用水时序曲线

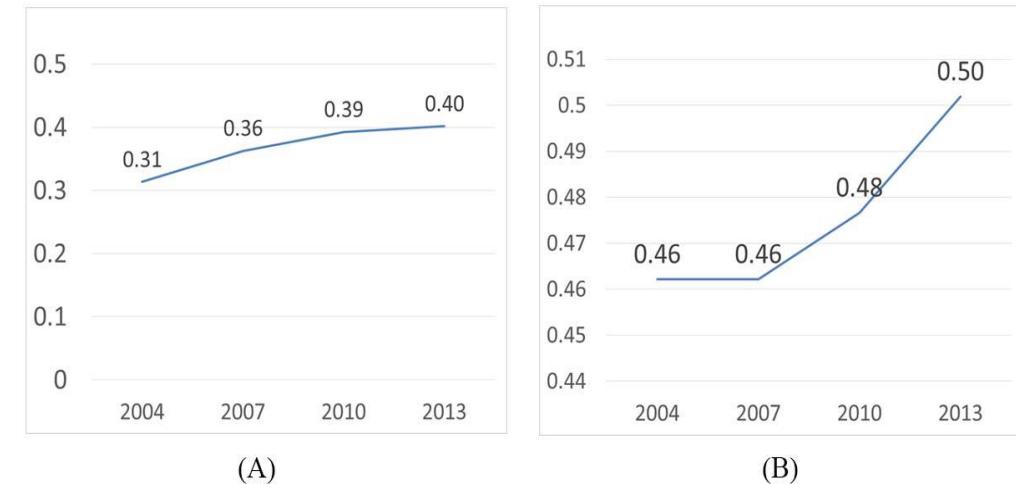
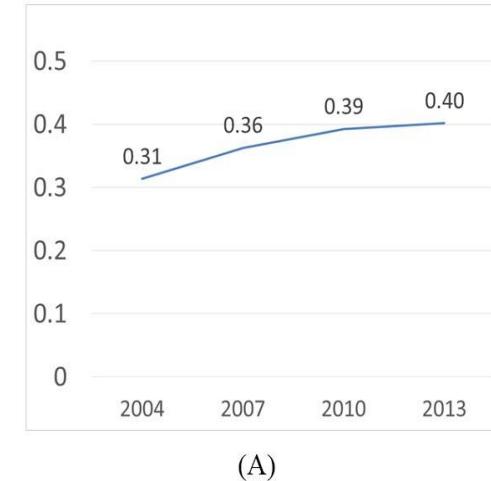
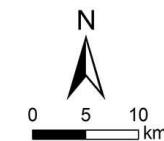
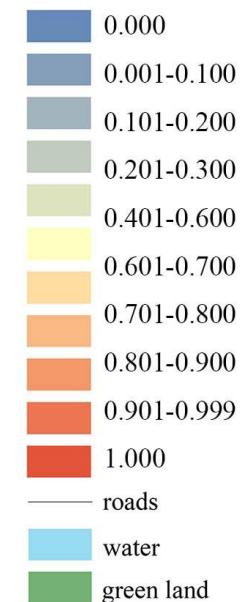
- (1) 居住用地曲线相似，夏季用水高
- (2) 居住和非居住用地曲线形态和数量相异。
- (3) 非居住用地用水量曲线具备较不规则，受功能影响。（C附近为材料工厂，D为电子公司）

用水量和城市土地利用类型具有一定的相关性。

## ■ 居住/非居住空间结构及变化

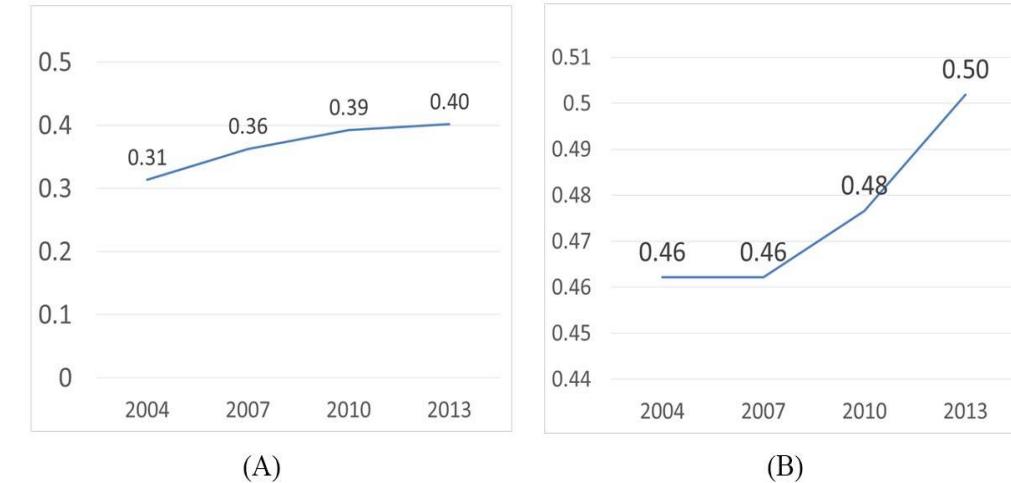
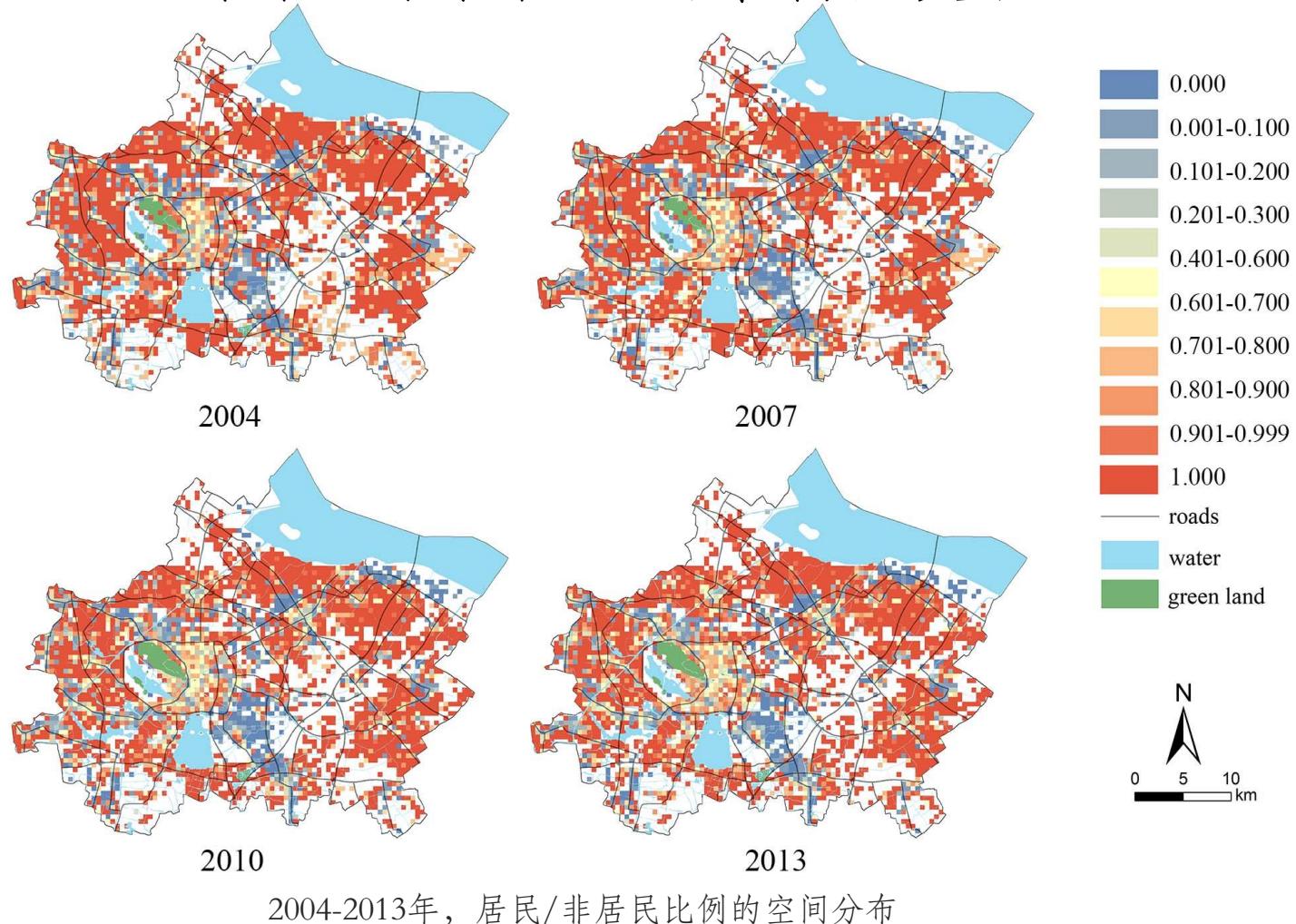


2004-2013年，居民/非居民比例的空间分布



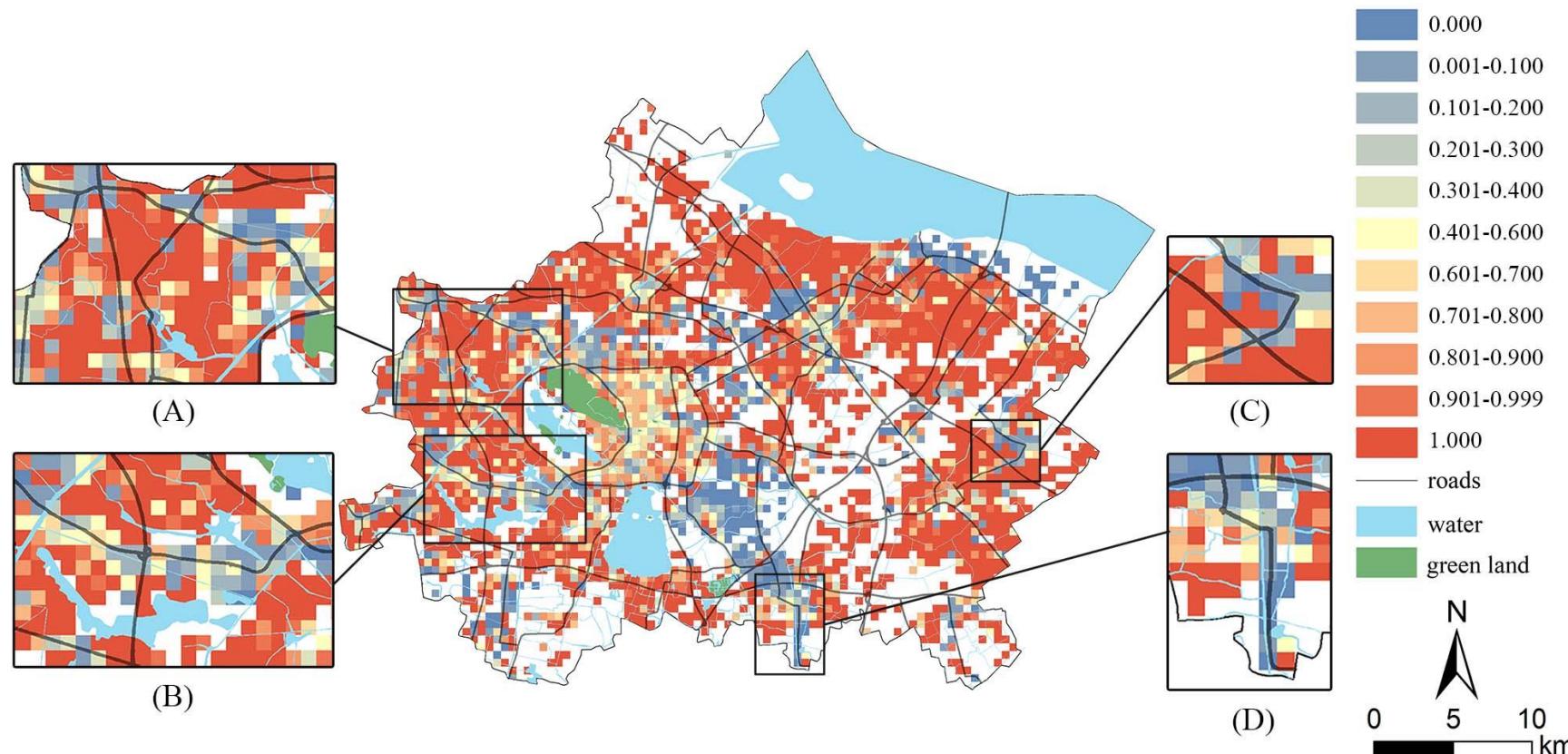
(A) 居民、非居民混合功能的网格占比。(B)所有的混合网格的比值的平均值。

## ■ 居住/非居住空间结构及变化



- (1) 空间分布方面：中心区域居民与非居民混合程度较高；四周区域有些以居民用水为主，有些混合度较高。
- (2) 非居民主导区域：如东北角、中心区东南部等。
- (3) 时间上，居民与非居民混合程度逐渐增大。

## ■居住/非居住空间结构及变化



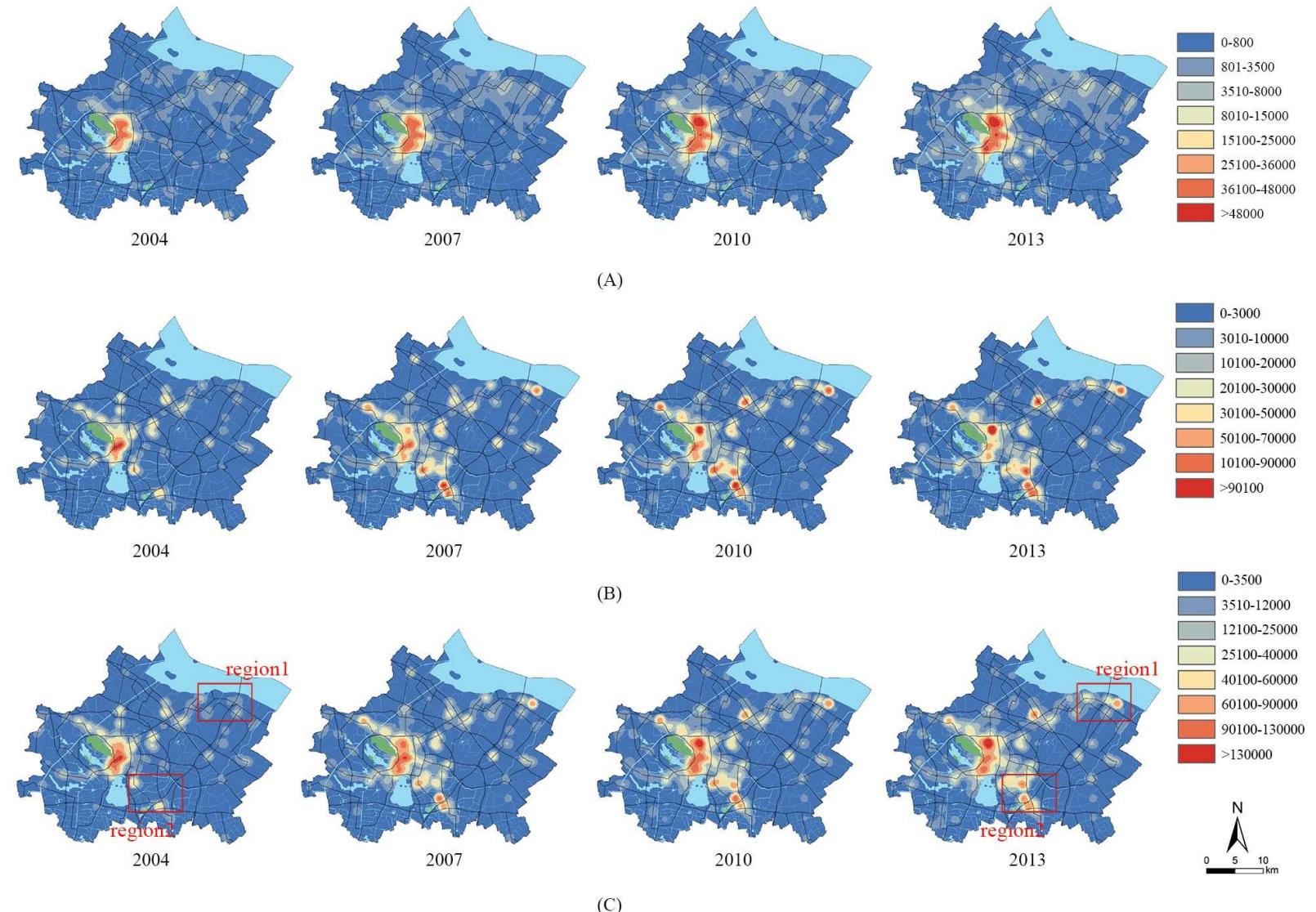
局部典型居住非居住用地分布特点

“居民沿河分布”、  
“非居民沿路分布”现象：  
重要道路附近混合度较高或  
非居民用水较高的现象。

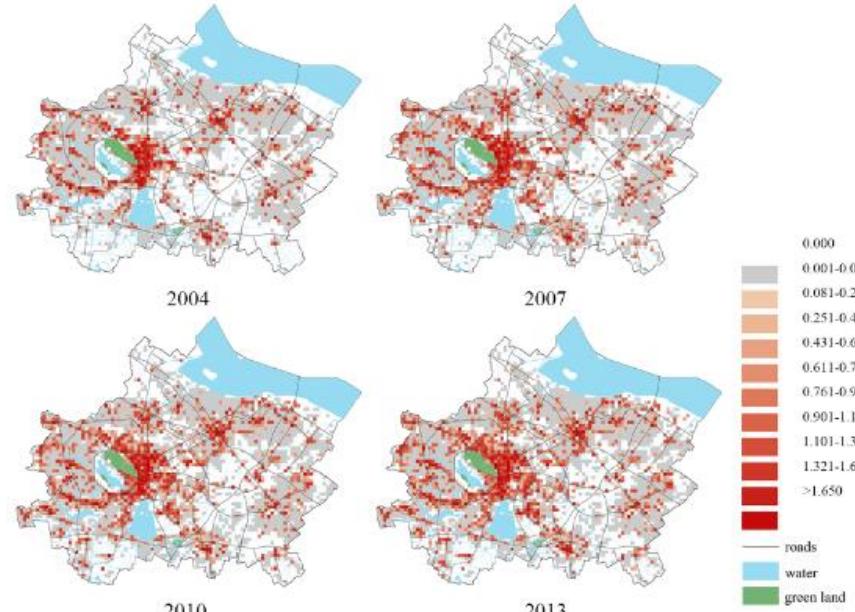
## ■ 城市中心区提取

- (1) 居民：中心区为一个大的主要中心和多个小的次级中心结构，主中心规模逐渐扩大；
- (2) 非居民：主中心聚集程度更大，东北部、中心区东南部的次级中心（方框区域）逐渐扩大规模；
- (3) 整体：演变与非居民相似，说明非居民用水量影响远大于居民；

原因：虞山镇经济发达；开发区的发展；拆迁等



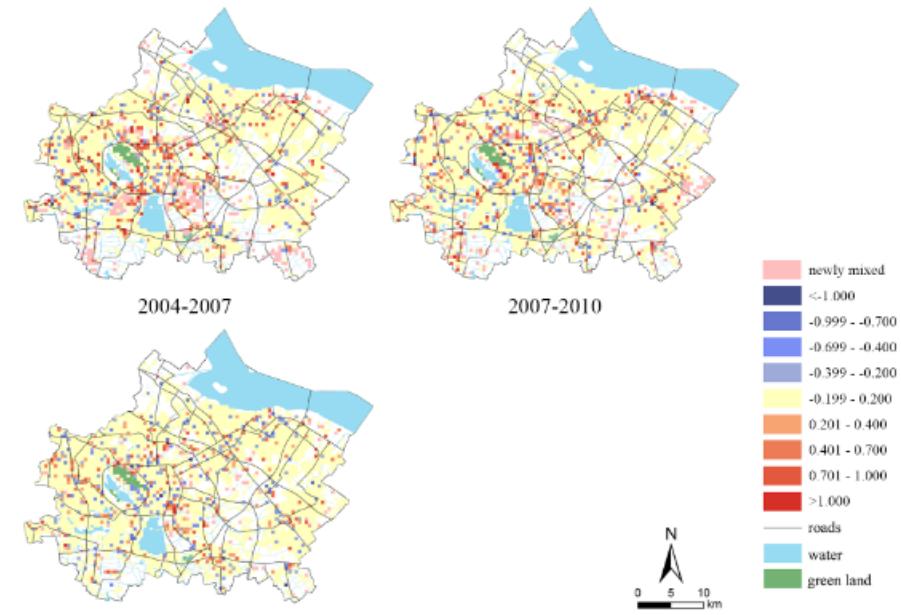
## ■ 土地利用混合度时空分布



(A) Mixed land-use patterns from 2004 to 2013 delineated by the land-use diversity index.

