城市大数据的分析与统计

Analysis and statistics for urban big data

龙瀛，博士
清华大学建筑学院

2016年10月9日

Approaching the Human City: Beijing Studio
September 11 - 23, 2016
COURSE SYLLABUS

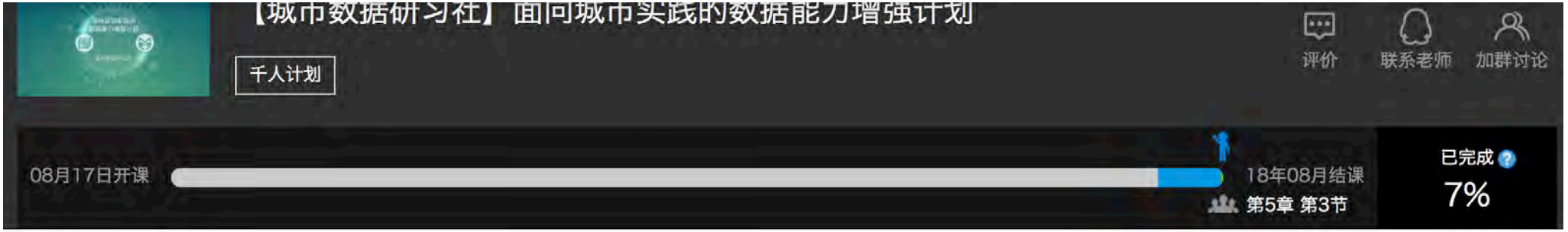
A joint workshop between the Human Cities Initiative at Stanford University
& Tsinghua University Academy of Art and Design and the School of Engineering

We invite you to participate in this experiment while we are holding class. Allow yourself to be fully present in the room, so you can listen to your classmates and what they have to say and share with you. Experience the freedom of not having to have your attention diverted or your mood instantly altered by whatever email or text message should come your way. You may take a phone call if it seems particularly urgent. But for most circumstances, we urge that you give yourself permission to be in control of your own time and energy— to actively choose where you want to direct your attention, as opposed to a portable device making that decision for you.

Please turn your mobile devices to "silent" or "do not disturb" mode, and do not take them out for the duration of the class. We strongly encourage you to take notes using pencil and paper— as research shows that this helps with memory retention— but If you must have a laptop to take notes, **please do not check your e-mail or browse the internet at any time.**

There will be plenty of opportunities to plug in once you leave the classroom. Let's treat our classroom as a sacred space to enjoy the moment.


腾讯课堂
首页
课程
精选合辑
学团
课程
搜索课程
机构管理
帮助
【城市数据研习社】面向城市实践的数据能力增强计划
千人计划
评价
联系老师
加群讨论
08月17日开课
18年08月结课
第5章 第3节
已完成
7%
付费课程均提供直播回放能力，即使你错过直播时间，也不用担心错过学习~
第1章 课前准备模块
第1节
课程总体介绍（购前必看，免费学习）
【录播】课程总体介绍 (26分钟)
第2节
入群方式及资料下载（课前必看）
【录播】入群方式及资料下载（课前必看） (8分钟)
第2章 规划师数据基础技能模块
第1节
城市POI数据爬取及相关知识
【录播】POI数据概念及爬取方法 (15分钟)



第3节 OSM开源地图数据爬取

【录播】OSM开源地图数据爬取 (24分钟)

第4节 Excel数据分析基础技能

【录播】Excel数据分析基础技能 (23分钟)

第5节 BDP在线数据可视化

【录播】BDP在线数据可视化 (30分钟)

第6节 百度图说图表可视化

【录播】百度图说图表可视化 (19分钟)

第7节 Powermap数据可视化专题

【录播】坐标数据的Powermap制图 (19分钟)

【录播】Powermap时间轴设置 —— 动态图表 (15分钟)

《大数据与城市规划》教学大纲

1. 大数据与城市规划概论（整合进9月23日）
2. 大数据在城市规划中应用的研究进展（9月23日）
3. 城市大数据的获取（9月30日）
4. 城市大数据的分析与统计（10月9日）
5. 城市大数据的可视化（10月14日，ArcGIS可视化、三维、在线可视化平台）
6. 城市大数据挖掘：空间句法（10月21日）
7. 城市大数据挖掘：城市网络分析（10月28日）
8. 学生作业中期汇报与点评（11月4日）

9. 大数据与城市规划的结合 （11月11日）
10. 数据增强设计（11月18日）
11. 战略及总体规划中的大数据应用（11月25日）
12. 控制性详细规划中的大数据应用（12月2日）
13. 城市设计中的大数据应用（12月9日）
14. 参与式规划中的大数据应用（12月16日）
15. 大模型：新数据环境下的城市研究新方法（12月23日）
16. 学生作业终期汇报与点评（12月30日）

上一堂课的回顾

- 结构化网页的抓取
- 利用API抓取开放数据
- 共享的第一个版本的北京旧城数据
- 课外阅读

运用互联网数据进行城市规划与研究所使用的主要工具及流程

- **STEP1：定位查找数据源的网络地址**

Chrome（谷歌浏览器）

- **STEP2：将获取的网络开放数据保存在本地**

火车采集器（V8）

- **STEP3：数据的清洗、预处理**

Microsoft Excel 2007

- **STEP4：坐标系统转换（地理坐标→平面坐标）**

万能坐标转换器

- **STEP5：空间分析、空间统计及可视化表达**

IBM SPSS 19.0

ArcGIS 10.x软件包

来源：郑晓伟

分组与大作业

- List：人名/学号、作业题目、组长及其联系方式
- 没有分组的明天晚上之前由课代表协助完成，建议课后留下来自我组织
- 大作业的主题/题目可以两周内上报

本讲大纲

1. 数据分析与统计概述
2. 基于ArcGIS的数据分析
3. 基于SPSS的数据统计
4. 案例介绍：基于街景图片的街道绿化研究

一、数据分析与统计概述

本节部分幻灯片来自西安建筑科技大学的郑晓伟老师，在此表示感谢

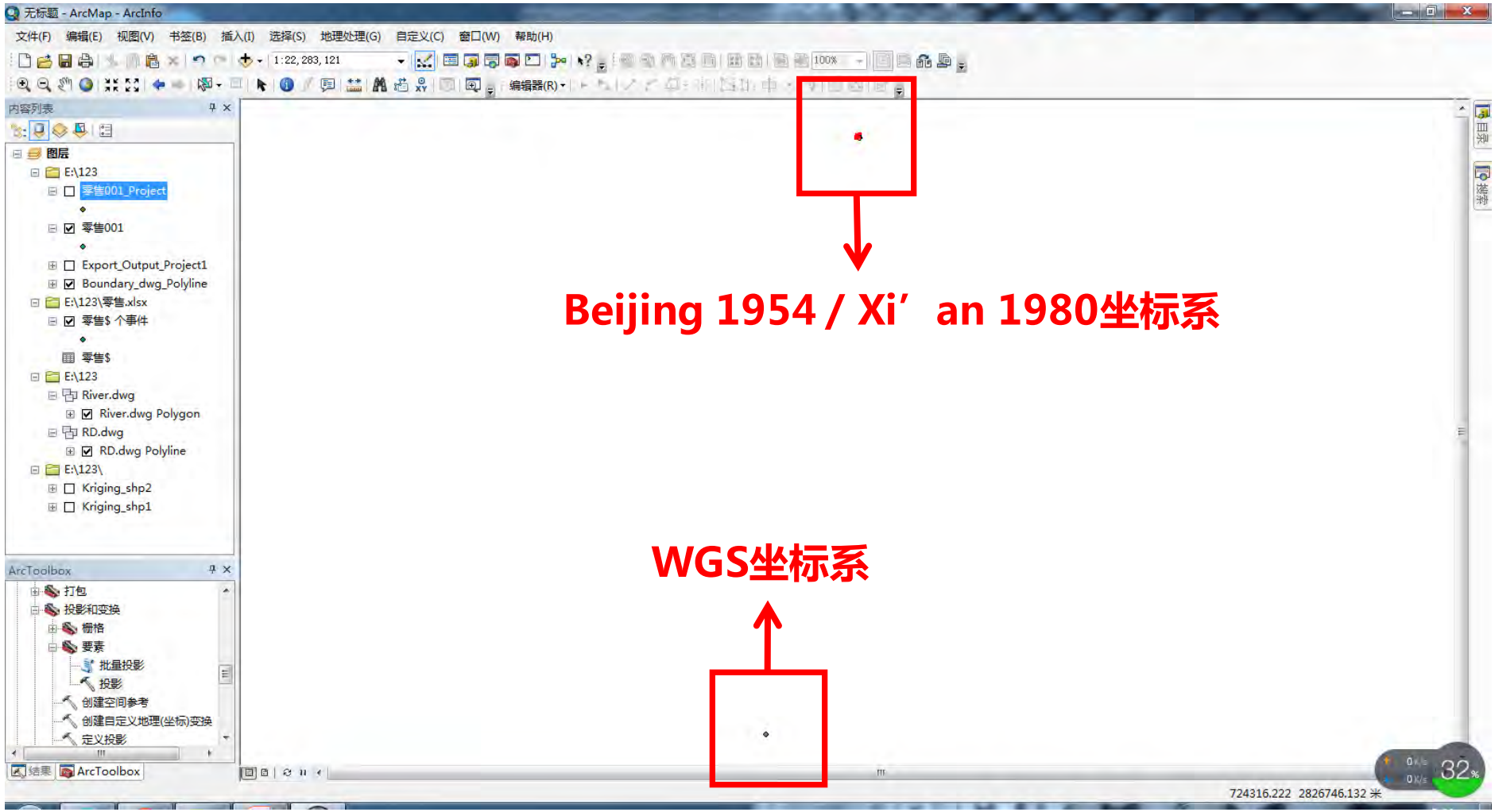
为什么要做数据分析与统计

- 数据→信息→知识→智慧

数据预处理

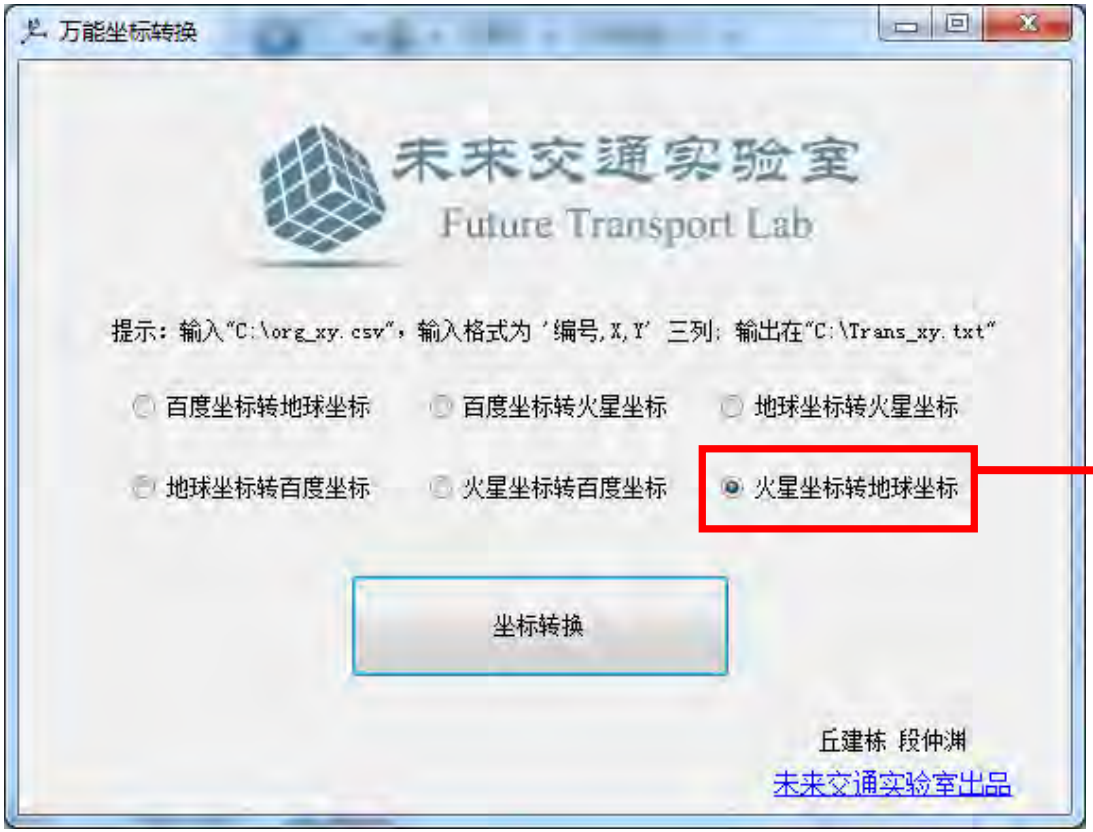
- Excel、Access、ArcMap、SPSS无论哪个软件，都需要进行数据预处理
- 数据预处理做什么（个人经验、先后顺序）？
  1. 是否有明显的数据缺失（是否完整）
  2. 冗余字段的删除
  3. 保留字段改名称
  4. 增加唯一ID字段并计算其数值（可追溯）
  5. 各个字段中的异常或者空值的处理
  6. 空间对象是否重叠/重复（如天安门的例子）
- 需要有强迫症的思维