2004-2013 混合用地模式

- 高度混合的网格集中在**城市和城镇的中心**
- 部分高度混合的网格沿**主要道路分布**

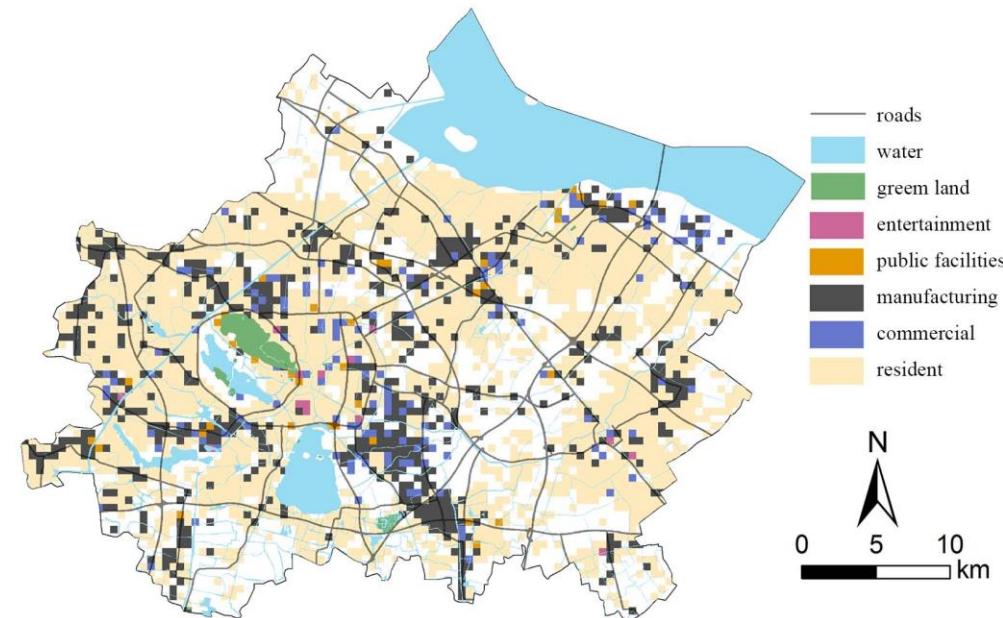


(B) Change of the land-use diversity index.  
Positive values indicate an increase, and a negative value indicates a decrease.

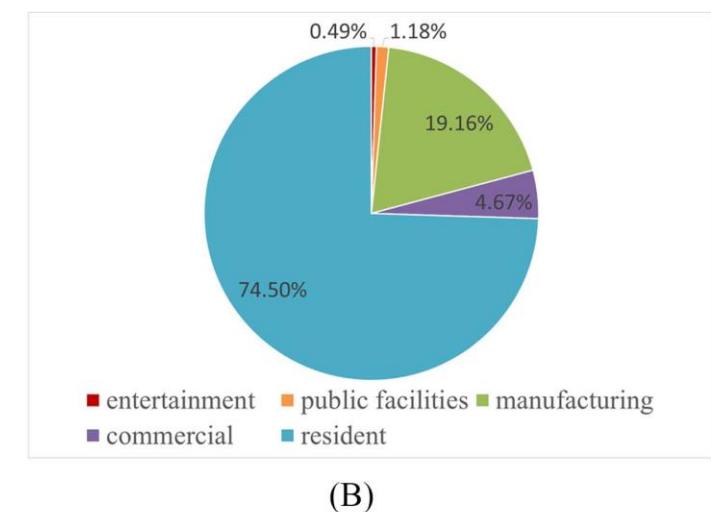
- 混合网格的比例从**33%**增加到**42%**，平均土地利用多样性指数从**0.26**增加到**0.34**
- 混合土地利用从**中心区域扩展到其周围**

## ■不同功能土地利用分布

- (1) 居住用地分布广泛。
- (2) 商业娱乐用地仅占城市空间很小一部分，主要分布在城市中心区域。
- (3) 公共基础设施用地，占比也相对较小。与娱乐用地较为相似，分布在城市中心地段，但整体上相对均匀。
- (4) 工业用地比例仅次于居住用地，是非居住用地中比重最大的一类。主要集中在几个位置，城市次级中心区域（开发区）、一些道路的沿路、城市边界等。
- (5) 商务用地与工业用地较为相似，两种用地常常相互结合而分布。



(A)



(B)

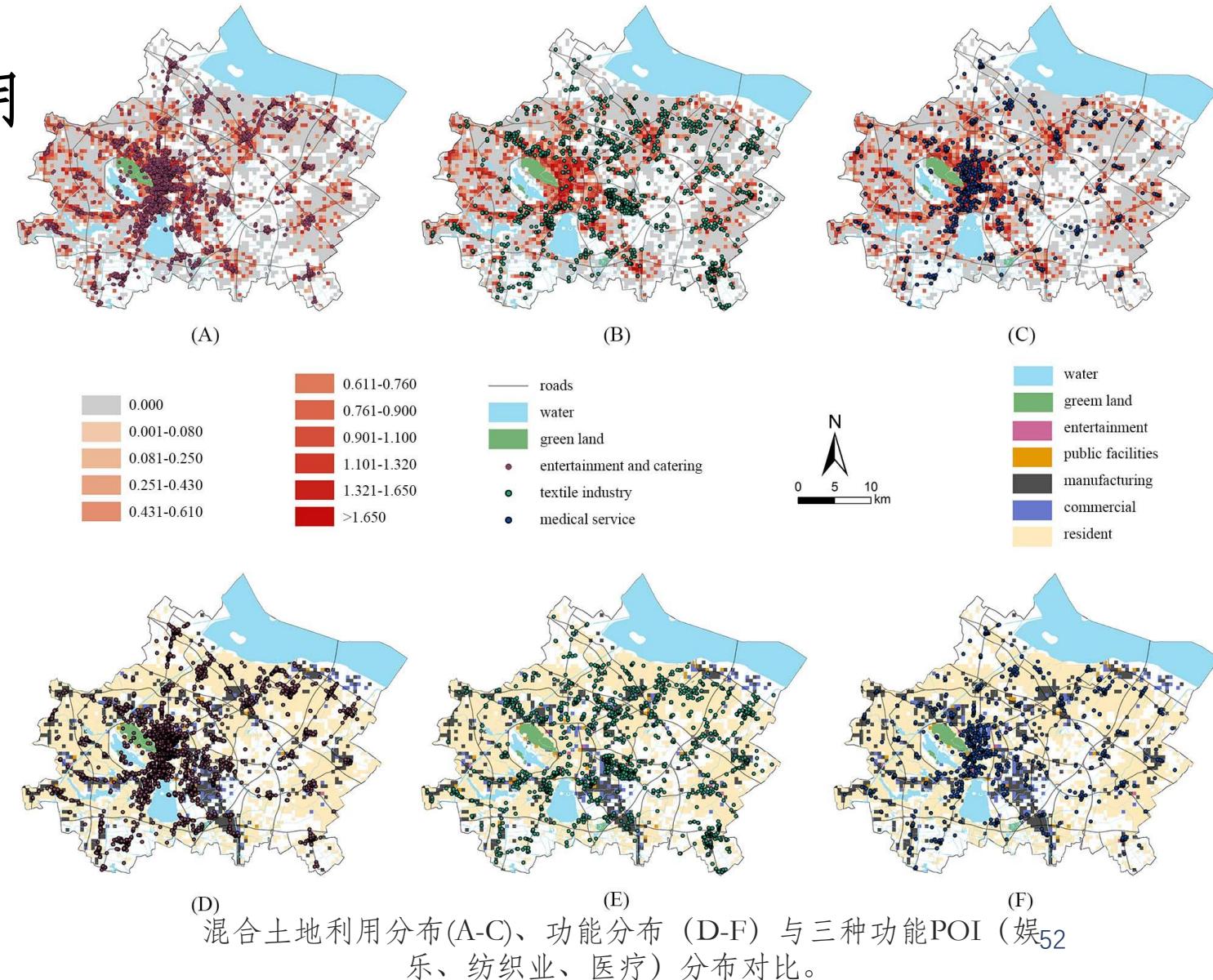
(A) 不同功能土地利用空间分布

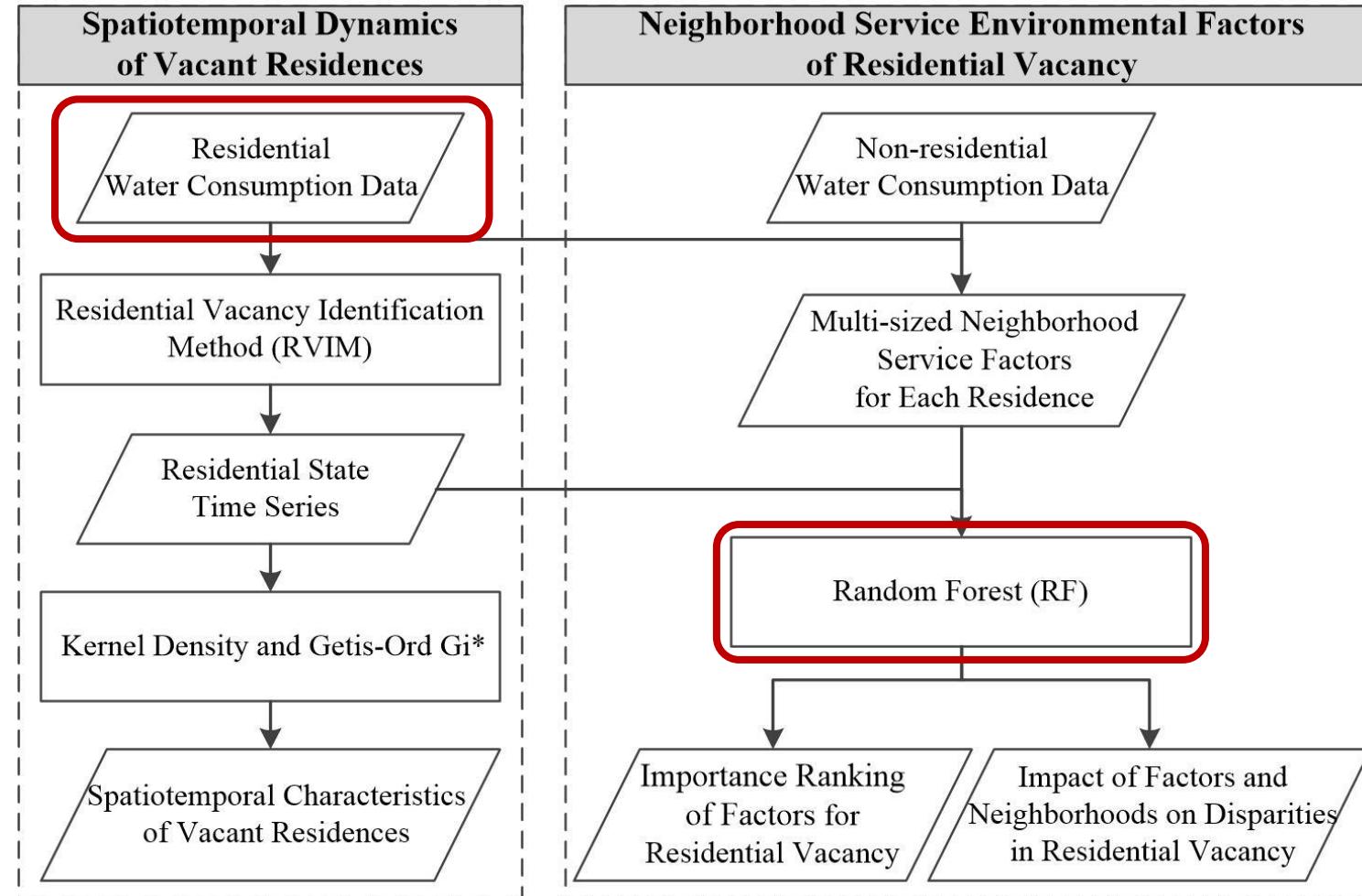
(B) 五种土地利用类型所占城市总用地的比重。

## ■ 设施分布与混合土地利用对比分析

- (1) 娱乐设施大多集中在混合度较高的区域。
- (2) 纺织业分布广泛，功能混合区、中心区、道路沿线、城市边界。
- (3) 医疗设施与娱乐设施相似，但就整体分布相对平衡。

→混合度较高的区域设施较为健全。  
→纺织业与居民生活结合紧密。



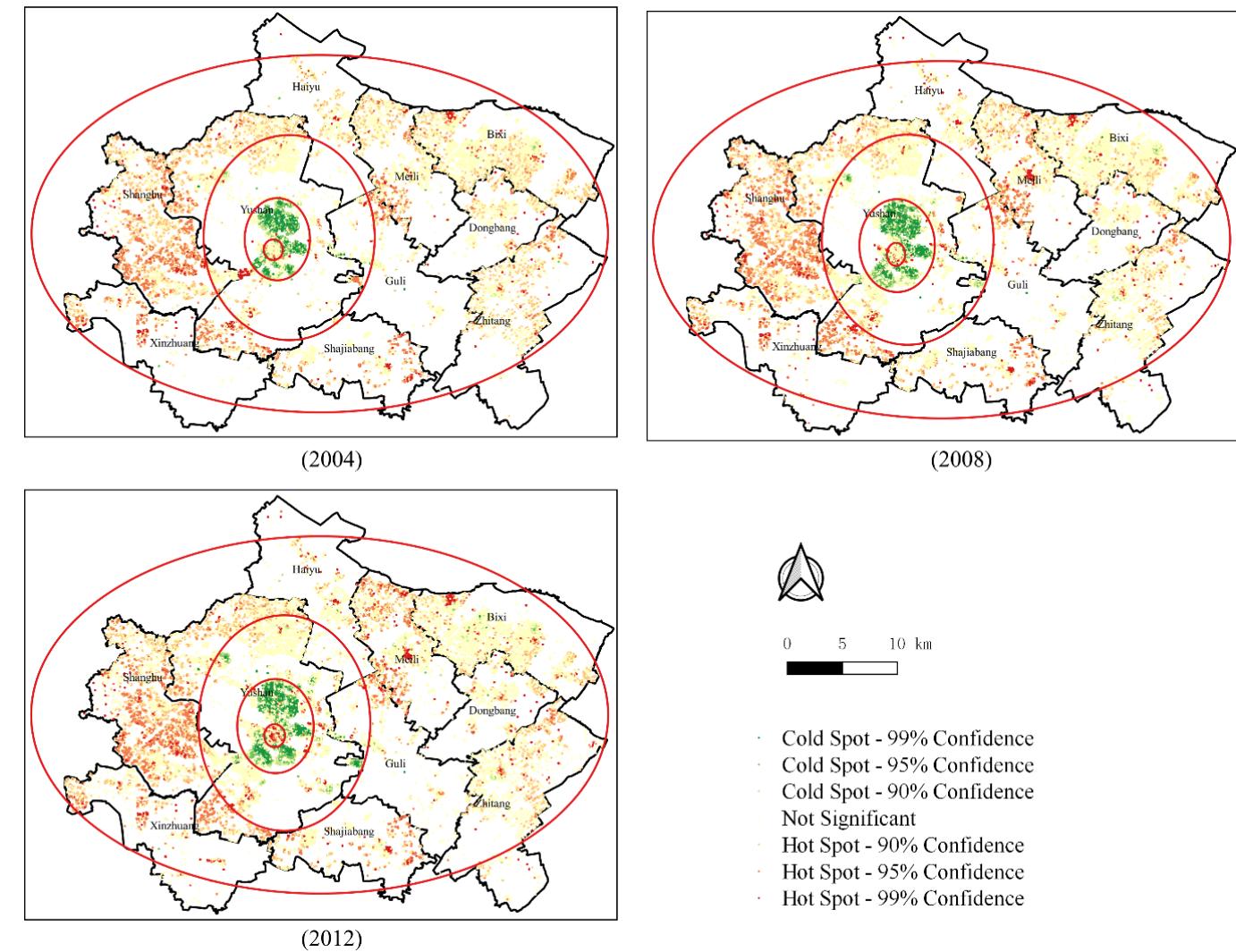


*Yongting Pan, Wen Zeng, Qingfeng Guan\*, Yao Yao, Xun Liang, Hanqiu Yue, Yaqian Zhai, Junyi Wang. 2020. Spatiotemporal dynamics and the contributing factors of residential vacancy at a fine scale: A perspective from municipal water consumption. Cities. DOI:10.1016/j.cities.2020.102745.*

## ■ 空置住宅时空特征

### ➤ 同心环结构

- 衰退的内城
- 繁荣的中心城区
- 近郊区
- 远郊区

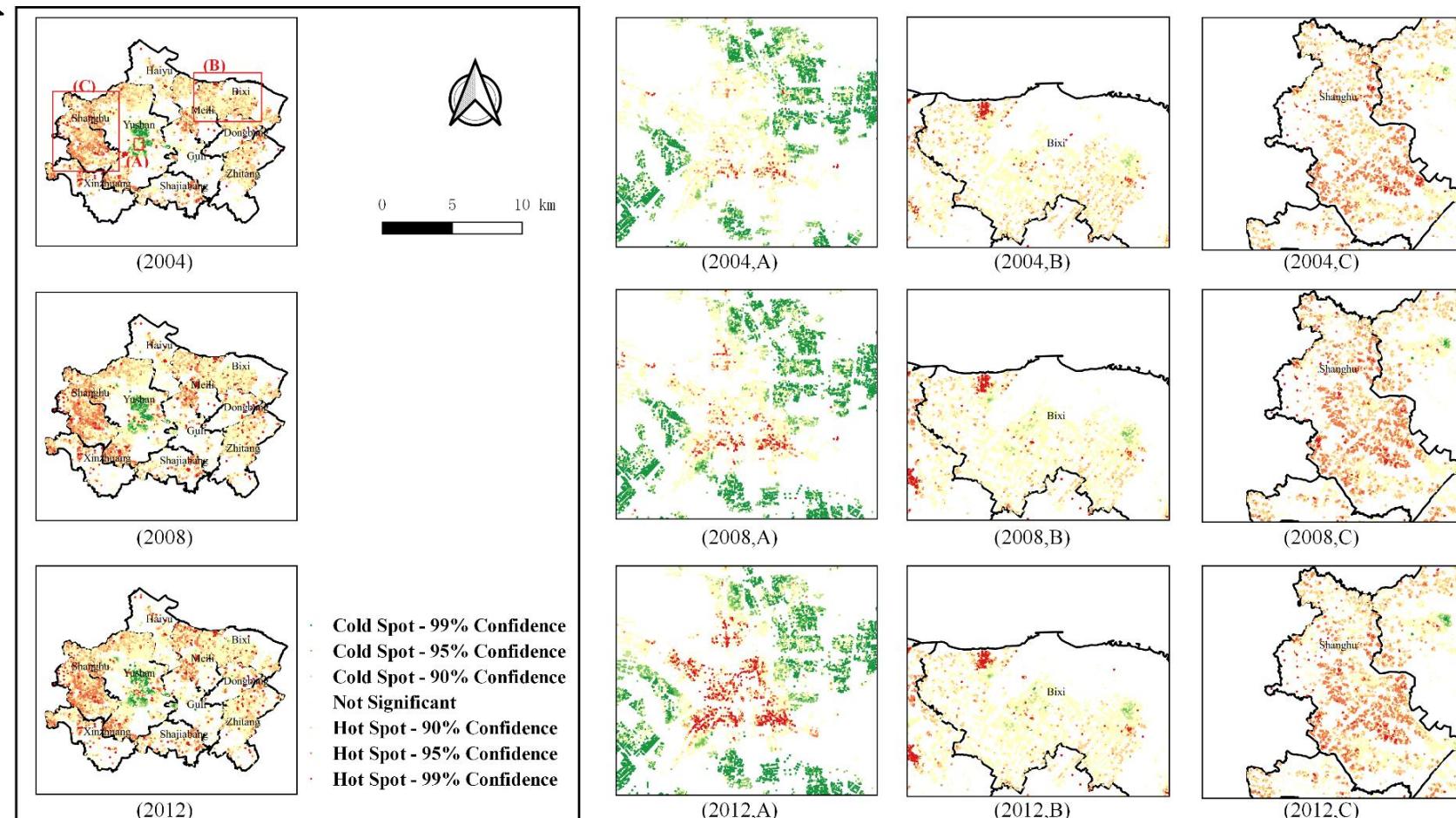


2004年、2008年和2012年常熟市的住宅热/冷点分布

## ■ 空置住宅时空特征

### ➤ 住宅形态改变

- 内城衰退(A)
- 副中心的崛起(B)
- 替代(A和B) 或增生(C)