位置纠偏



- 来自新浪、微博和腾讯等国内网站的地理位置多是火星坐标系
- 坐标系统的不同，会导致空间位置无法完全匹配，需要纠正到地球坐标
  - 保证一个项目的所有图层的坐标处于同一坐标体系（地球与火星）

位置纠偏



	A	B	C
1	1	108.9476	34.24587
2	2	108.9476	34.25782
3	3	108.9485	34.24925
4	4	108.938	34.25243
5	5	108.934	34.25051
6	6	108.9232	34.23627
7	7	108.9473	34.25789
8	8	108.9486	34.25013
9	9	108.9551	34.25435
10	10	108.9489	34.25114
11	11	108.9392	34.2558

Trans_xy - 记事本

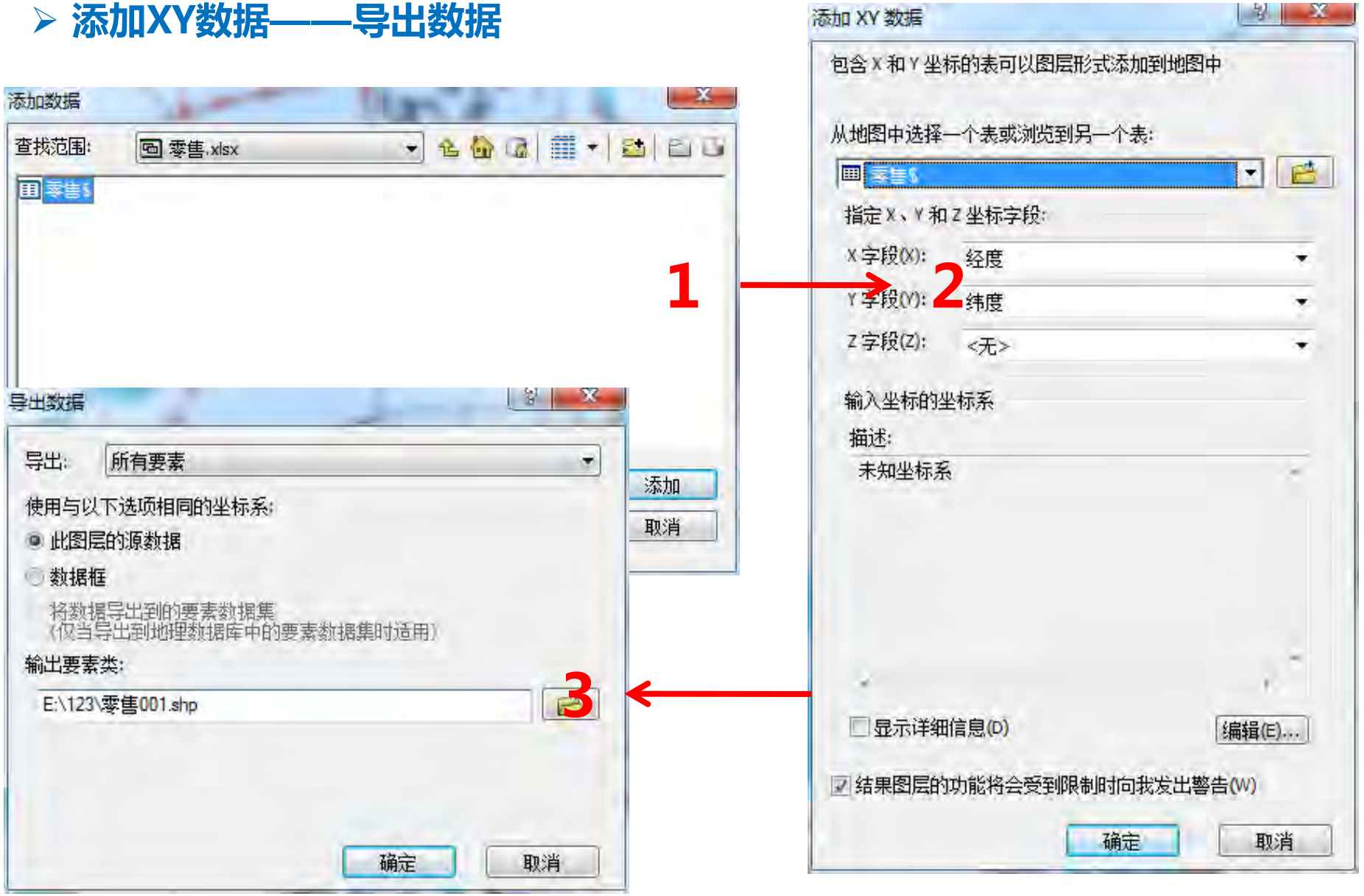
文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)

1,108.942977537453,34.2474656784029
2,108.942982355667,34.2594063777044
3,108.943816694545,34.2508402191483
4,108.933362489567,34.2540299722615
5,108.929354541146,34.252120020051
6,108.918577355224,34.2378962093459
7,108.942653270121,34.2594769410353
8,108.943971782149,34.2517137948719
9,108.950401159916,34.2559277029528
10,108.944228366813,34.2527287260284
11,108.934582343306,34.2573997981637