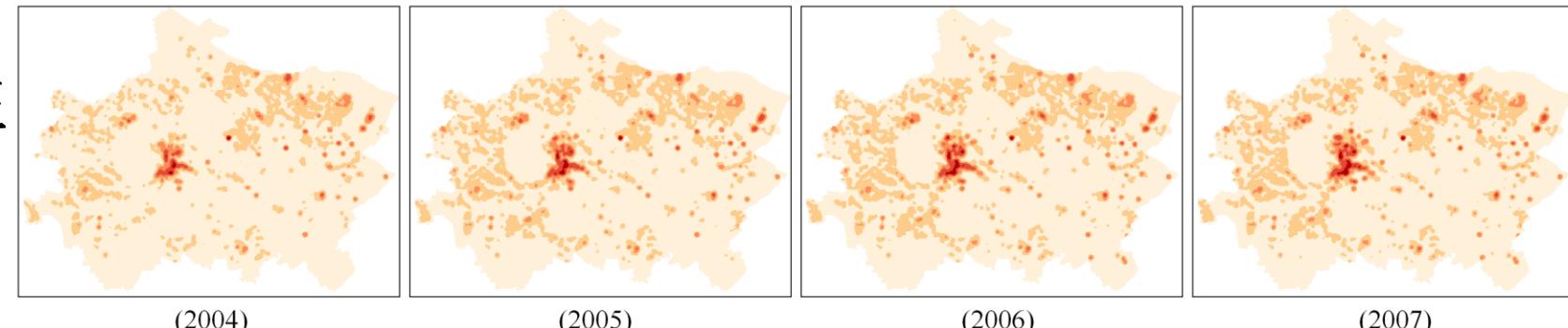


2004年，2008年和2012年常熟市的住宅热/冷点分布

## ■ 空置住宅时空特征

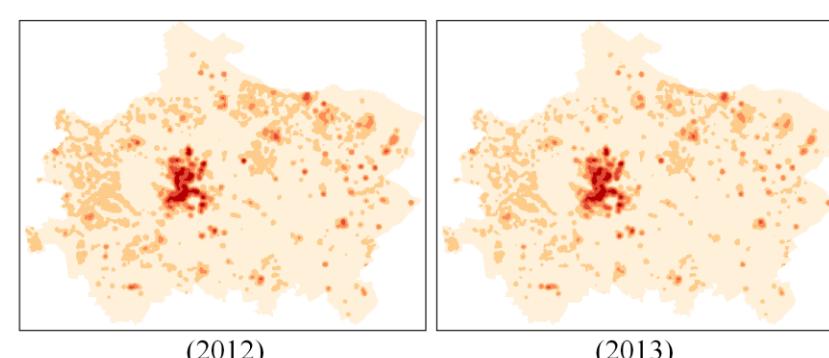
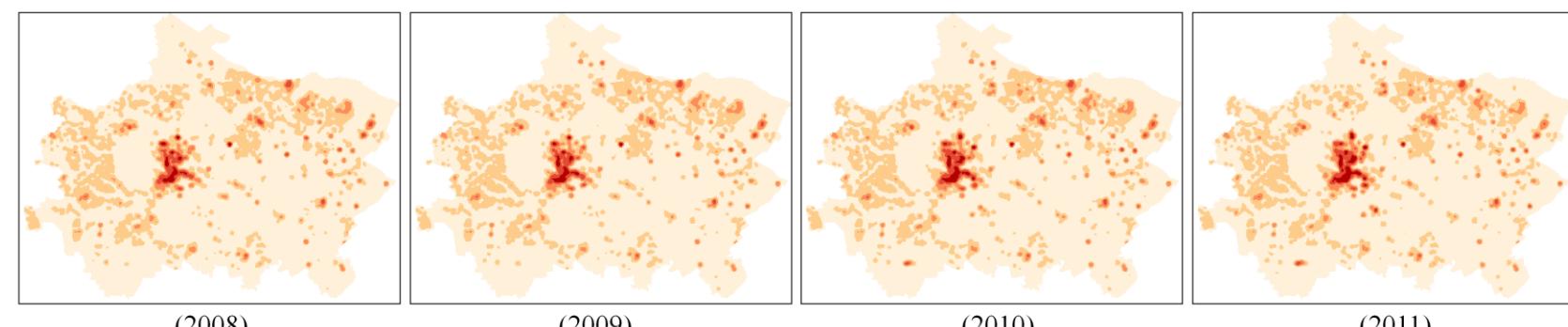
### ➤ 二元结构

- 中心
- 外围



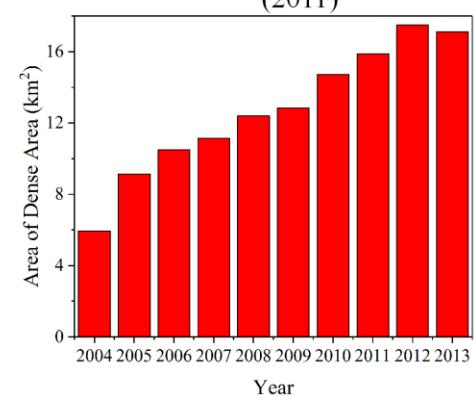
### ➤ 空置住宅的密集区

- 密集区**稳步扩大**
- 密集区面积**逐年线性快速增加**



### Kernel density

- |               |
|---------------|
| 0 - 870       |
| 870 - 3615    |
| 3615 - 9550   |
| 9550 - 18515  |
| 18515 - 36900 |



2004年至2013年常熟居住空置核密度

## ■ 模型实施及精度评估

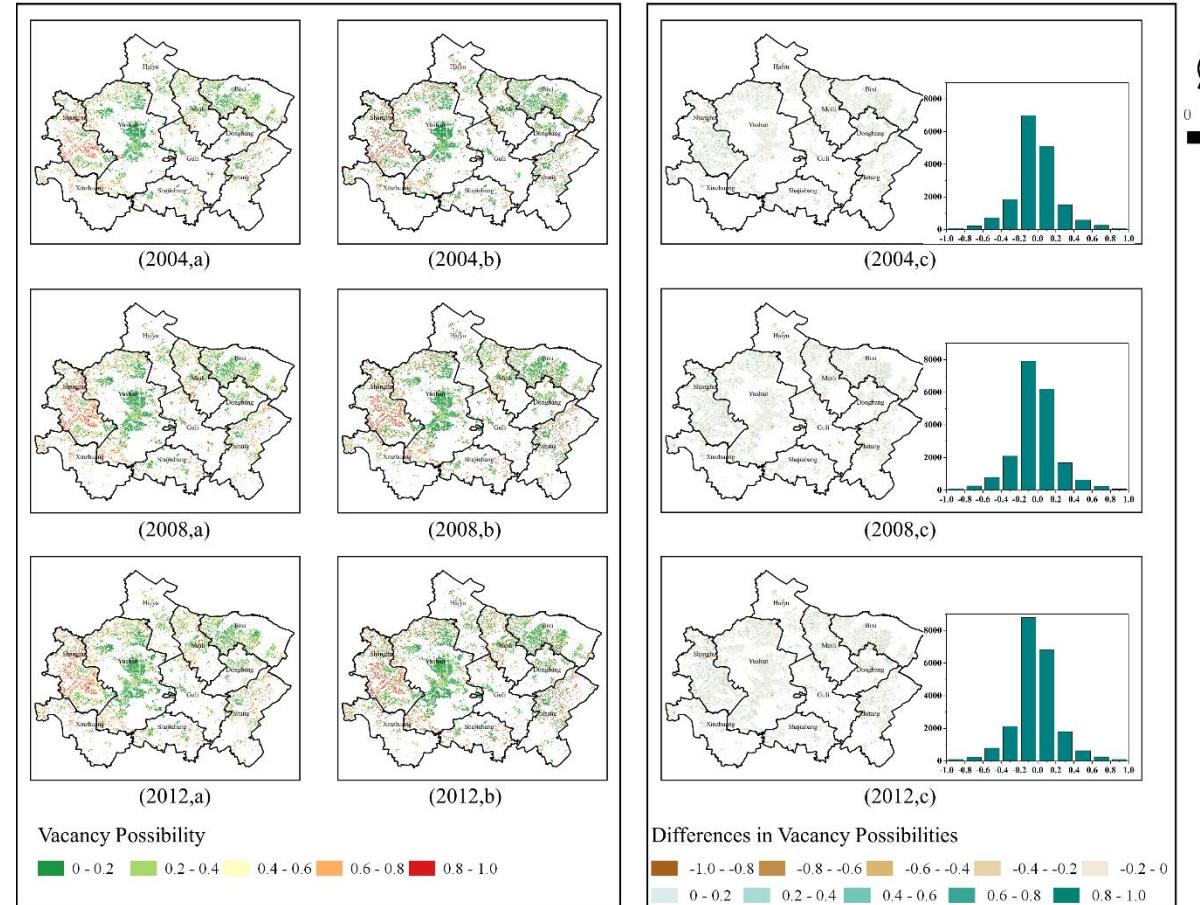
### ➤ 随机森林

- 训练集和测试集比例: 70%, 30%
- 树的数量(ntree) : 100
- 每次拆分时随机抽样的变量数(mtry) :  $\log_2(100)$
- 工具: Python + sklearn

➤ 总体精度: 0.758

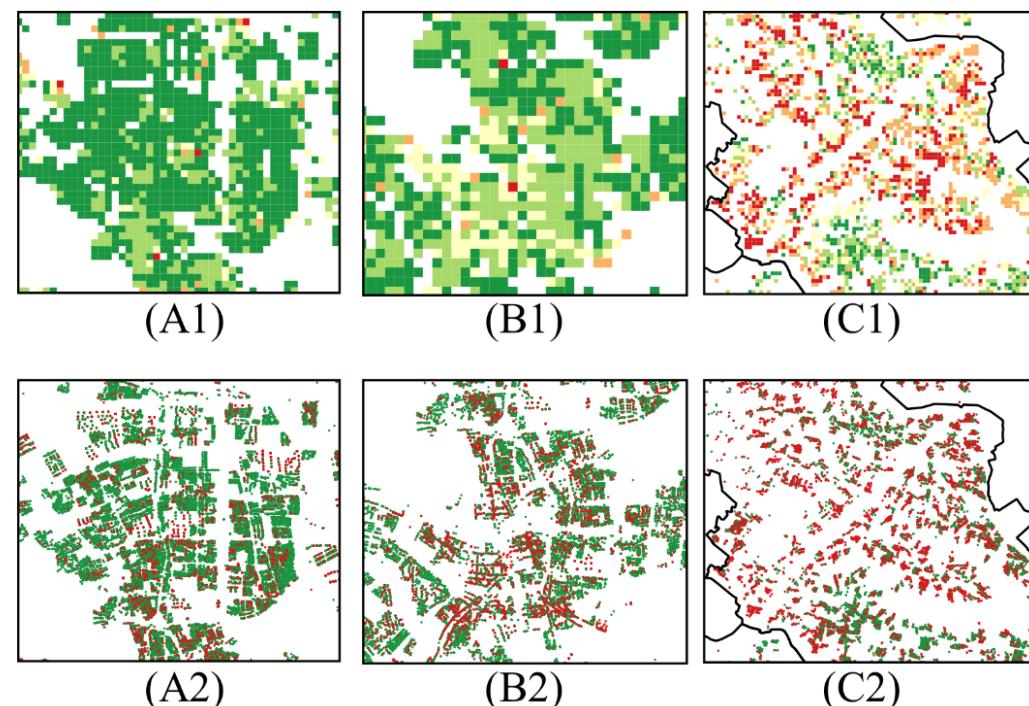
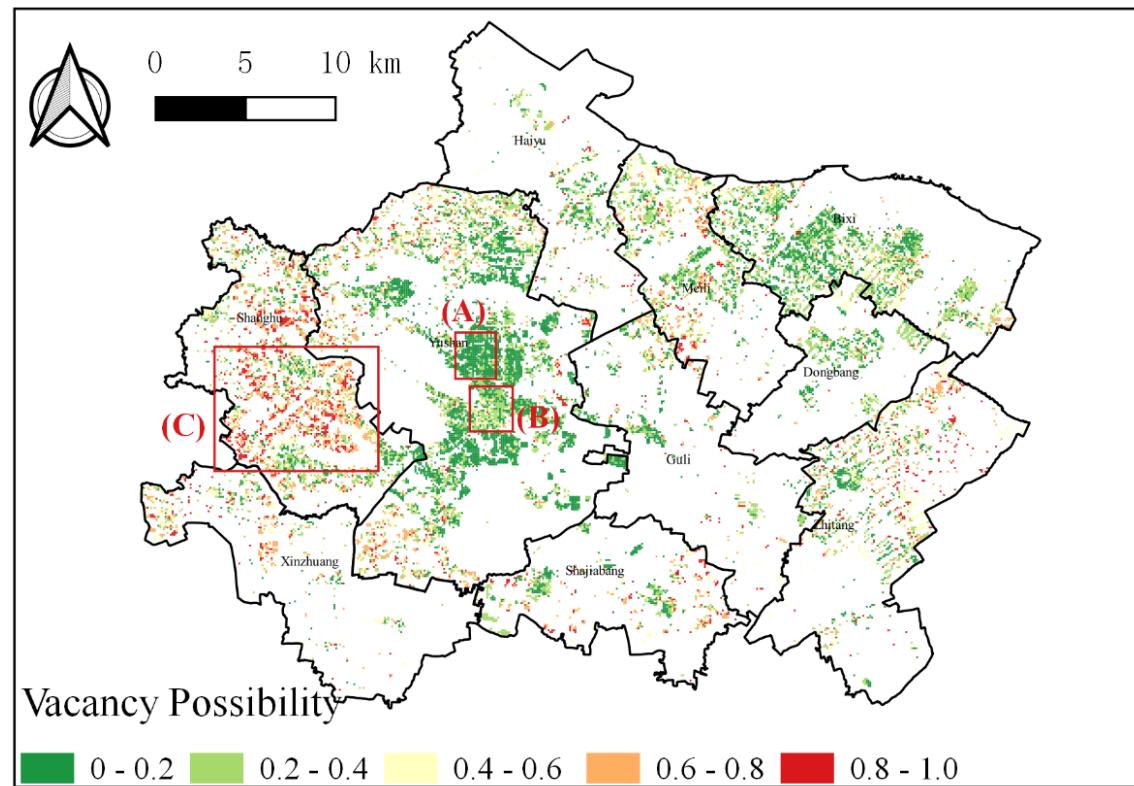
Year	Overall Accuracy	Year	Overall Accuracy
2004	0.772	2009	0.755
2005	0.765	2010	0.756
2006	0.763	2011	0.755
2007	0.761	2012	0.752
2008	0.756	2013	0.738

### ➤ 空间精度



2004年, 2008年和2012年常熟市100m \* 100m分辨率预测空置  
(a), 实际空置 (b) 及差异 (c) 的空间分布 57

## ■ 模型实施及精度评估



2012年常熟市三个典型区（A、B和C）的100m \* 100m分辨率的预测空置长度（#1）和每个住宅实际空置长度（#2）的细节图。（A）空置长度小于3个月，（B）空置长度为5-7个月，（C）空置长度大于10个月。

## ■ 住宅空置驱动因子

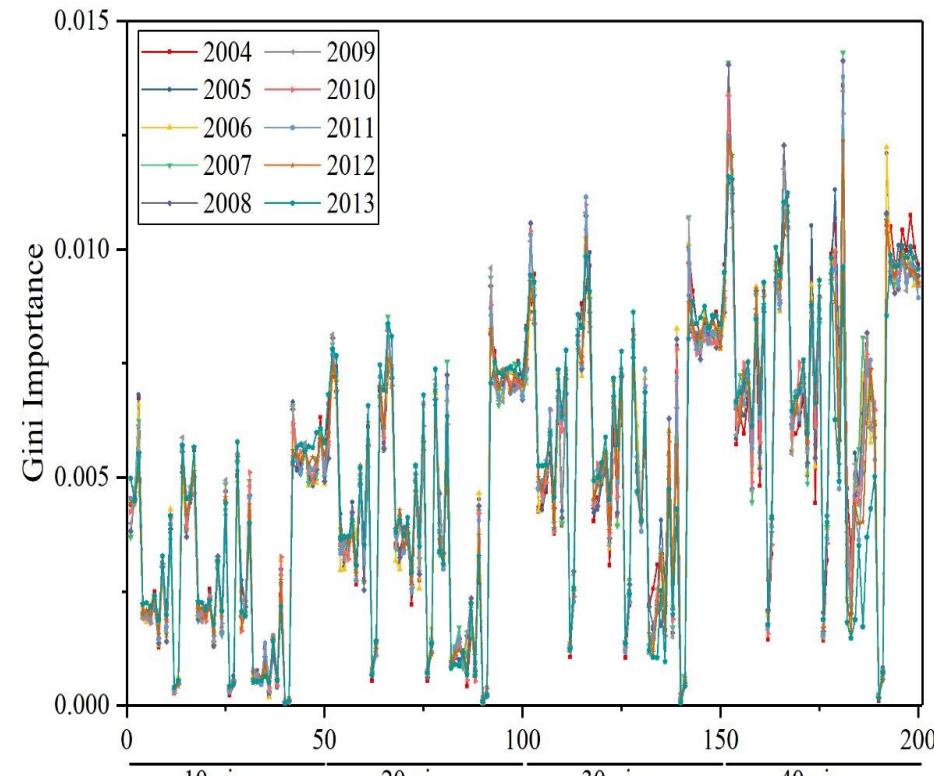
➤ 多样性是最重要的因素。

➤ 最重要的3个FASs:

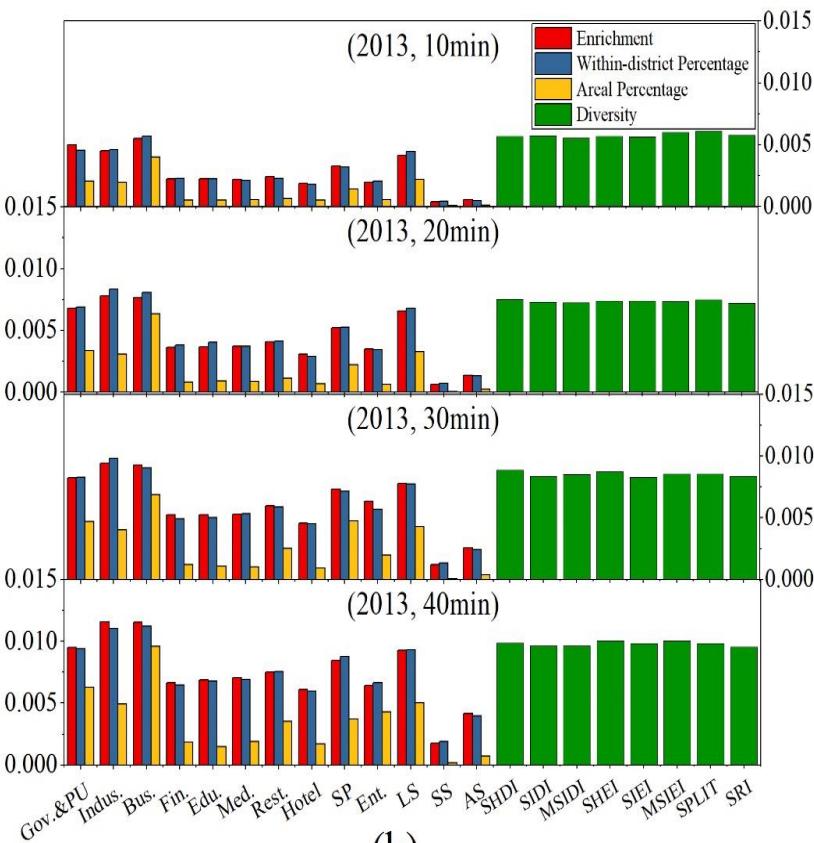
- **BUS, IND and LIF** (10 / 20  
/ 30分钟邻域内)
- **BUS, IND and GOV** (40分  
钟邻域内)

➤ 最不重要的2个 FASs:

- **TOU and AUT**



(a)



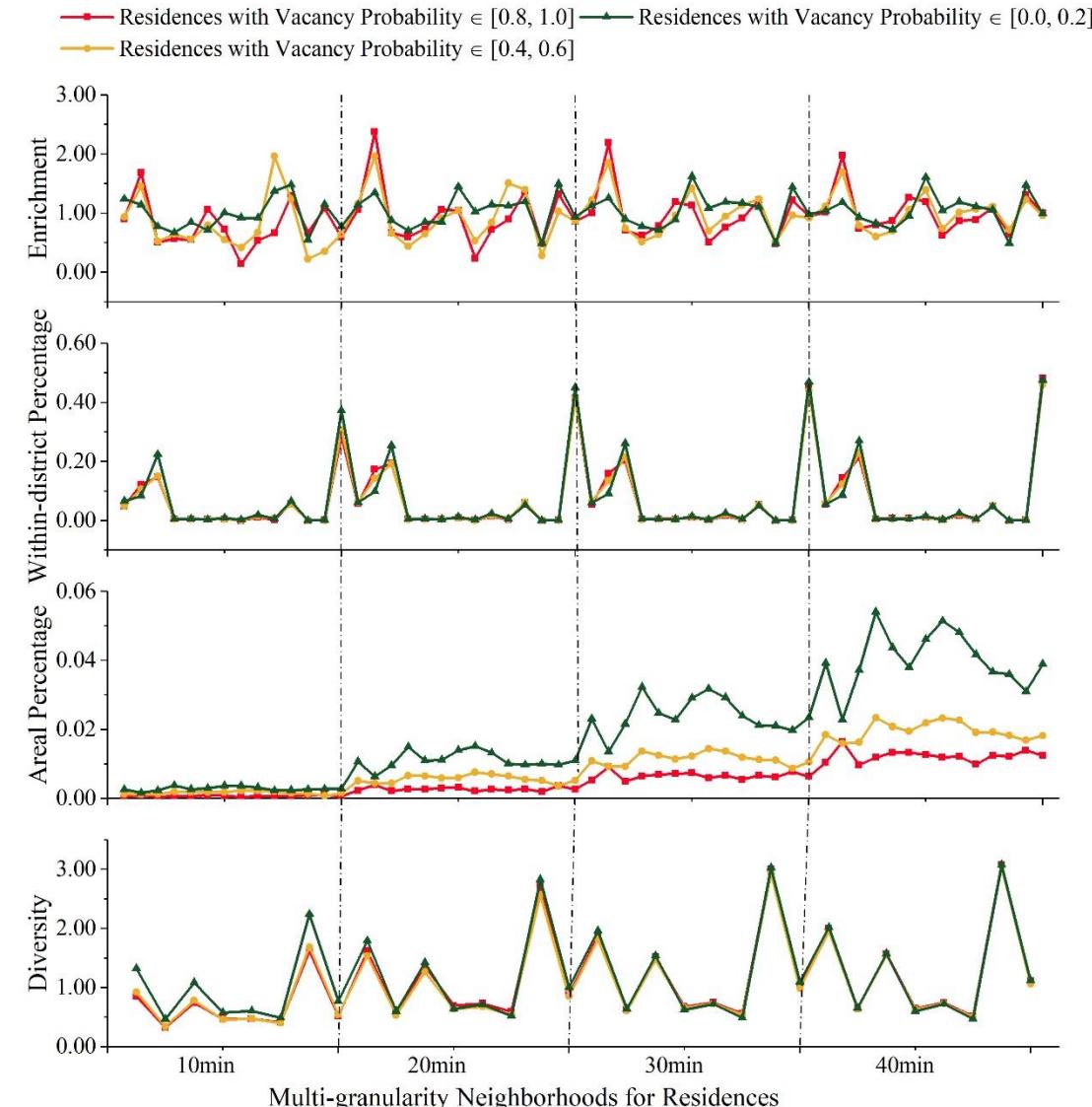
(b)

常熟市2004-2013年10分钟，20分钟，30分钟和40分钟邻域内影响住宅空置的因素  
重要性排名(a)和2012年细节展示(b)

## ■ 住宅空置驱动因子

住宅空置长度越低

- 富集因子越接近1
- 类型占比和多样性越高
- 邻域内占比变化不大



导致住宅空置相异的因素差异与邻域变化

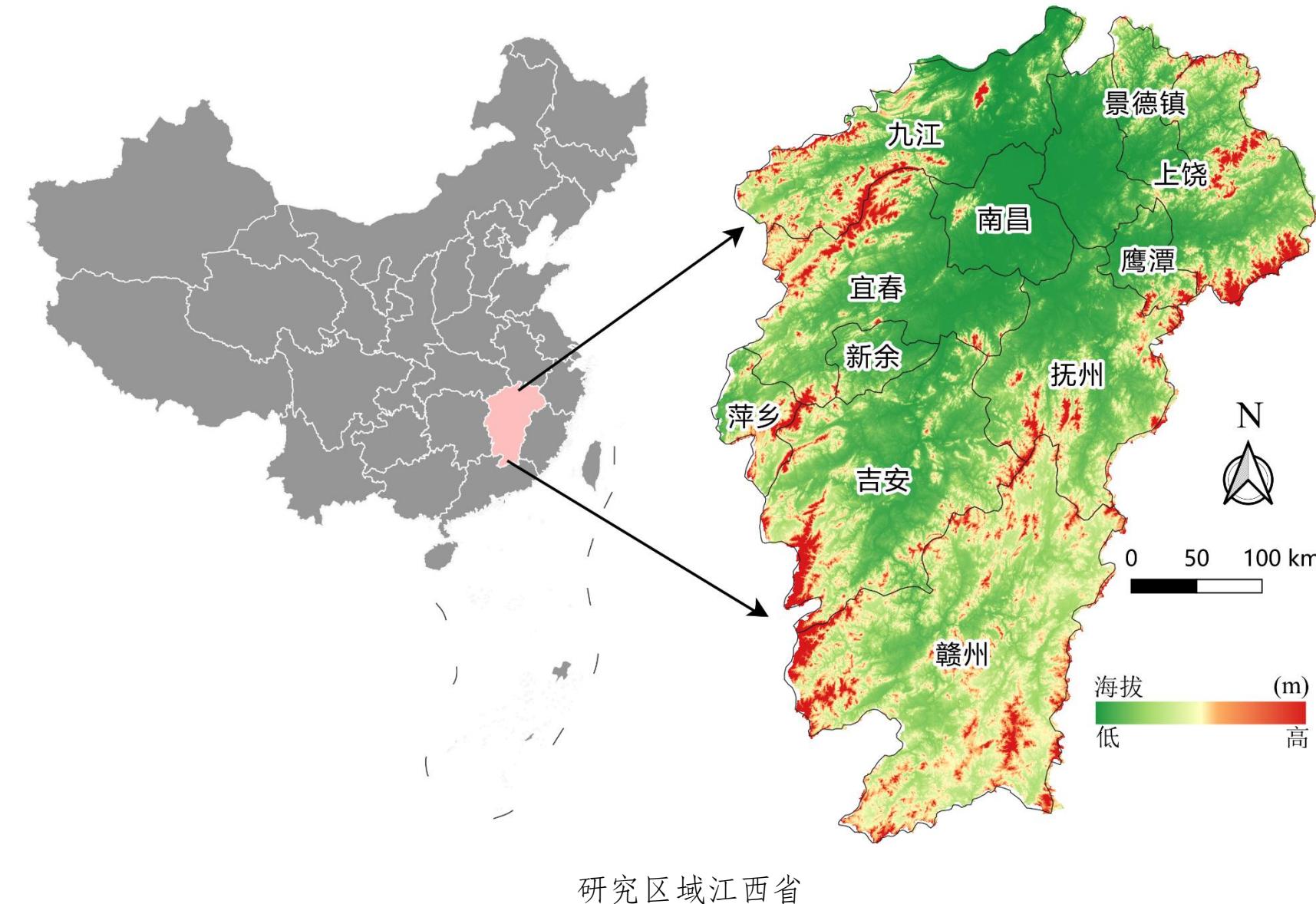
# 04

---

## 智慧电网感知城市社会经济发展



High-performance Spatial Computational Intelligence Lab @ CUG



本研究的研究区域为中国的江西省。目前江西省下辖11个地级市行政区。江西省是自然条件优越的农业大省，第一产业在GDP产值的比重高于全国平均水平达到17%。江西省不同地市产业结构差异明显。（中华人民共和国国家统计局 2011）

## ➤ 电网数据

变电站电力消耗数据样例

站点号	采集时间	结束采集时间	电流 (A)	电压(V)	电功率(w)
23**30	2018/1/1 0:00	2018/1/2 15:51	0.371	237.800	0.242
13**58	2018/1/1 0:00	2018/1/2 15:51	0.830	228.500	0.537
10**96	2018/1/1 0:00	2018/1/2 15:51	0.353	235.100	0.212
:	:	:	:	:	:

变电站站点信息数据样例

市公司	县公司	台区 编号	变电站 名称	线路名称	配变名称
国网南昌供 电公司	国网江西 南昌县	23**46	黄湖35kV变 电站	10kV五丰线915开 关间隔	10kV五丰线蒋巷所五丰外范01 号公变
国网南昌供 电公司	国网江西 南昌县	23**44	黄湖35kV变 电站	10kV五丰线915开 关间隔	10kV五丰线蒋巷所五丰治安片 01号公变
国网南昌供 电公司	国网江西 南昌县	23**43	黄湖35kV变 电站	10kV五丰线915开 关间隔	电网_10kV五丰线蒋巷所五丰 十八队01号公变
:	:	:	:	:	:

电力消耗时序数

据是本实验最重要的数据，本研究使用了江西省2018年18万个变压器1-8月每隔半小时所记录的用电消耗数据。

变压器地址数据

是电网数据中的一部分。本研究共对江西省18万个电站进行地址匹配。

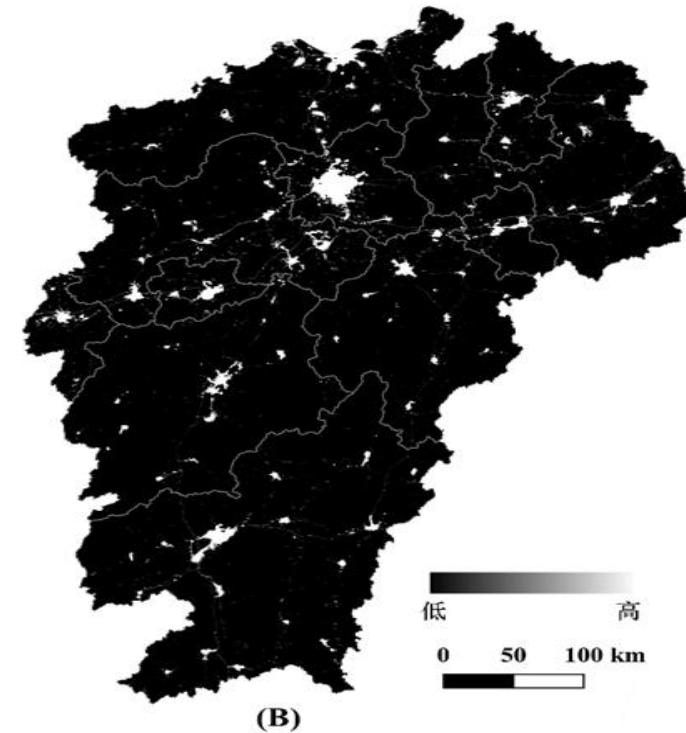
## ➤ 夜间灯光数据

夜间灯光数据主要来自于由美国国防气象卫星搭载的可见光成像线性扫描业务系统（DMSP/OLS）和国家极轨卫星搭载的可见光近红外成像辐射仪（NPP/VIIRS）获取的夜间灯光影像数据源。



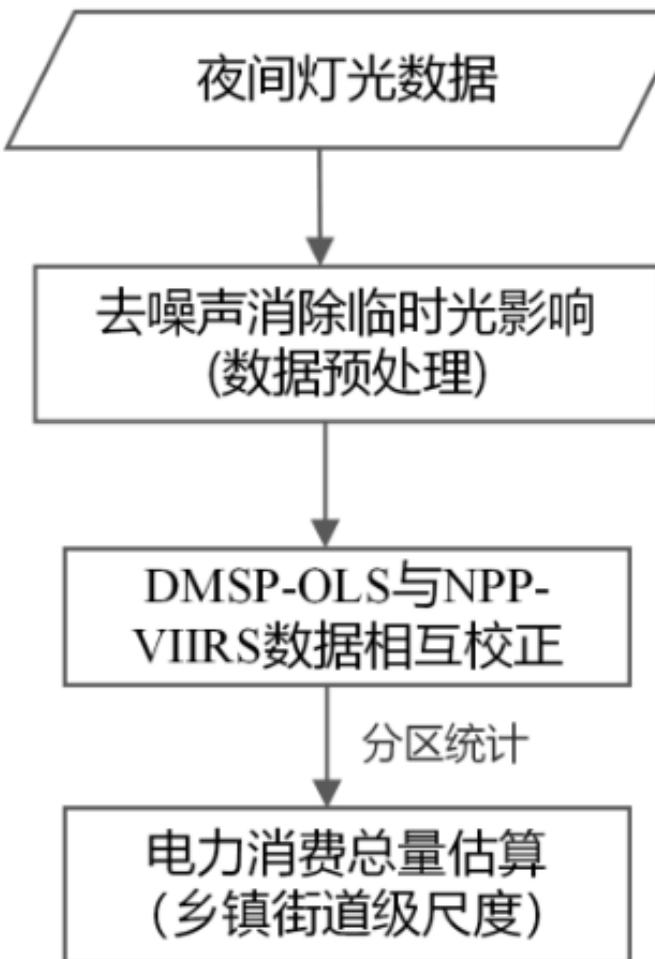


(A) 2010年DMSP-OLS数据  
传感器：DMSP-OLS  
空间分辨率：1KM

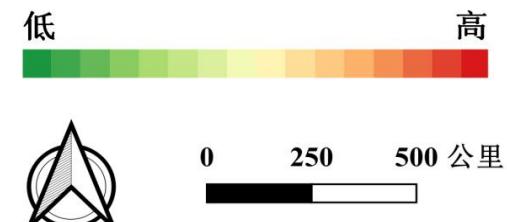
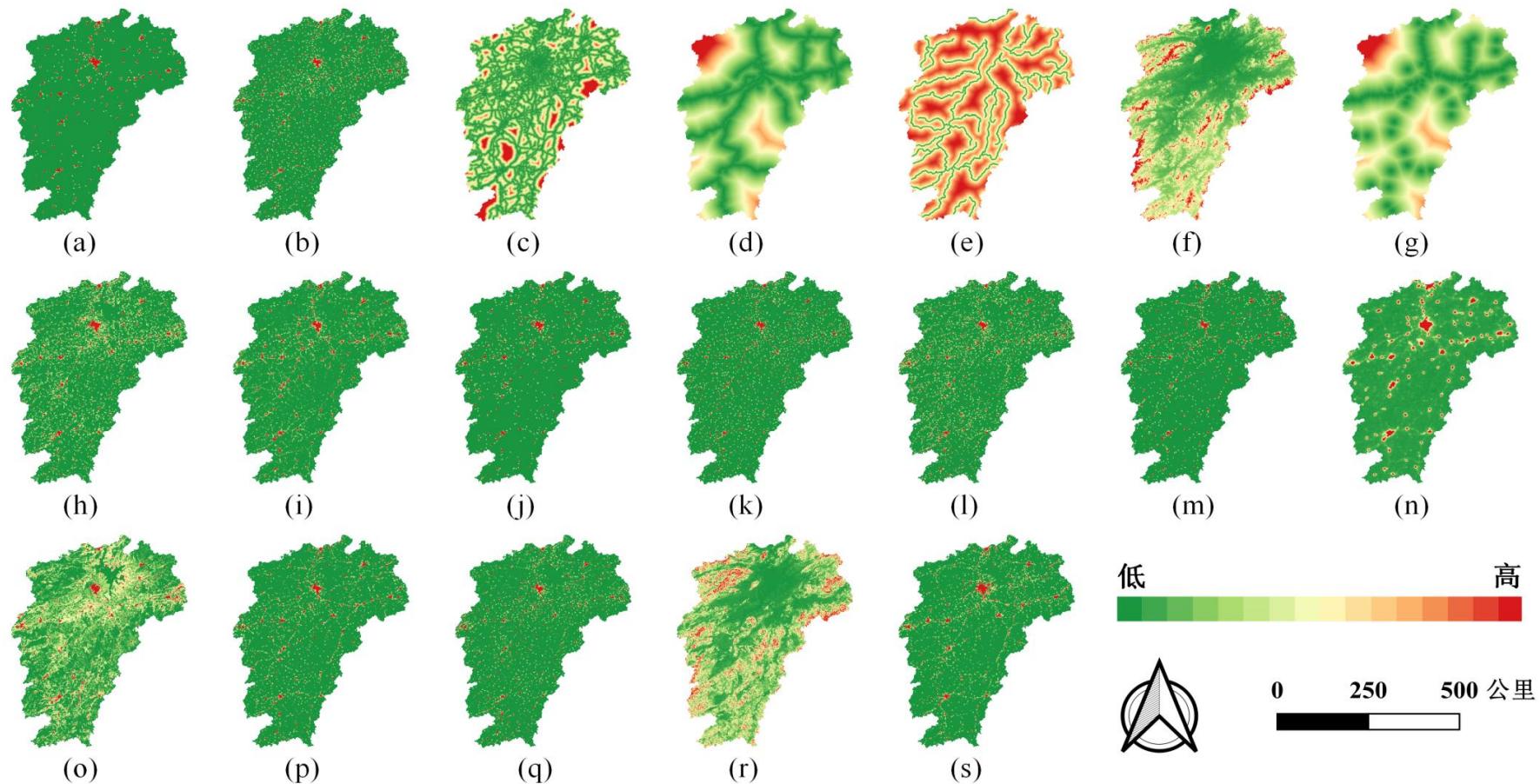