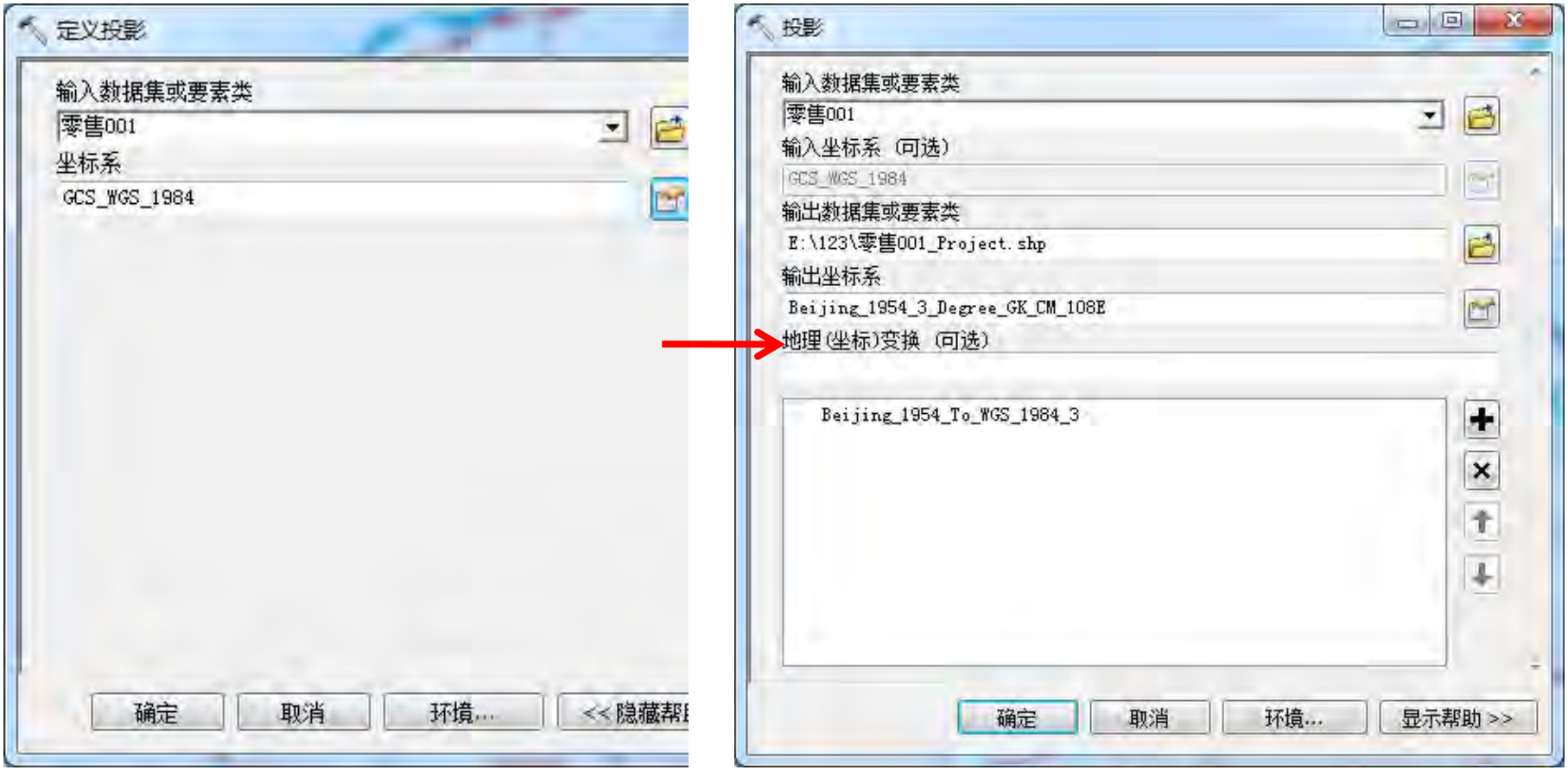
- 推荐利用未来交通实验室的万能坐标转换工具进行纠偏
  - 百度等也提供了纠偏/加偏的API
- Win7以上系统以管理员模式打开
- .net Framework版本需在4.0以上