(B) 2015年NPP-VIIRS数据  
传感器：NPP-VIIRS  
空间分辨率：500M

预处理：



## ➤ 多源空间数据



多源空间数据集

## ► 江西省各区县GDP数据

江西省各区县GDP构成数据示例

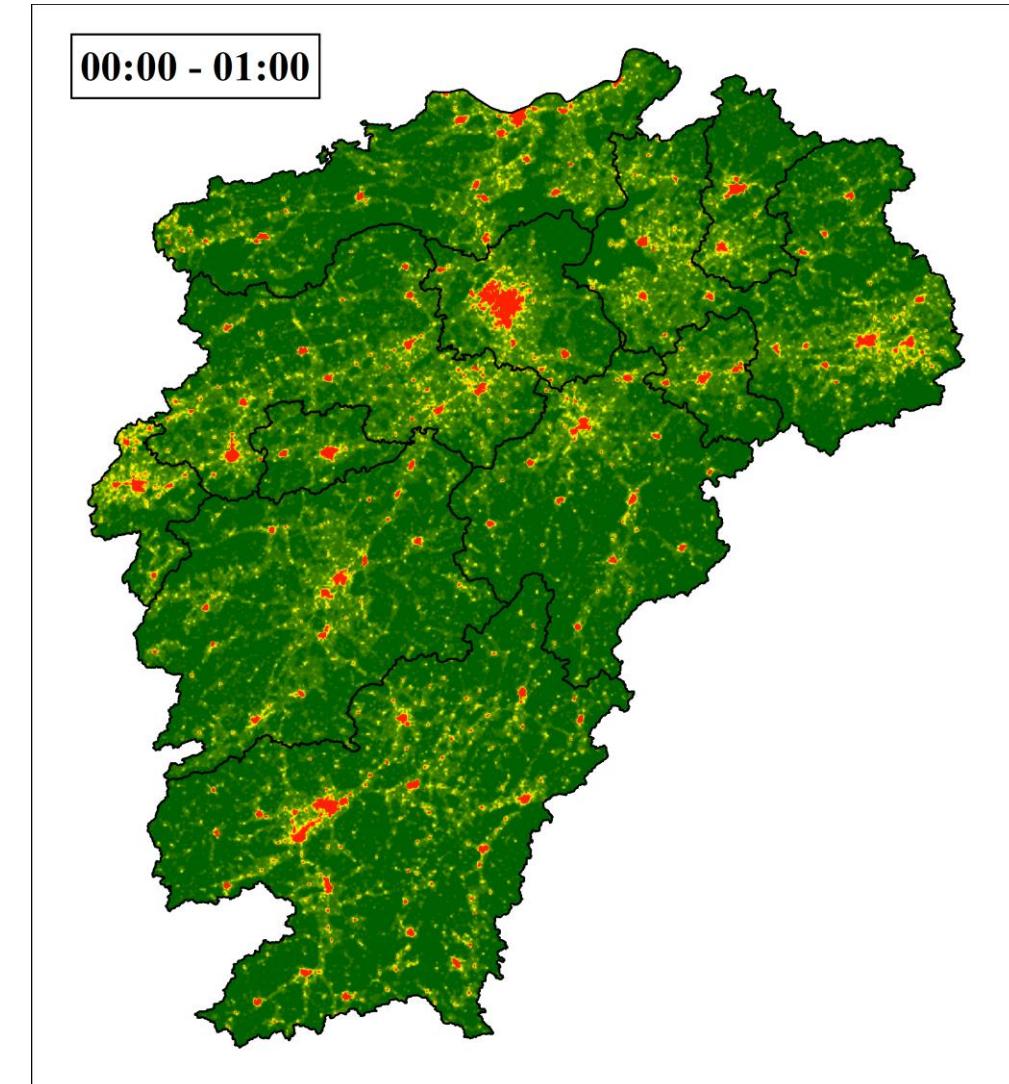
省名	地级市名	区县名	全年GDP总额	第一产业增长总额	第二产业增长总额	第三产业增长总额
江西省	抚州市	乐安县	624,481	106,605	218,776	299,100
江西省	抚州市	金溪县	837,158	129,862	314,897	392,399
江西省	抚州市	广昌县	629,978	106,586	251,918	271,474
江西省	抚州市	南城县	1,221,477	175,612	484,264	561,601
江西省	抚州市	宜黄县	678,640	95,386	321,789	261,465
:	:	:	:	:	:	:

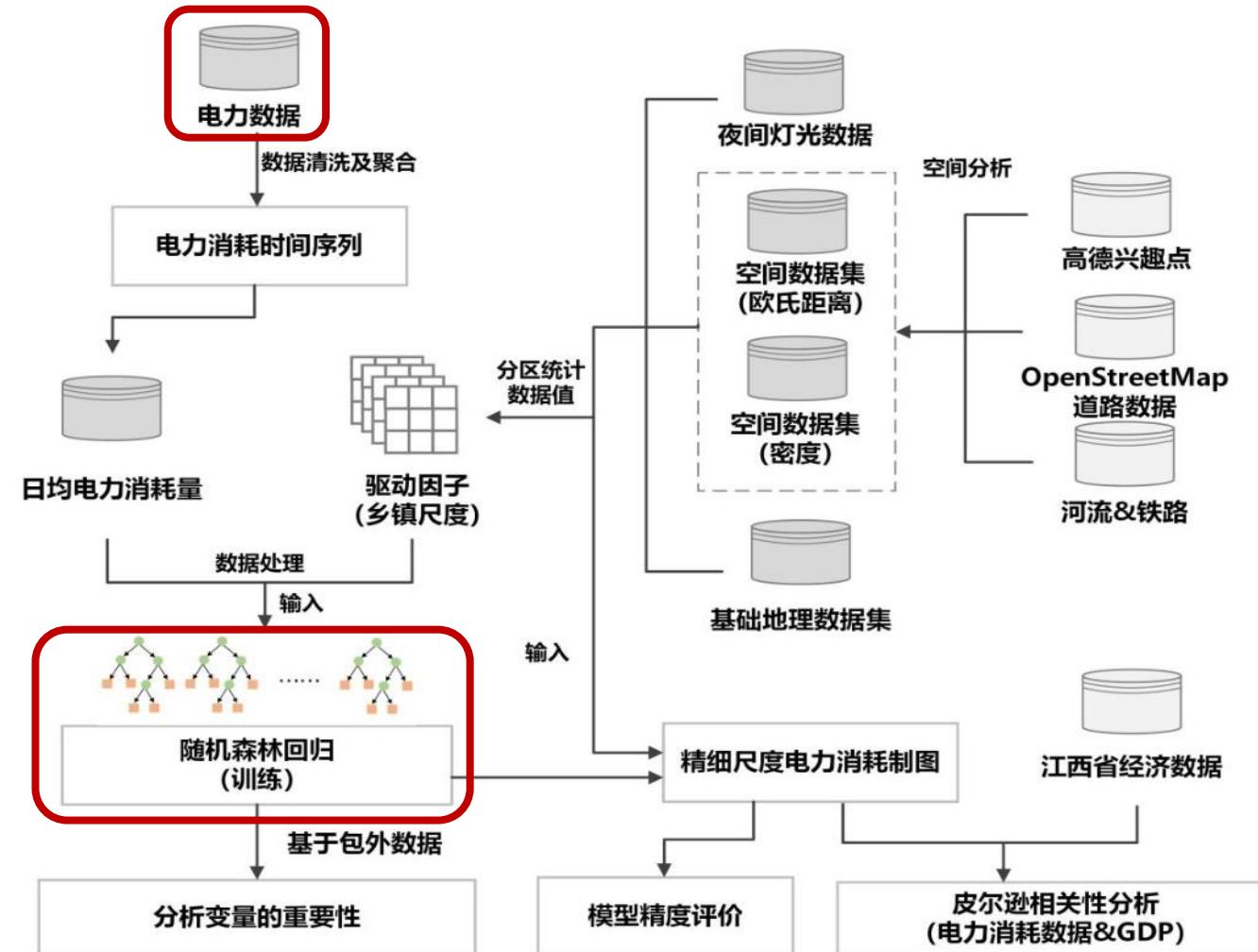
GDP常被公认为是衡量国家经济状况的最佳指标。

数据来源：GDP数据一般来源于国家统计局网站和各地区统计局网站。

➤ 社交媒体用户分布时序数据

- 右图为江西省2015年某工作日腾讯用户数量24小时的空间分布情况；
- 可以看到用户数量的空间分布随一天内会发生变化，这在一定程度上反映了不同区域的职能情况，因此可以基于每个乡镇内的用户数量变化情况进行职能划分。

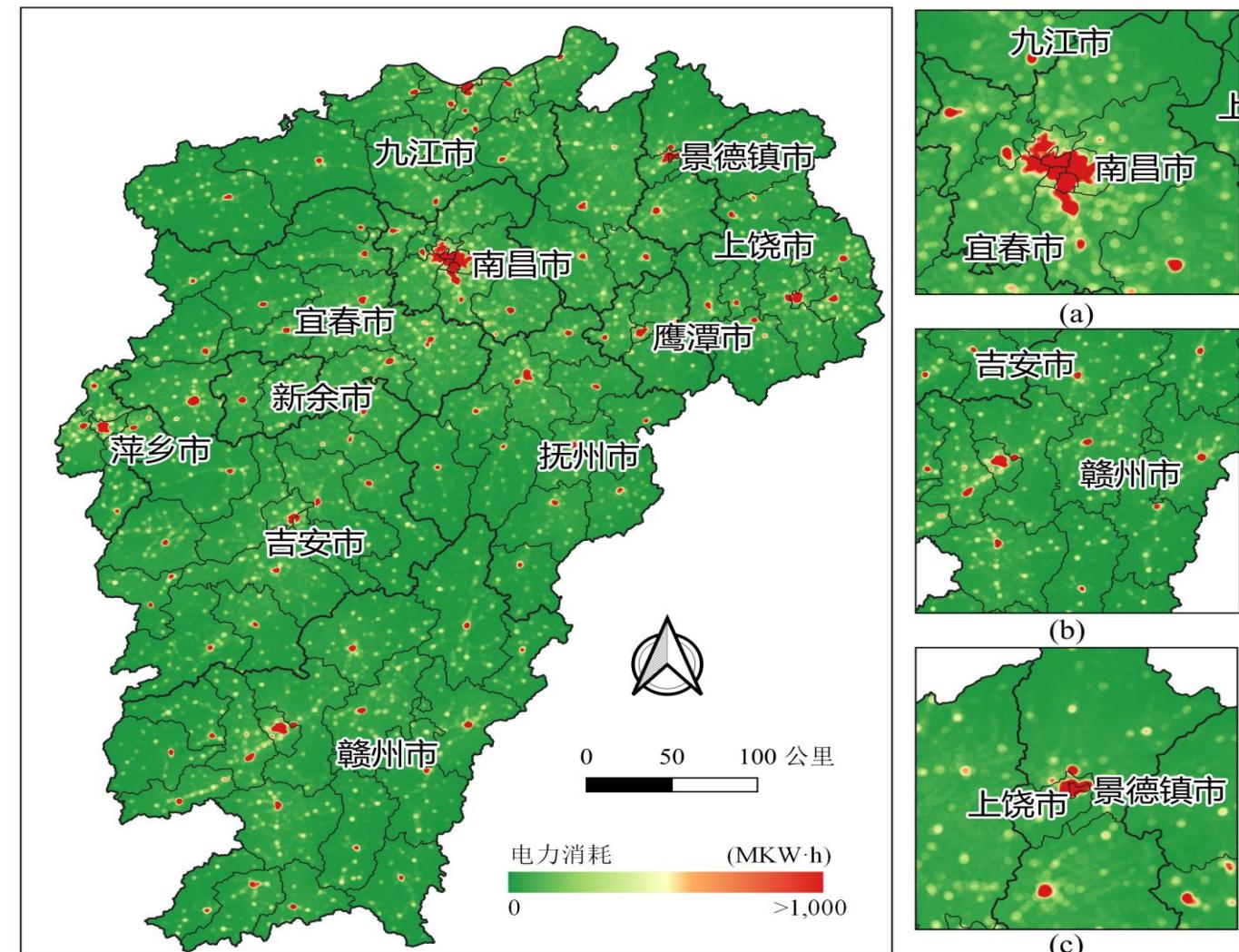




耦合江西省电力消耗数据和空间数据集建立江西省电力消耗模型

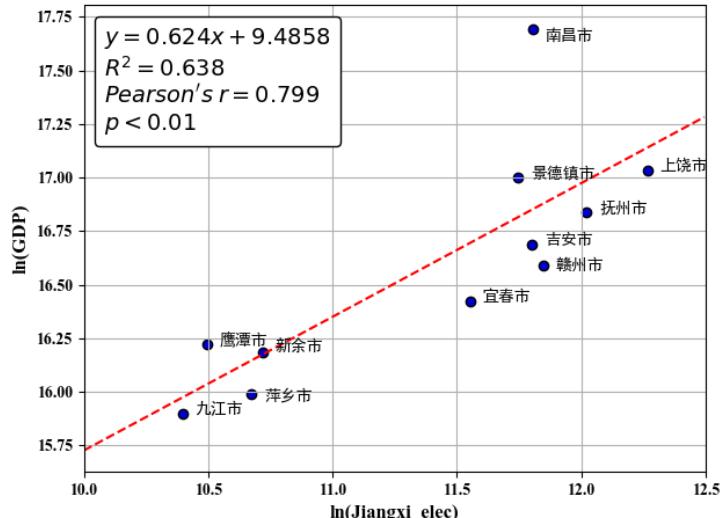
## ■ 精细尺度的电力消耗制图

- ① 江西省电力消耗总体呈现出较为显著的“南低北高”的空间分布格局。
- ② 用电消耗较多的区域均位于发达的中心区域，如江西省省会南昌。
- ③ 赣州市发展水平较低，电力消耗量均低于平均水平。