数据空间化（GeoCoding）

➢ 添加XY数据——导出数据

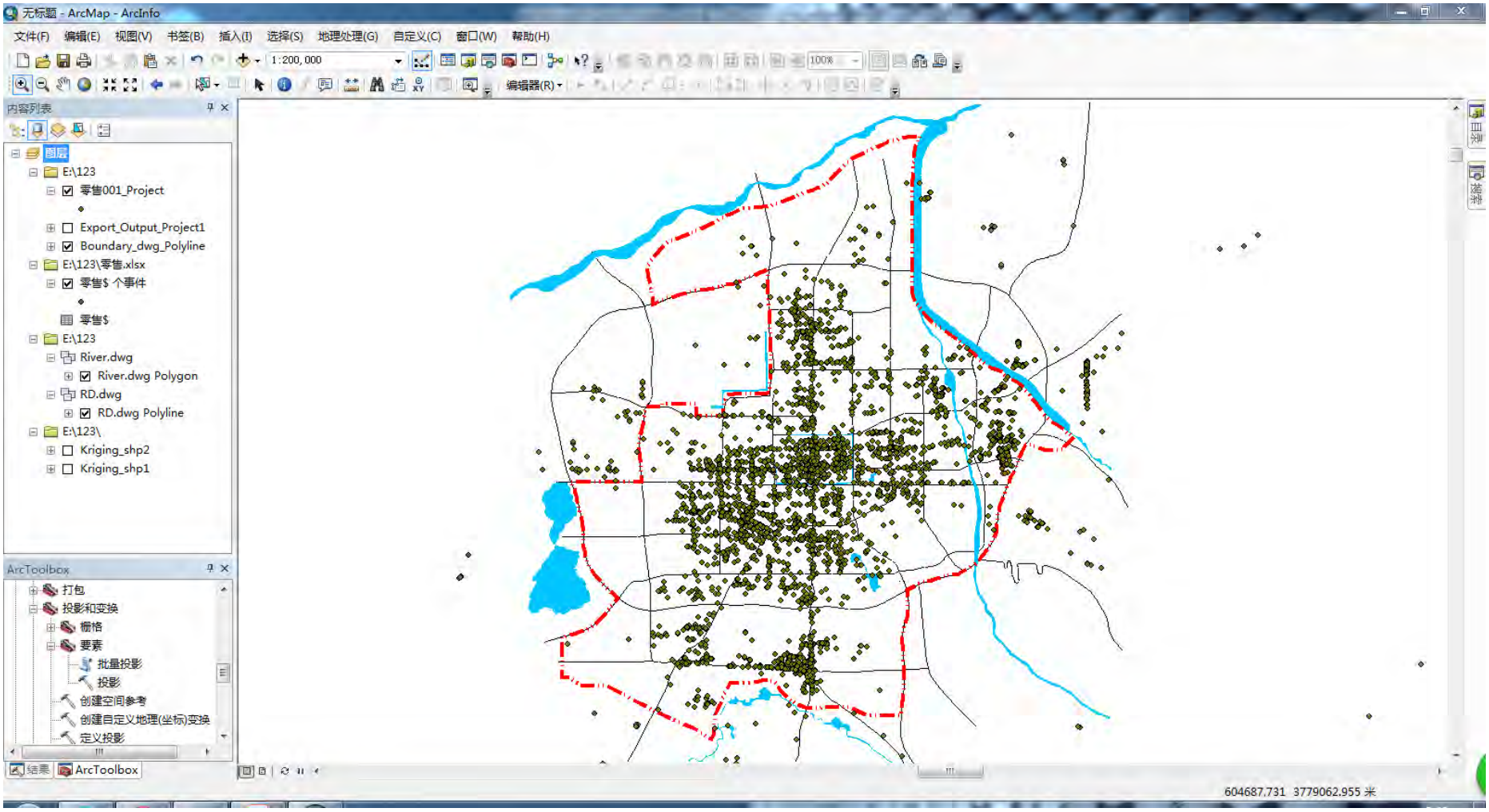


数据空间化（GeoCoding）



- 定义投影系统

数据空间化（GeoCoding）



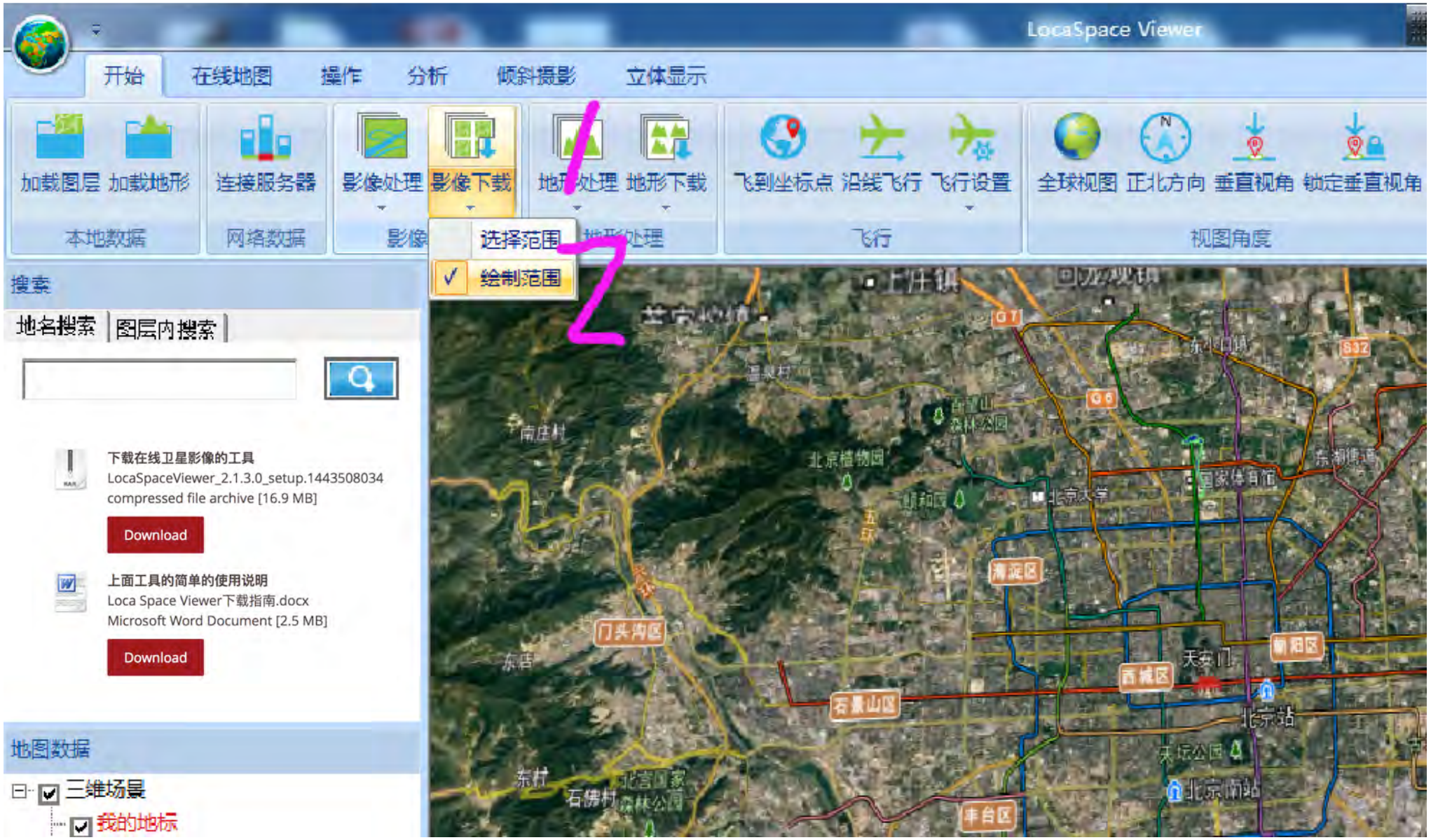
- 空间化后的数据一览
- 点状地物可以用于后续的不同空间单元的分析与统计

常用的数据分析与统计的软件

- 办公自动化软件
  - Access、Excel
- 数据库平台
  - SQL Server、ORACLE
- 地理信息系统软件
  - ESRI ArcGIS、GeoDA
- 统计分析工具
  - SPSS、STATA
- 大数据软件
  - 火车采集器（也能进行数据预处理！）、Tableau

- 多数工作都能在Excel中完成！

自动下载影像地图的工具 Loca Space Viewer



- **http://www.beijingcitylab.com/projects-1/22-urban-design-course/**

二、基于ArcGIS的数据分析

GIS vs CAD

GIS vs Big Data

点、线与面

ArcCatalog

- 建立GeoDatabase（GDB）
  - 如果工程不大，建议Personal GDB，物理格式为mdb，与微软Office的Access通用，便于属性数据预处理、分析与统计
  - 栅格图层、Toolbox、网络分析和拓扑检查等也可以存入并在GDB中操作
- 数据管理（类似Windows的资源管理器或Mac的Finder）
  - 导入图层、过程图层存储、删除图层
- 调用ArcToolbox
  - 便于操作图层