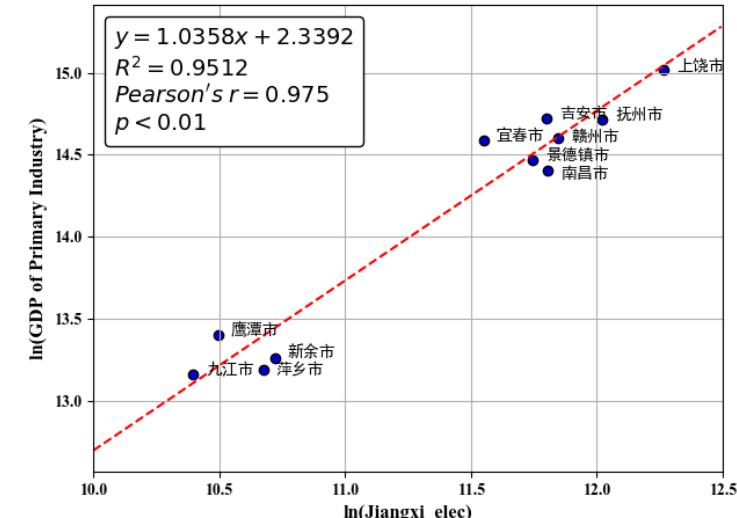


江西省电力消耗分布图，(a)代表南昌地区；(b)代表江西南部地区；(c)代表江西东北部地区

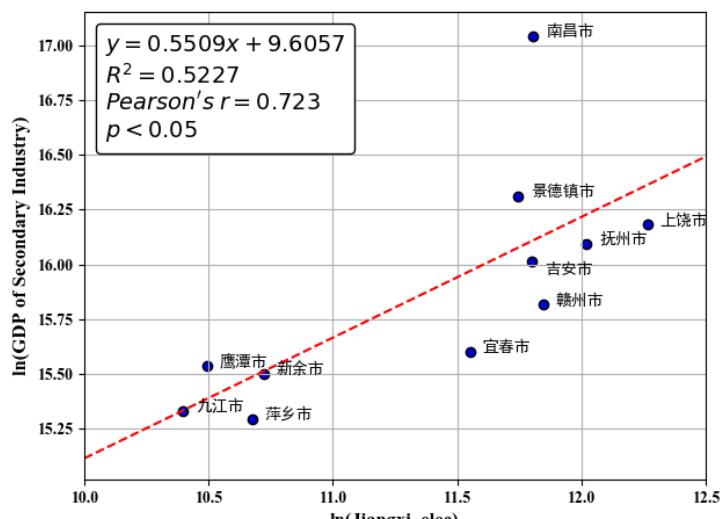
## ■ 电力消费与产业发展



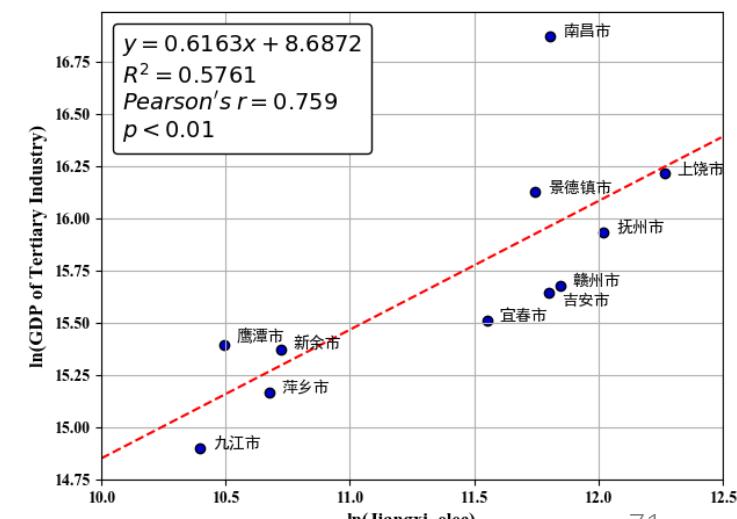
江西省各市电力消耗和GDP总量



江西省各市电力消耗和第一产业GDP



江西省各市电力消耗和第二产业GDP

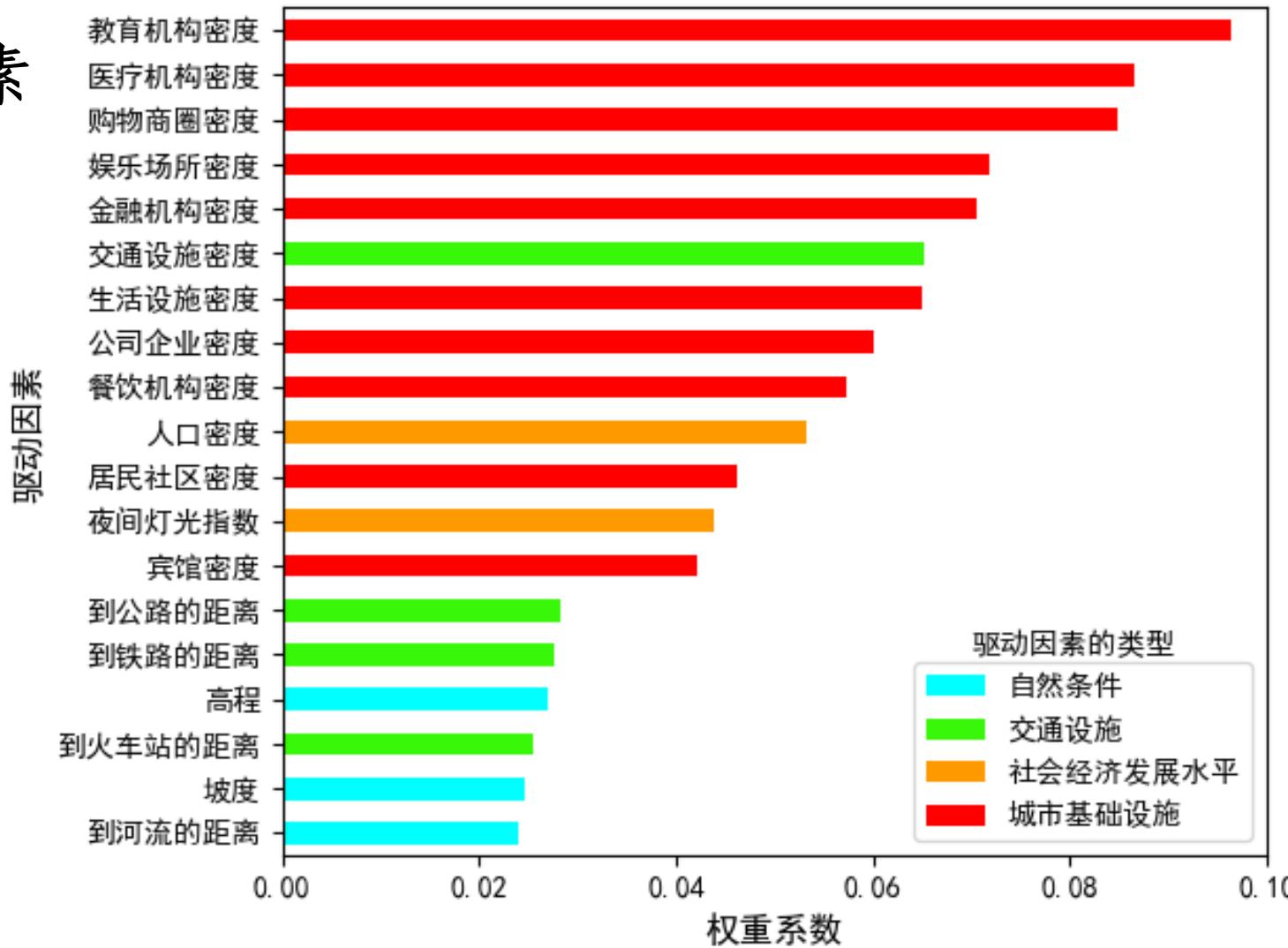


江西省各市电力消耗和第三产业GDP

## ■ 江西省电力消费驱动因素

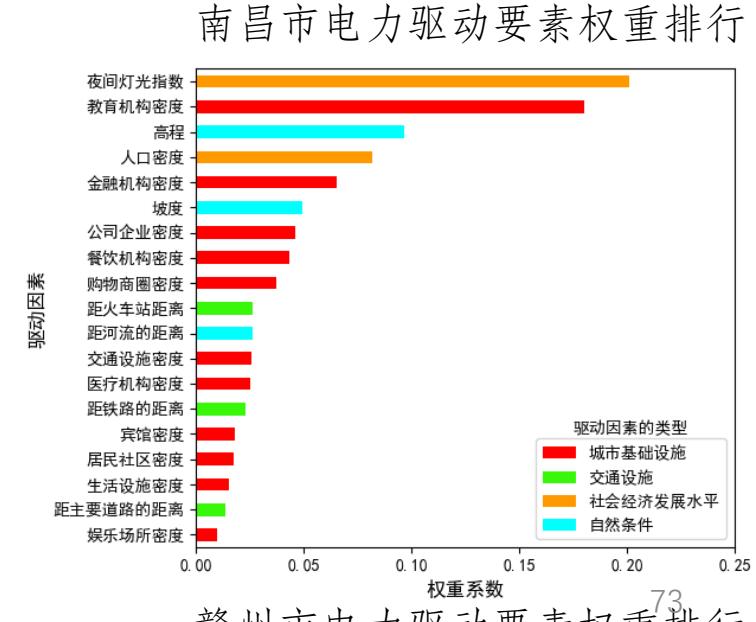
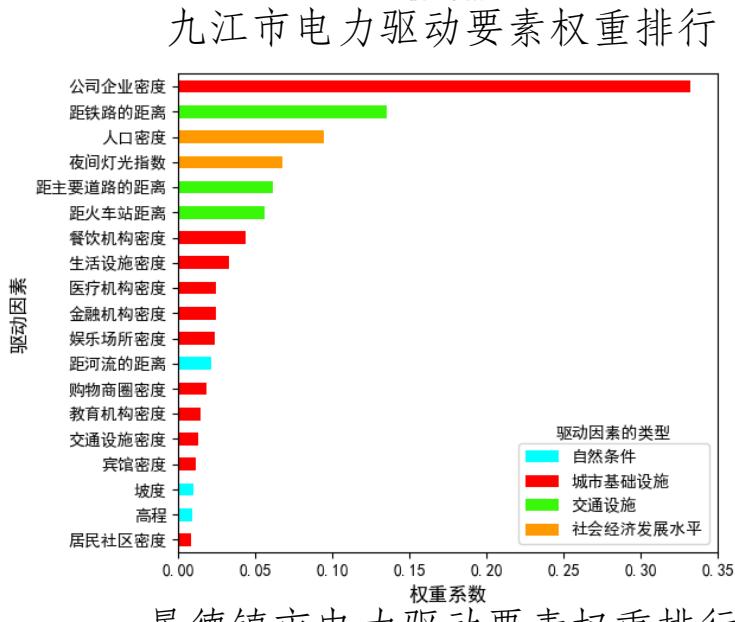
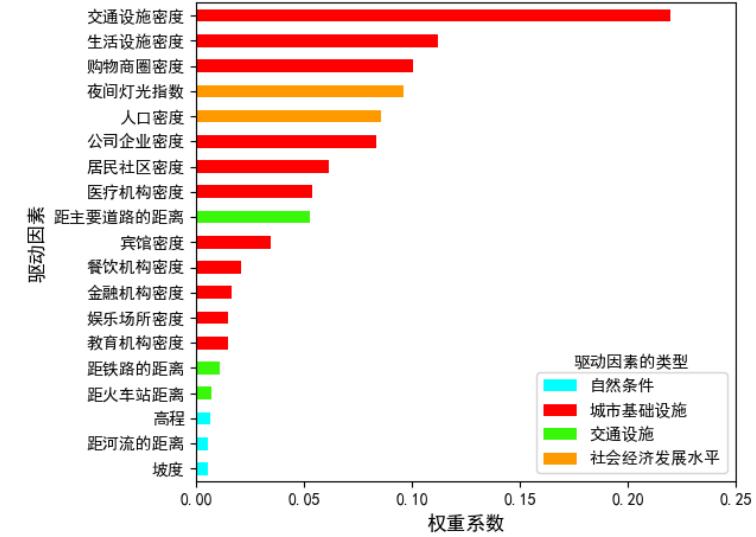
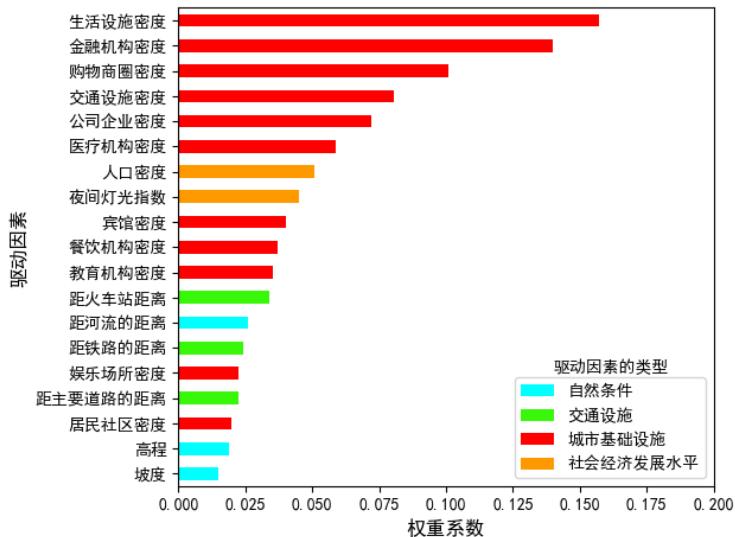
深入分析与人群活动有关的因素可以发现工作区域（47.00%）对能源消耗的影响要显著大于居住区域（21.07%）。

四大类驱动因素对江西省电力消耗的驱动重要性排行由大到小依次为：城市基础设施，交通设施，城市经济发展水平，自然条件。



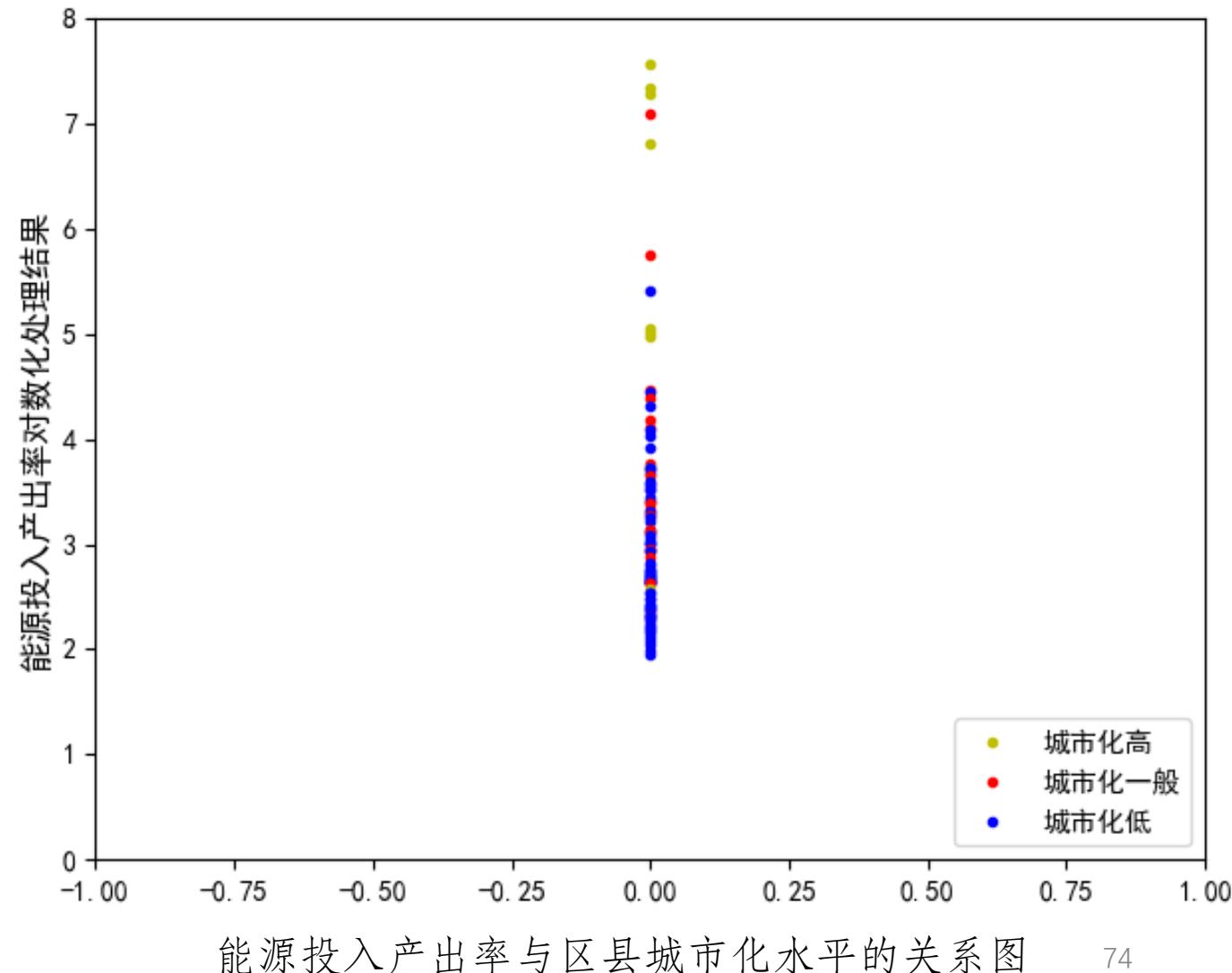
江西省电力消耗驱动因素权重统计图

## 典型地市电力消费驱动因素

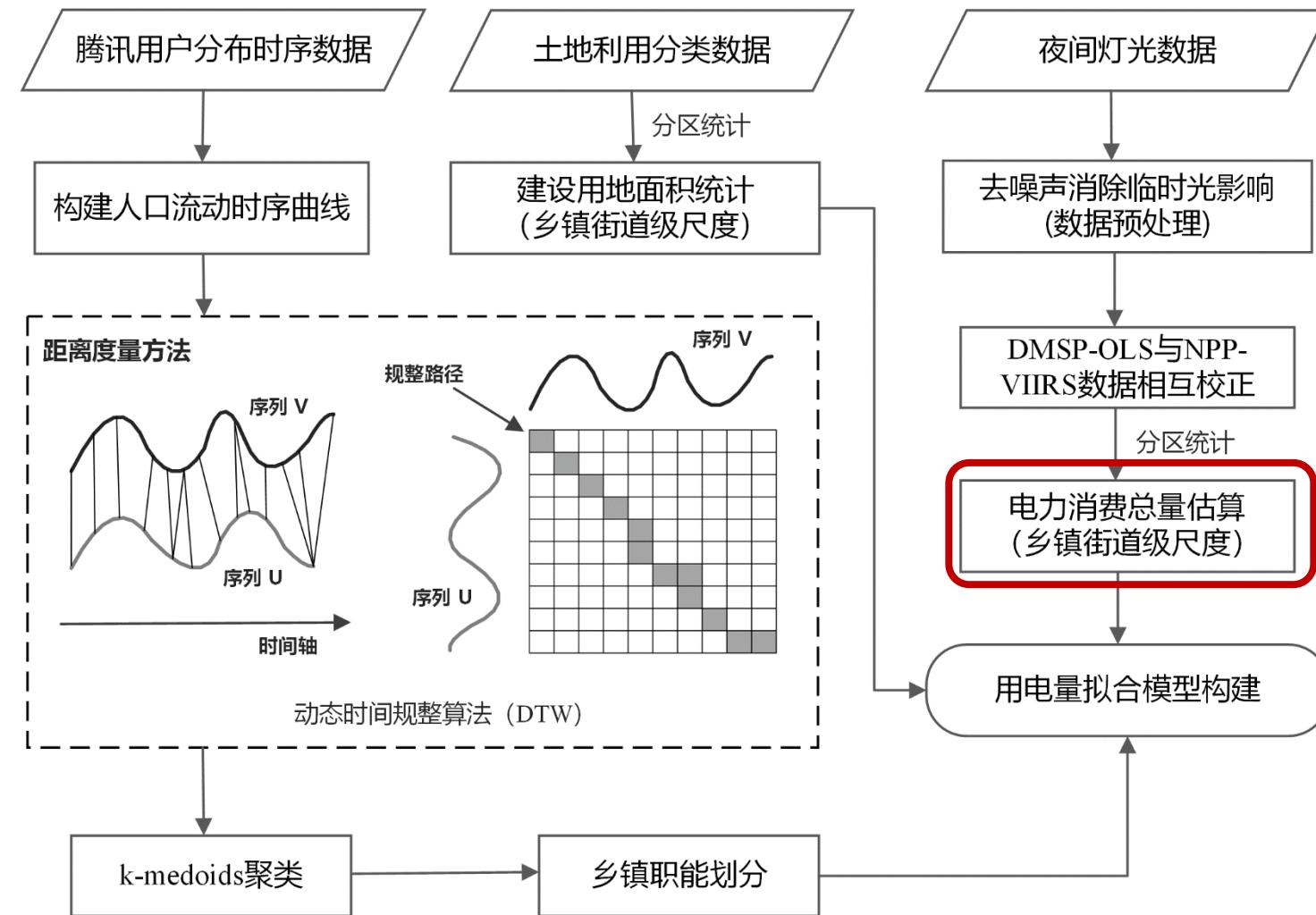


## ■ 电力消费与江西省区县发展

处于城市化不同阶段的城市，由于产业结构的差异导致能源的投入产出率具有较大的差异。城市化程度较高的地级市行政中心，其能源的投入产出率普遍较高。城市化程度较低的区县，能源投入产出率较低。



能源投入产出率与区县城市化水平的关系图

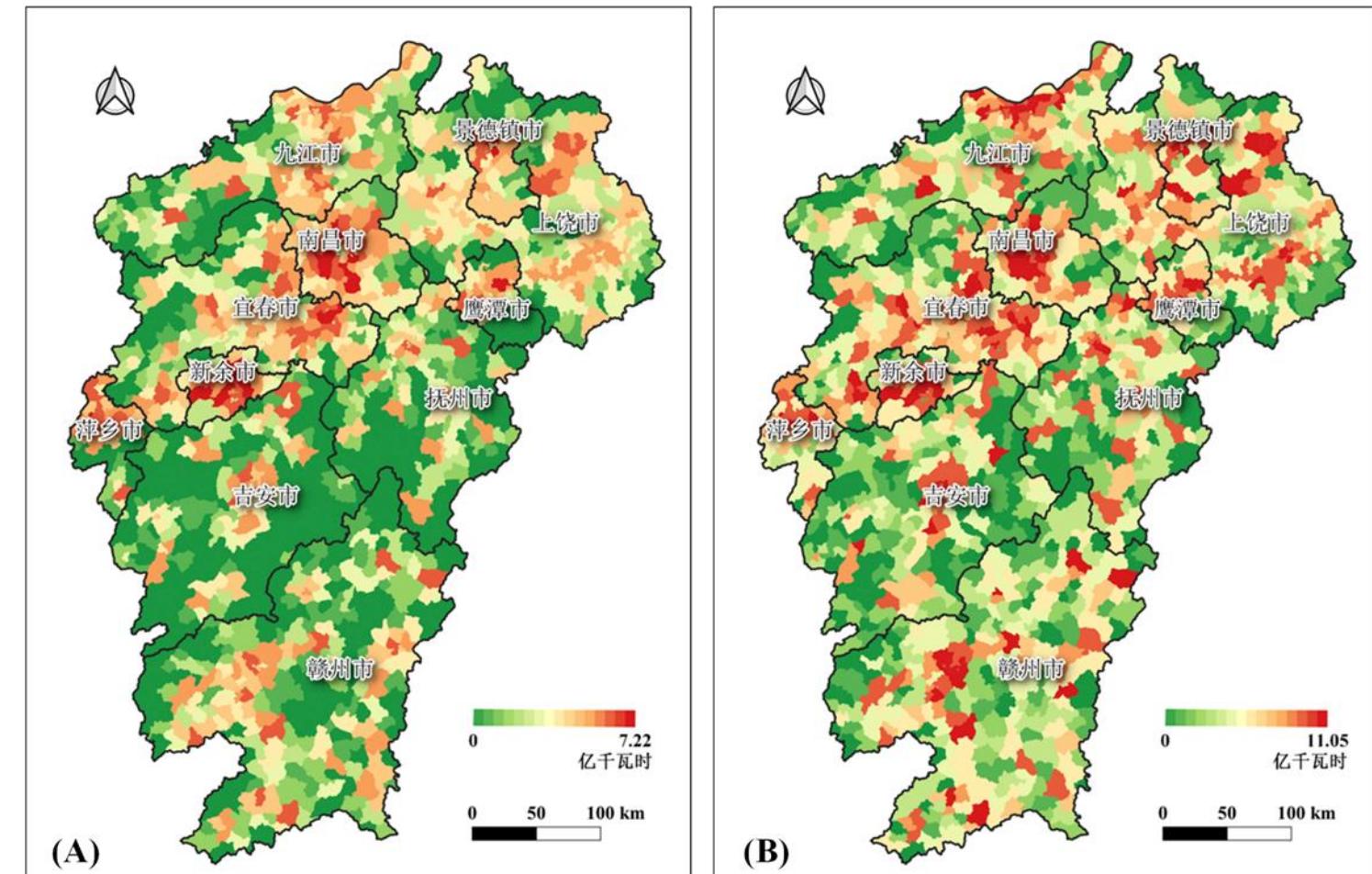


# 4.2 | 电力需求与城市变化的关联分析



## ■ 电力消费估算

- 根据2010年和2015年的夜间灯光数据，对两个年份的总用电量分配到乡镇街道级尺度。
- 2010年江西省总用电量为682.61亿千瓦时，2015年江西省总用电量为1,023.67亿千瓦时。
- 从图中可以看到2015年的用电量总体比2010年多，大部分乡镇的用电量都增多了。



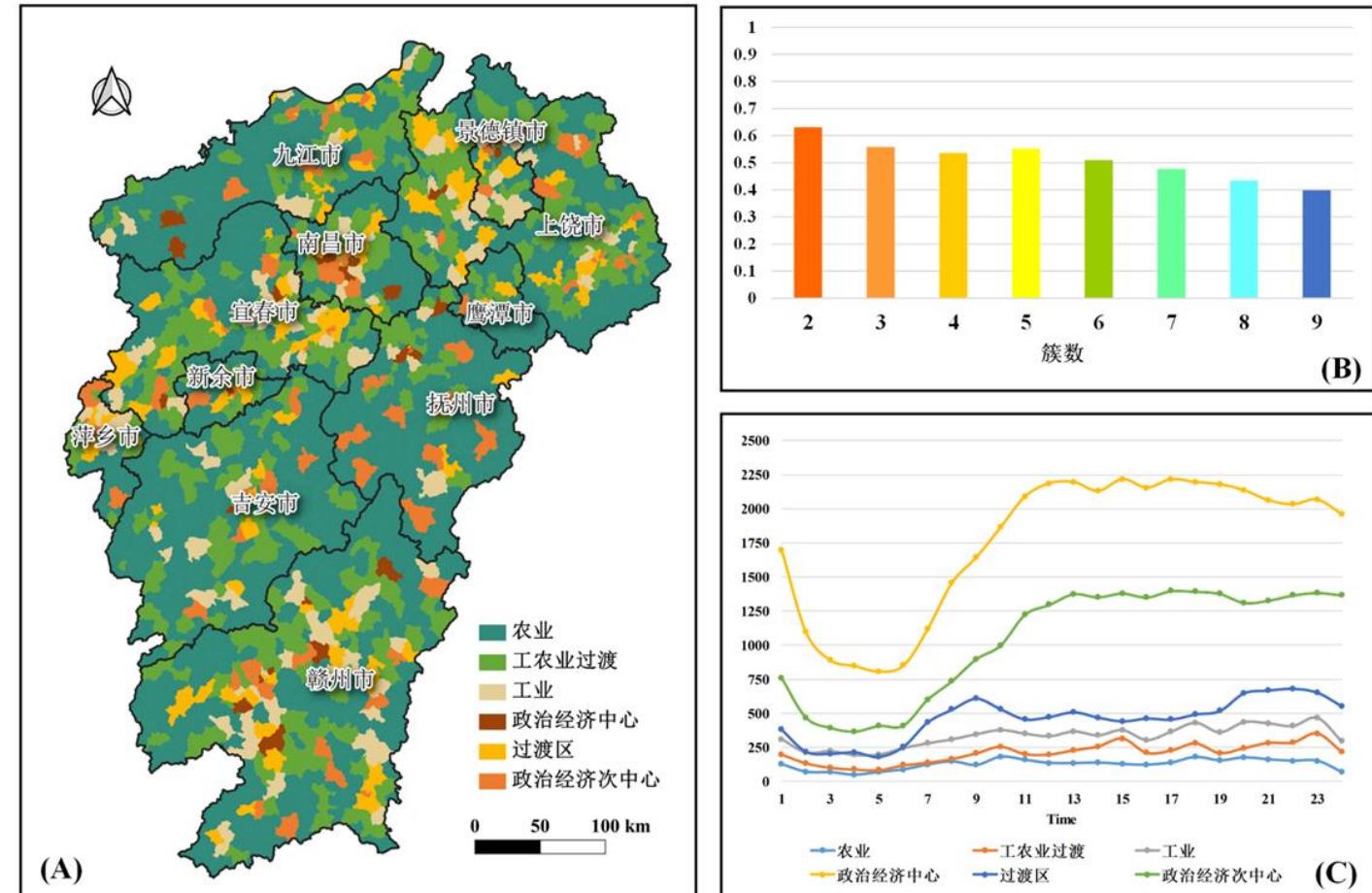
江西省用电情况分布图: (A) 2010年; (B) 2015年

## 4.2 | 电力需求与城市变化的关联分析



### ■ 区域职能聚类

- 利用了基于动态时间规整距离的 k-medoids 聚类算法对社交媒体用户分布时序数据进行聚类分析。
- 采用轮廓系数来评估不同k值的聚类效果。若轮廓系数大于0.5，则说明聚类效果良好。
- 根据该实验结果与江西省实际情况，将各乡镇街道级行政单元划分为6类职能区域。

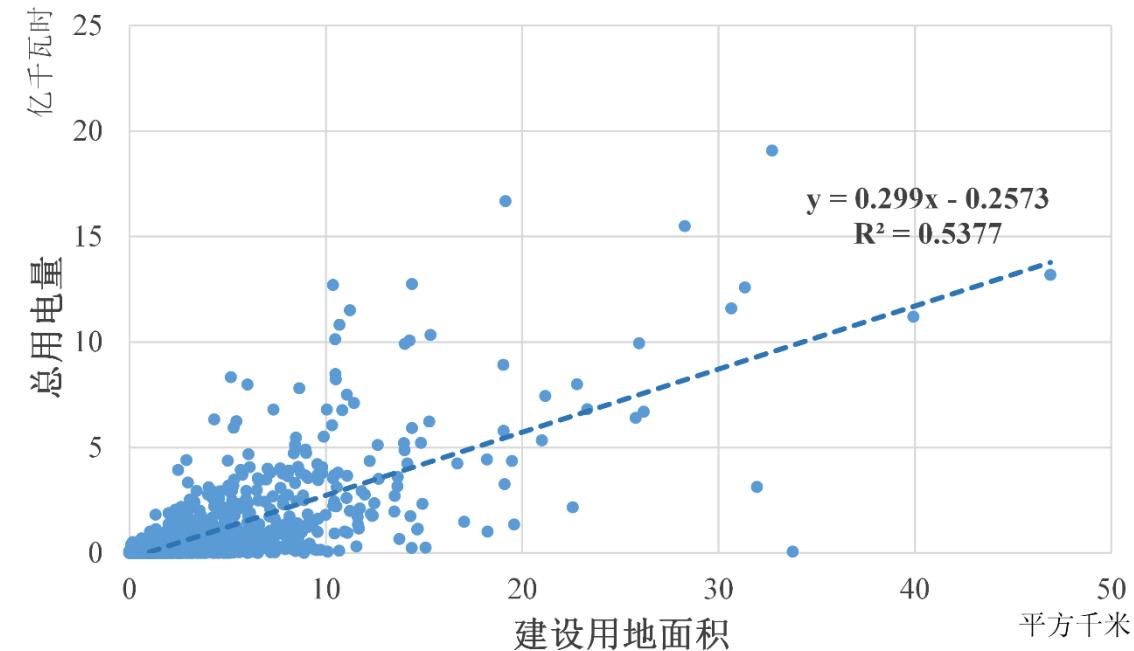
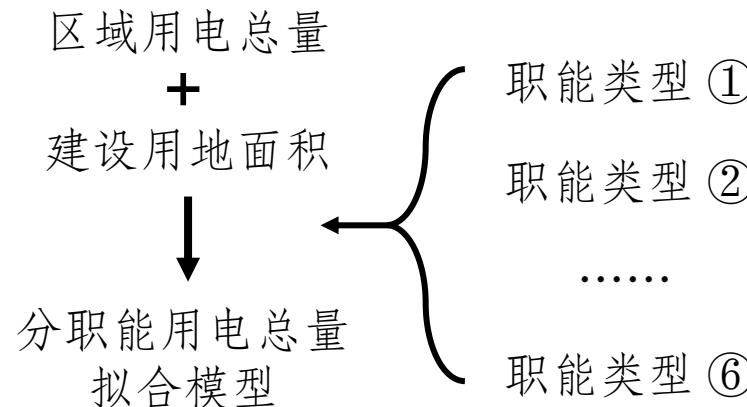


乡镇聚类结果：(A) 聚类结果的空间分布；(B) 聚类的轮廓系数；  
(C) 每个聚类中心的人口流动时序曲线

## ■ 分职能区域用电总量拟合模型

由于电力能源消费存在着显著的空间异质性，区域功能属性的差异在很大程度上会影响城市化与电力消费之间的关系。

因此本文将基于前面的区域聚类分析结果，通过分析不同职能下城市建设用地面积与用电总量的关系来构建用电拟合模型。

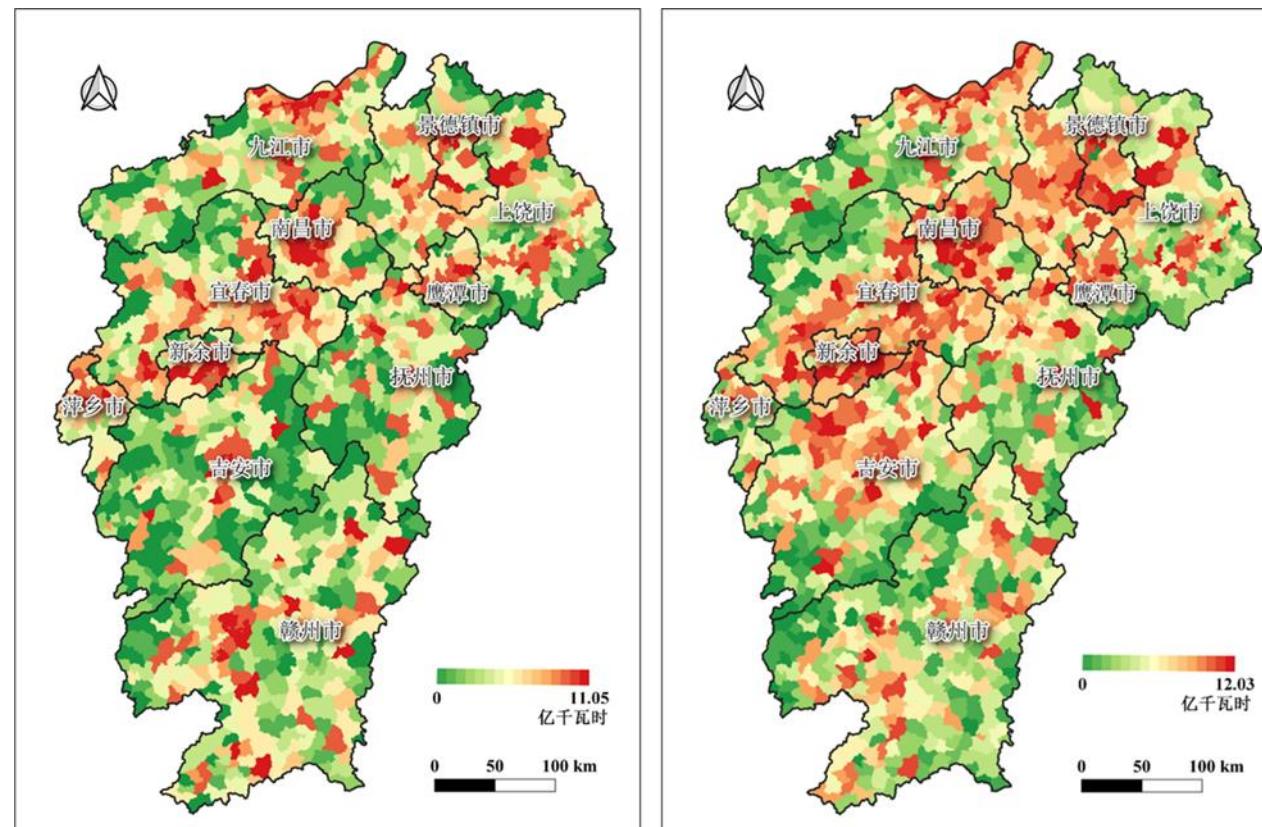


对2010年全省所有乡镇级行政单元的用电总量与城市用地面积进行回归分析。实验结果显示回归线的拟合优度  $R^2$  达到0.5以上，拟合精度较高，由此说明城市建设用地面积能够有效地反映区域用电情况。

# 4.2 | 电力需求与城市变化的关联分析



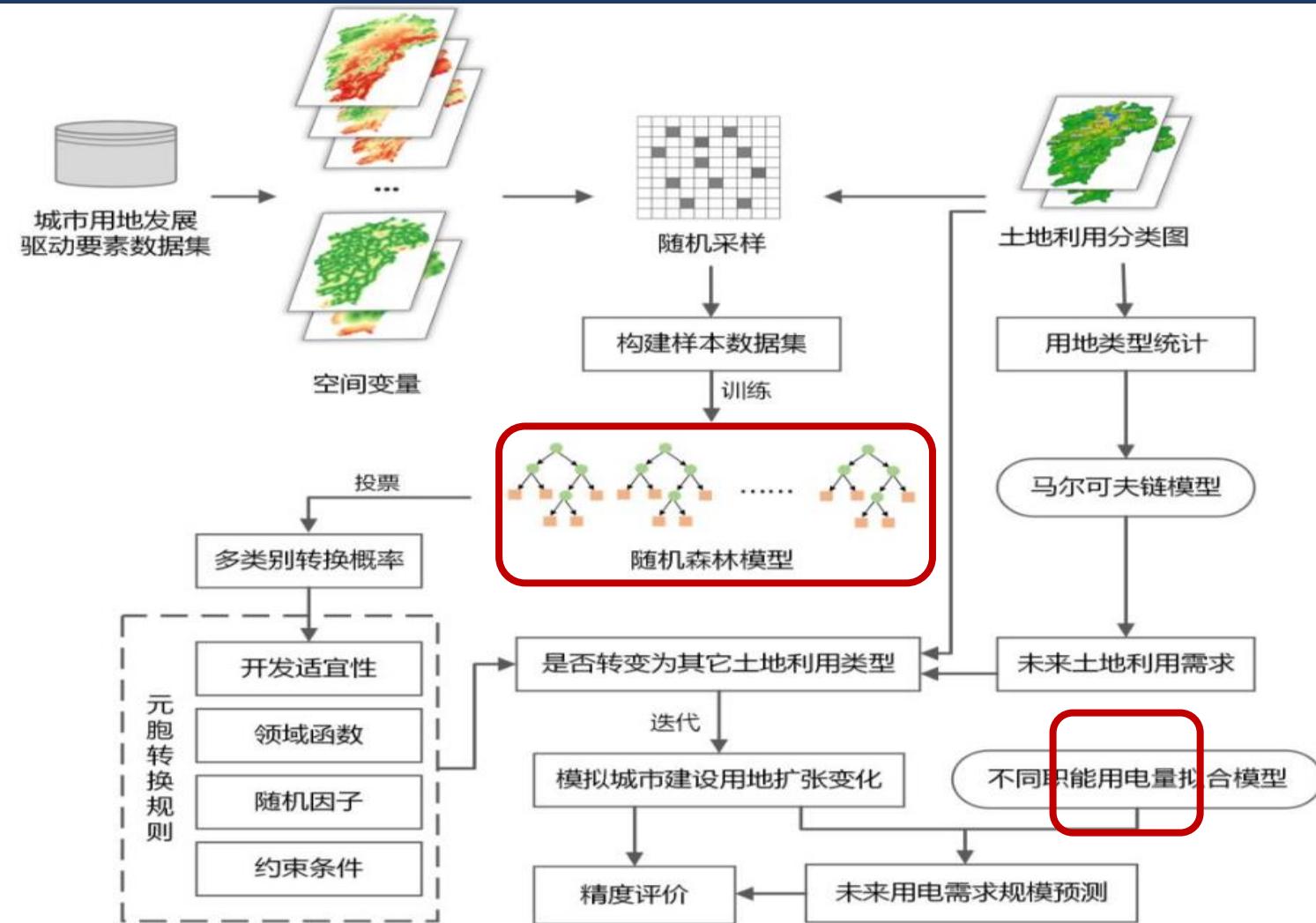
## ■ 全省用电量真实与拟合情况对比



江西省2015年用电量真实和拟合情况对比  
(A) 真实用电量分布; (B) 模型拟合结果

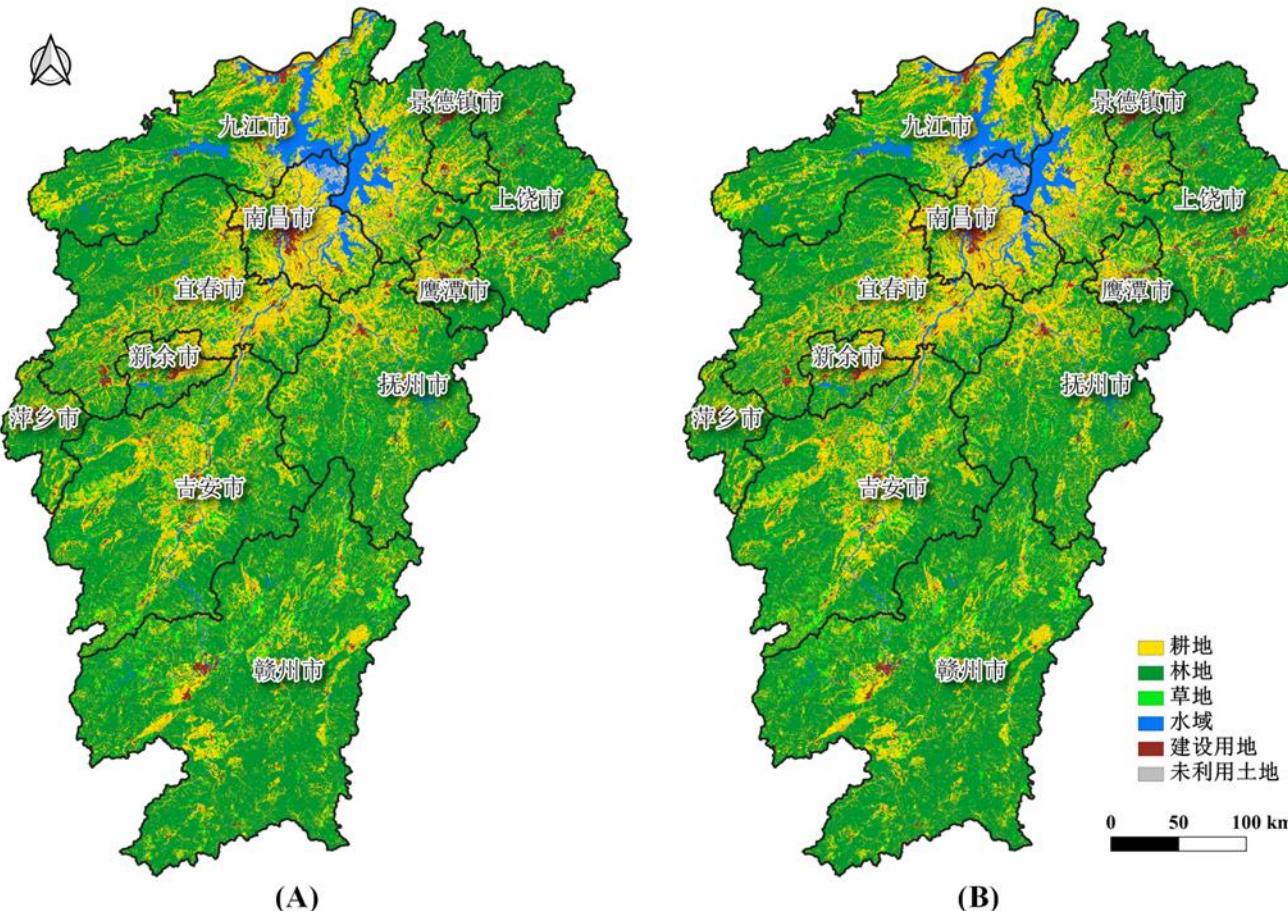
用电量真实和拟合统计结果及对比精度

区域	真实总用电量 (亿千瓦时)	拟合总用电量 (亿千瓦时)	相对精度
抚州市	56.01	70.40	74.30%
赣州市	138.30	104.66	75.67%
吉安市	66.60	108.14	37.63%
景德镇市	40.99	50.80	76.06%
九江市	139.83	112.54	80.48%
南昌市	163.21	156.05	95.61%
萍乡市	56.87	30.48	53.59%
上饶市	115.50	128.12	89.07%
新余市	67.20	60.77	90.43%
宜春市	144.78	164.77	86.19%
鹰潭市	34.38	36.79	93.00%
江西省	1023.67	1023.53	99.99% 79