ArcMap

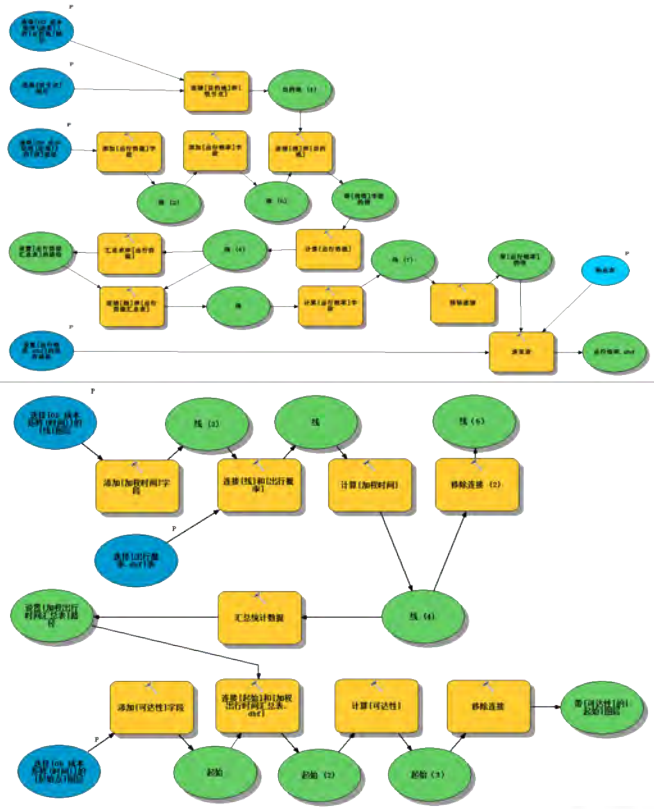
- 显示与管理
  - 加载、增加/删除字段、选择一部分对象
- 计算
  - 计算字段（属性与空间）、关联join、空间关联spatial join
- 分析与统计
  - 一般需要调用Toolbox
  - 字段的summarize功能
- 可视化
  - 符号化、layout视图

ArcToolbox

- 个人常用的工具箱展示
  - 空间分析（buffer、overlay、union）
  - 空间统计（全局自相关Moran's I、局部子相关LISA、最小二乘回归OLS、地理加权回归GWR）
  - 空间数据处理（repair）
- Model Builder

Aggregate Polygons
Append
Batch Project
Buffer
Clip
Create Personal GDB
Define Projection
Dissolve
Eliminate
Eliminate Polygon Part
Erase
Euclidean Distance
Extend Line
Feature To Line
Feature To Point
Feature To Polygon
Feature to Raster
Focal Statistics
Generalize
Generate Near Table
Geographically Weighted Regression

Geographically Weighted Regression
Identity
Intersect
Kernel Density
Multipart To Singlepart
Ordinary Least Squares
Point Statistics
Project
Raster to Polygon
Reclassify
Repair Geometry
Sample
Simplify Line
Sort
Spatial Join
Summary Statistics
Trim Line



ArcGIS Scripting using Python

- 每个Toolbox都可以利用脚本实现（具体参见每个工具箱的help）
- 推荐利用ArcGIS安装的默认Python版本和编译器（2.7.x而不是3.x）
- 课外参考资料待放到网站
  - ExtendingArcGISWithPython
  - Programming ArcGIS 10.1 with Python Cookbook

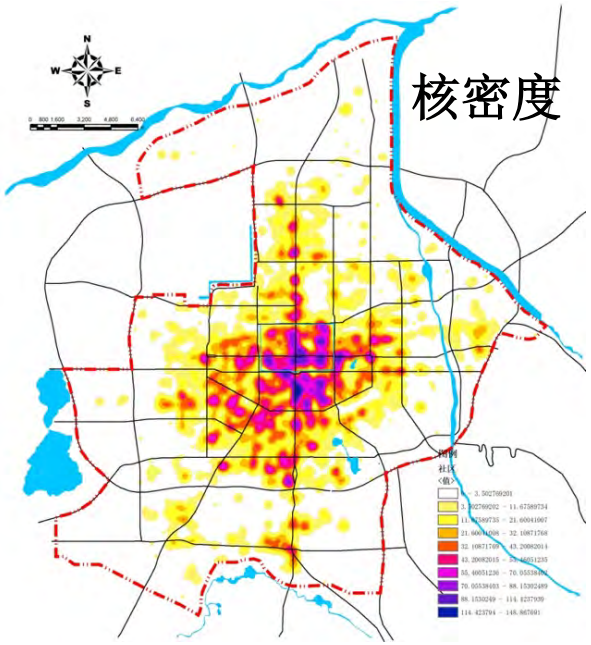
- 会了Python与不会所做的工作的巨大差异（个人经历的分享）

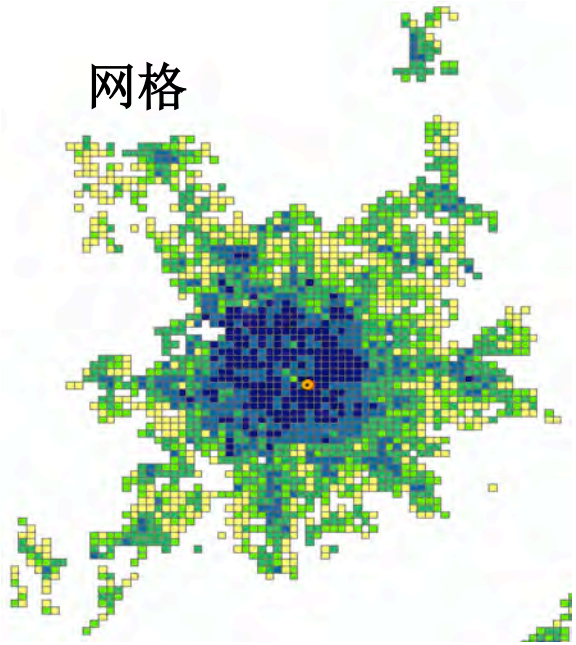
- 具体演示：

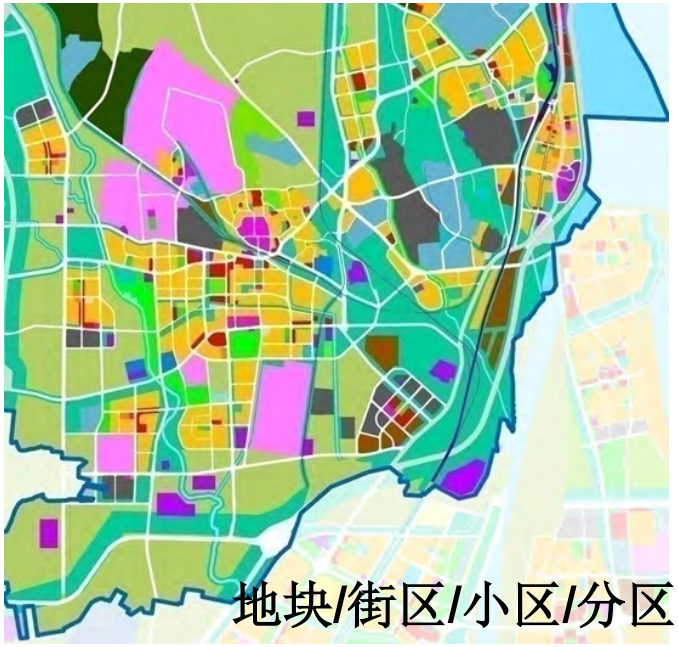
ArcGIS的几个常用技巧

- 推荐使用ArcGIS 10.x的英文版本
- 选择合适的分析单元，生成图层作为日后常用
- GeoDatabase（中的OBJECTID/OID会随着操作而变化，建议单独建立一个字段如BLOCK_ID/STREET_ID表示唯一的空间对象ID
- 能用属性来表示，就不用额外生成新的图层
- 尽可能地利用GeoDatabase（gdb）来管理空间数据，而不是ShapeFiles（如果工程不大，建议mdb）
- 不要随意删除mdb中空间图层的属性对象（行）
- Toolbox中的Repair Geometry是个好工具
- 数据库释放空间的方法（Compact database，在数据库上右键）

分析单元的选择（！！！我们看待城市的视角）







不同分析单元的数据分析

- 核密度：
  - kernel density工具箱
  - point density是另一种选择
- 网格：
  - Create fishnet生成不同尺度的网格
  - ArcMap中的spatial join
- 地块：
  - 利用道路网生成地块（单线与双线），详见所提供的Liu and Long 2016 EPB参考资料（利用兴趣点和道路数据推导地块主导功能、功能密度和功能混合度）
    - 所提供的Parcels2011AICP为2011年利用路网生成的扣除了道路空间的地块/街区
  - ArcMap中的spatial join
- 街道：
  - 路网数据经过必要的预处理（多线变单线、拓扑处理、细枝末节道路的删除等）
    - 所提供的Road_all_attributes经过了数据预处理
  - ArcMap中的spatial join

三、基于SPSS的数据统计

SPSS的很多功能可以用Excel实现

常用功能

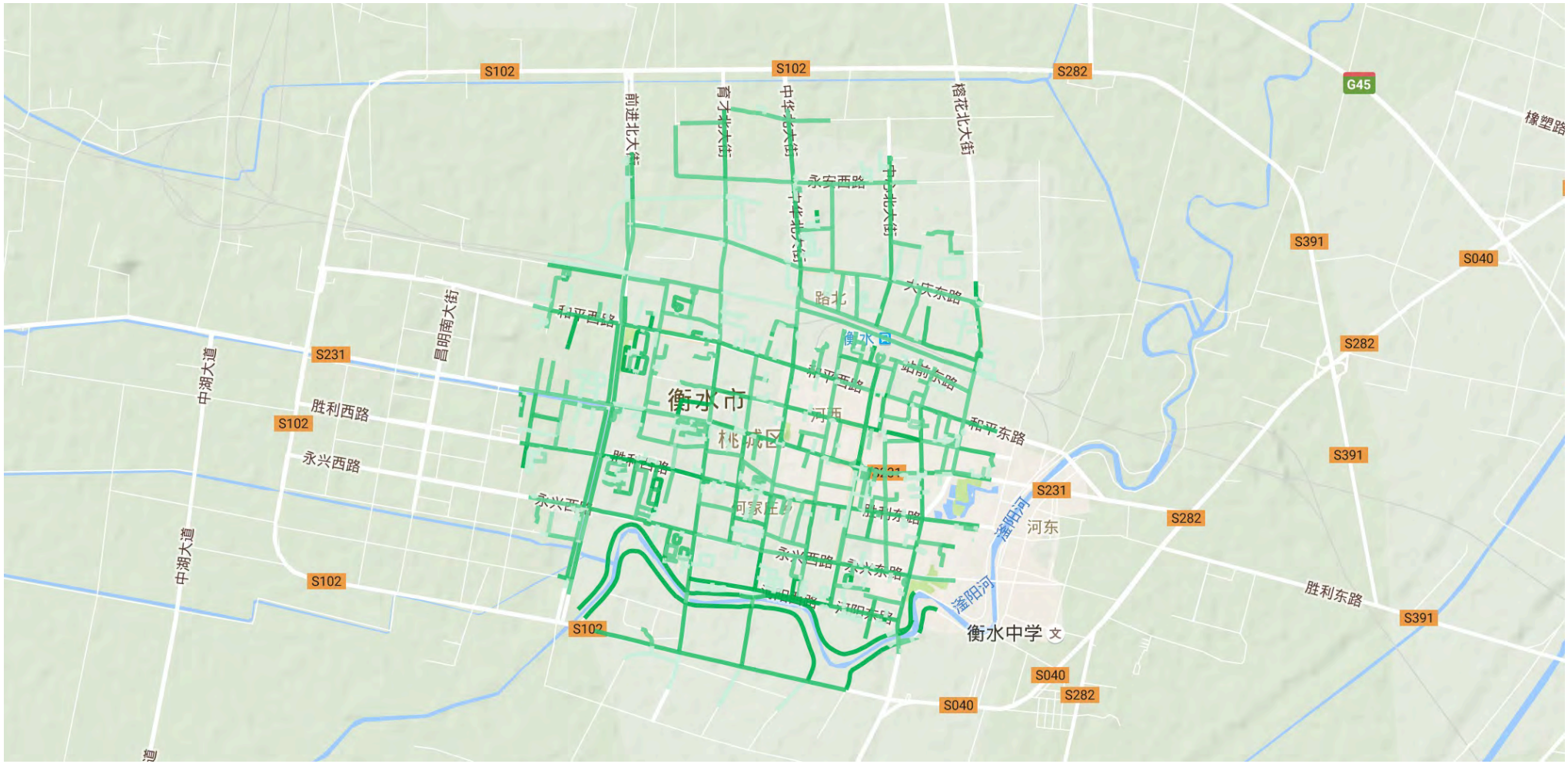
- 数据处理
  - 计算字段
- 统计性描述
- 相关分析
  - 0.8
- 回归分析
  - 回归前的自相关（autocorrelation）检查（VIF）
  - 二元/多元回归、线性/非线性回归
    - 部分数据的回归分析
    - 对数ln（如房价）
- 聚类分析
  - K-means