未来用电需求规模增长模拟与预测

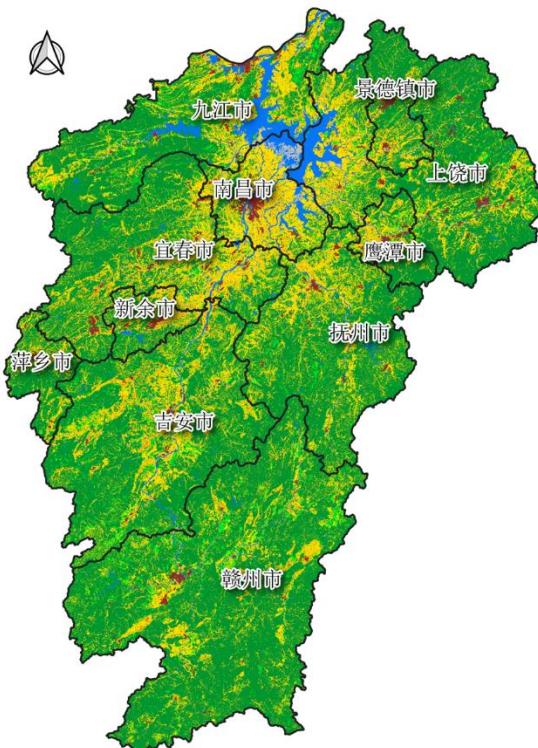
## ■ 城市化模拟结果分析



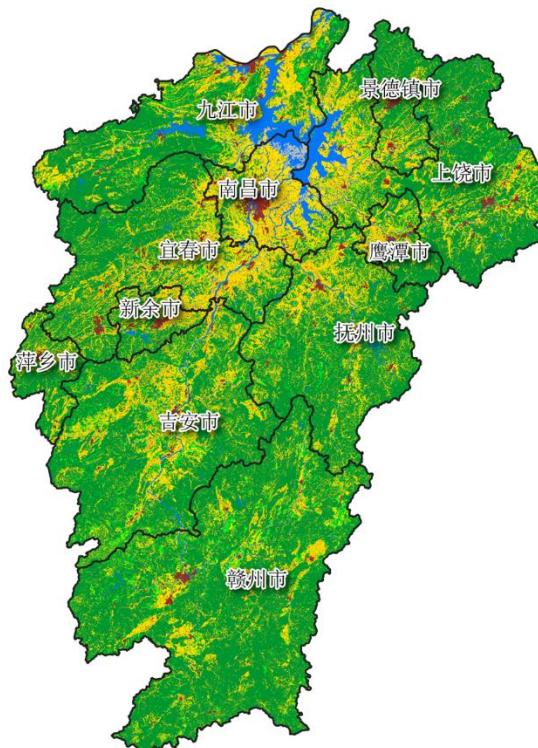
2015年真实与模拟的土地利用分布：(A) 真实情况；(B) 模拟情况

- 本研究通过整合RF-CA模型与MC模型，构建出土地利用变化模拟框架，有效地预测了未来城市发展变化情况。
- 左图展示的是2015年真实与模拟的土地利用情况。经过对比，模拟的总体精度为87.33%，Kappa系数为0.85。总体而言精度较好，图中对比两者的差异不大，说明模型是有效的，可以用来模拟未来的土地利用变化情况。

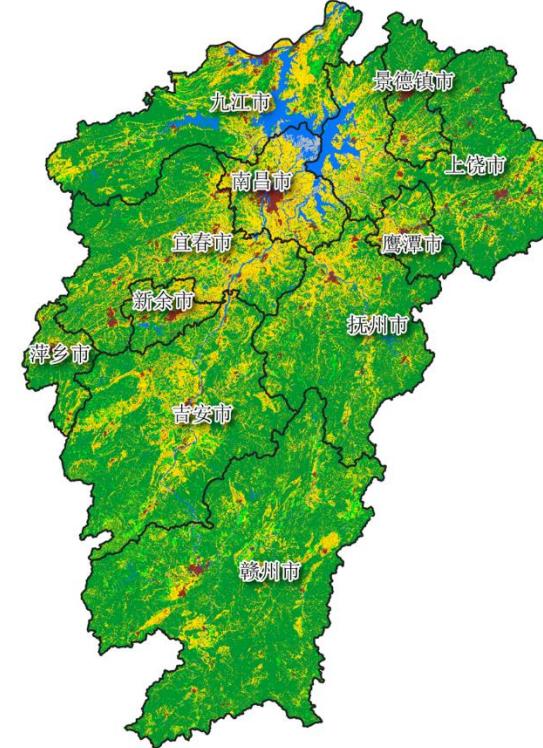
## ■ 城市化模拟结果分析



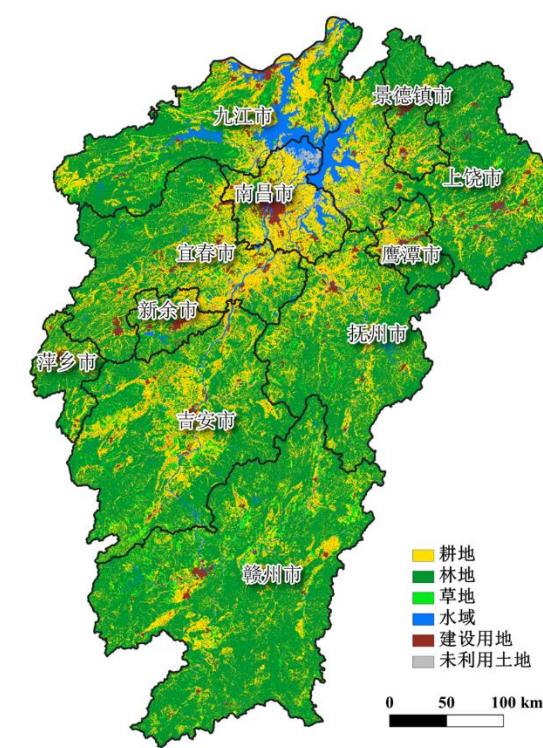
(A)



(B)



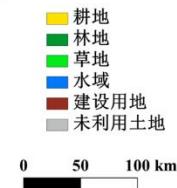
(C)



(D)

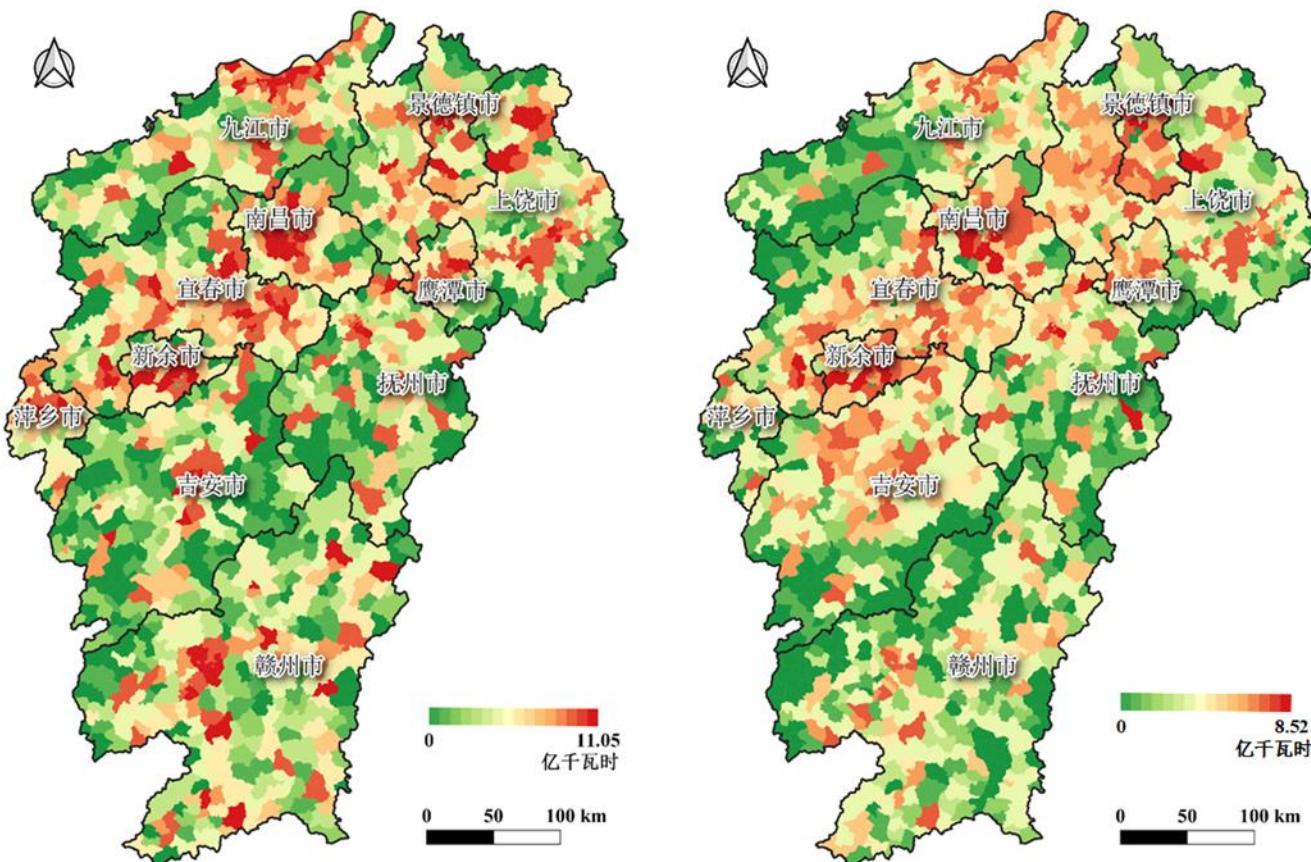
土地利用模拟情况: (A) 2020年模拟情况; (B) 2025年模拟情况; (C) 2030年模拟情况; (D) 2035年模拟情况

年份	2010	2015	2020	2025	2030	2035
建设用地面积 (平方千米)	3960.87	4793.99	5037.57	5280.35	5502.86	5721.71



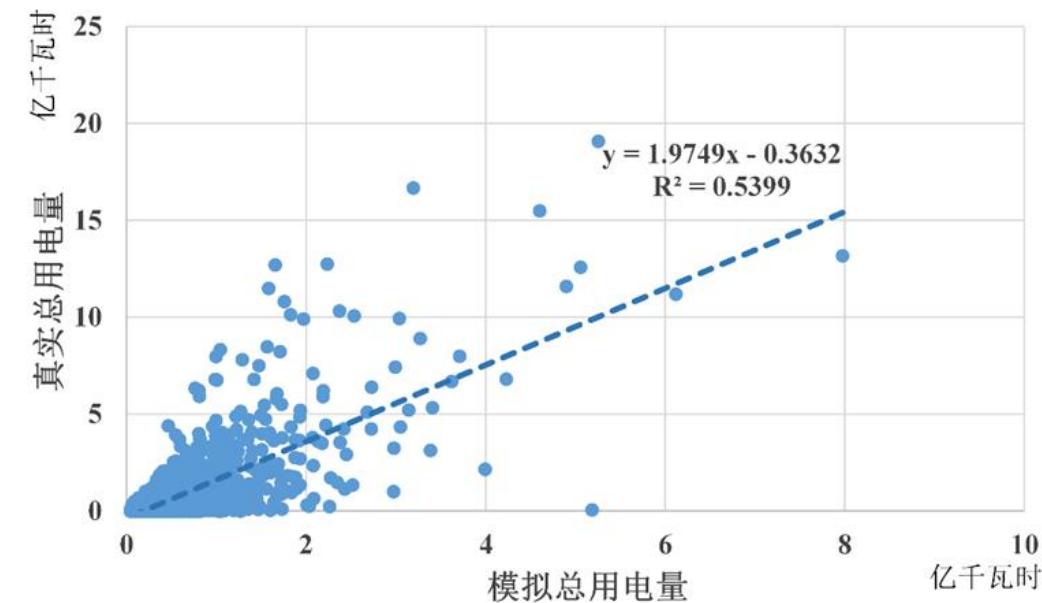
0 50 100 km

### ■2010-2015年用电消费变化模拟



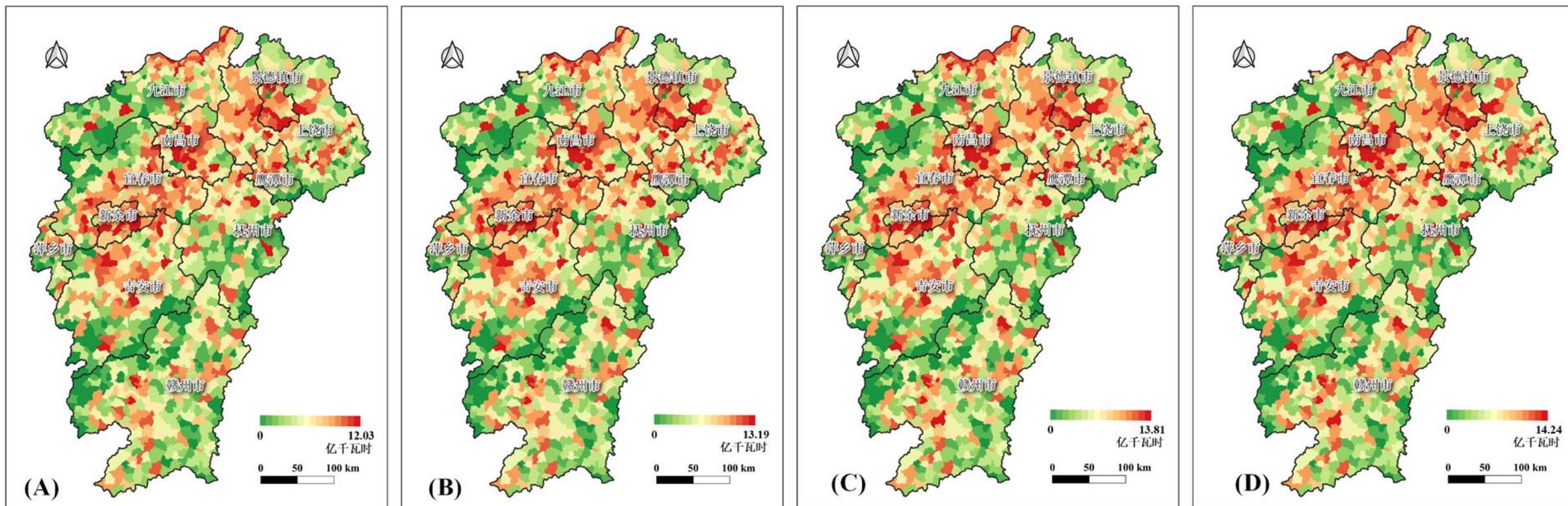
2015年模拟用电总量分布与真实分布情况：

(A) 真实分布； (B) 模拟分布



基于所构建的不同职能下用电总量与建设用地面积的拟合模型，结合模拟得到的2015年土地利用分布，最终得到2015年模拟的用电总量分布。从左图以及上图可以看出，在进行用电总量的模拟时，本文所提出的模型是较为有效的，拟合优度 $R^2$  达到了0.54。

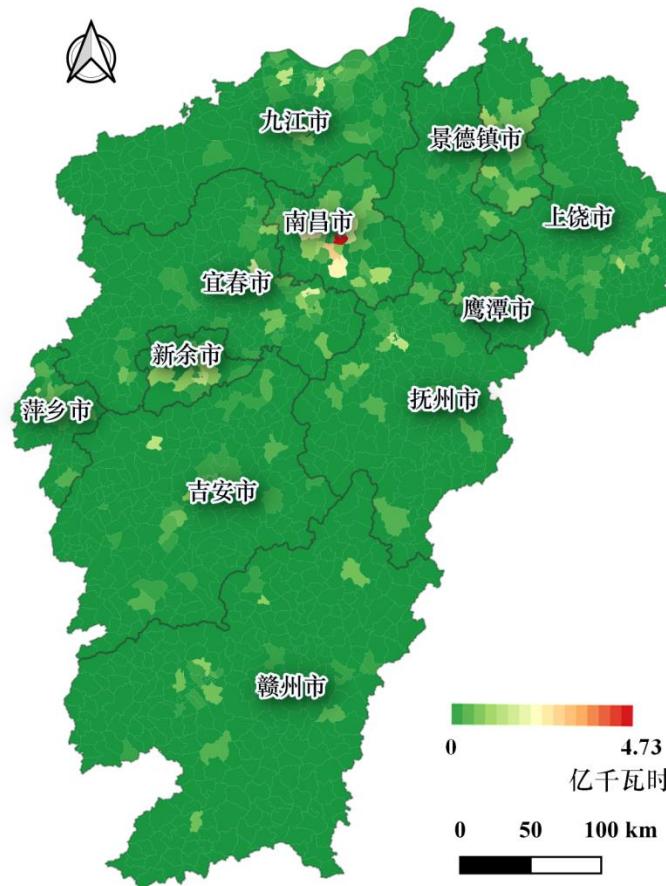
## ■ 电力需求分布中长期预测



未来用电总量预测：(A) 2020年；(B) 2025年；(C) 2030年；(D) 2035年

## ■ 电力需求分布中长期预测

未来用电情况预测统计结果



区域	2015年总用电量 (亿千瓦时)	2020年总用电量 (亿千瓦时)	2025年总用电量 (亿千瓦时)	2030年总用电量 (亿千瓦时)	2035年总用电量 (亿千瓦时)
抚州市	56.01	95.22	109.65	121.23	131.73
赣州市	138.30	147.29	168.59	185.94	206.82
吉安市	66.60	94.95	114.01	131.24	155.28
景德镇市	40.99	61.82	69.95	77.75	83.75
九江市	139.83	145.67	167.53	188.78	202.39
南昌市	163.21	179.38	205.00	233.71	291.51
萍乡市	56.87	87.91	94.64	100.71	117.19
上饶市	115.50	152.72	177.95	197.90	215.03
新余市	67.20	84.05	89.73	94.77	100.31
宜春市	144.78	188.43	214.17	235.26	252.93
鹰潭市	34.38	44.70	50.53	55.21	58.99
江西省	1023.67	1282.14	1461.76	1622.51	1815.94

# 本章总结



本章概述市政基础设施以及中国城市市政基础设施建设现状与展望，在此基础上，详述市政大数据的数据类型、来源、时空分辨率、优势、挑战及预处理手段等。

本章图解机器学习中经典的决策树、随机森林和旋转森林算法。

以水耗以及电力为例，市政大数据有助于理解城市空间结构、感知城市社会经济发展，包括但不限于：

- 感知混合城市土地利用模式
- 住宅空置的时空动态及驱动因子分析
- 用电时空分布及经济模式分析
- 电力需求与城市变化的关联分析
- 电力需求规模增长模拟和预测

市政大数据加深城市规划者、城市决策者对于城市系统的认知，为科学合理的城市规划、城市建设管理和安全保障等提供决策支持。



姚尧 博士, 副教授

地理与信息工程学院, 地图制图学与地理信息工程

阿里巴巴集团, 达摩院, 访问学者

Email: [yaoy@cug.edu.cn](mailto:yaoy@cug.edu.cn)

办公地点: 未来城校区地信楼522办公室

谢 谢!



High-performance Spatial Computational Intelligence Lab @ CUG