- 软件展示

四、案例介绍

基于街景图片的街道绿化研究

成果展示：GeoHey平台



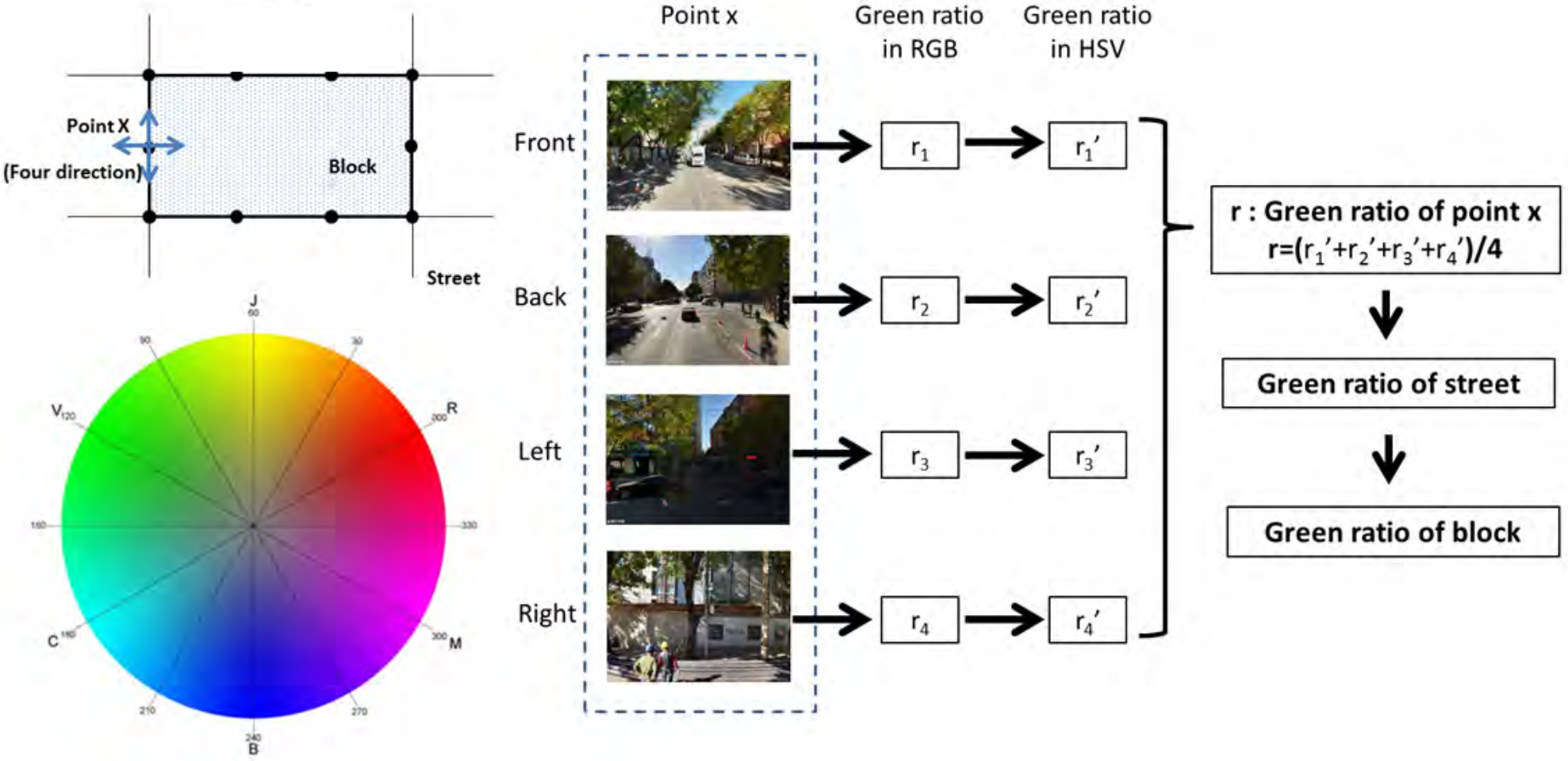
- https://geohey.com/apps/dataviz/357b07615c4b4e25b76dcdd1ca9cd8f2/share?ak=ZmYzNmY0ZWJhYjcwNGU2ZGExNDgxMWUxNmZiOWNhNGY

绿地率 vs 绿视率

- 绿化是建成环境的重要要素，具有净化空气、缓解紧张情绪等作用，是空间规划关注的重要对象（如长久以来对田园城市的追求）
- 国家住房与城乡建设部的国家园林城市多批名单
- 平面的绿化（绿地率）与立体的绿化（绿视率）

清华大学 BCL

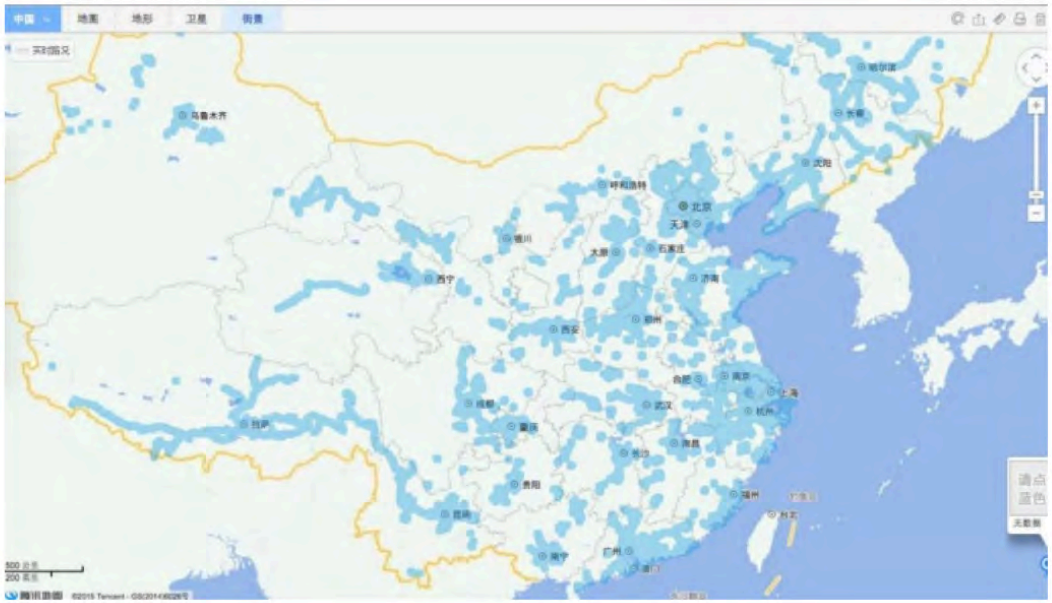
基于街景图片评价街道绿视率的技术路线



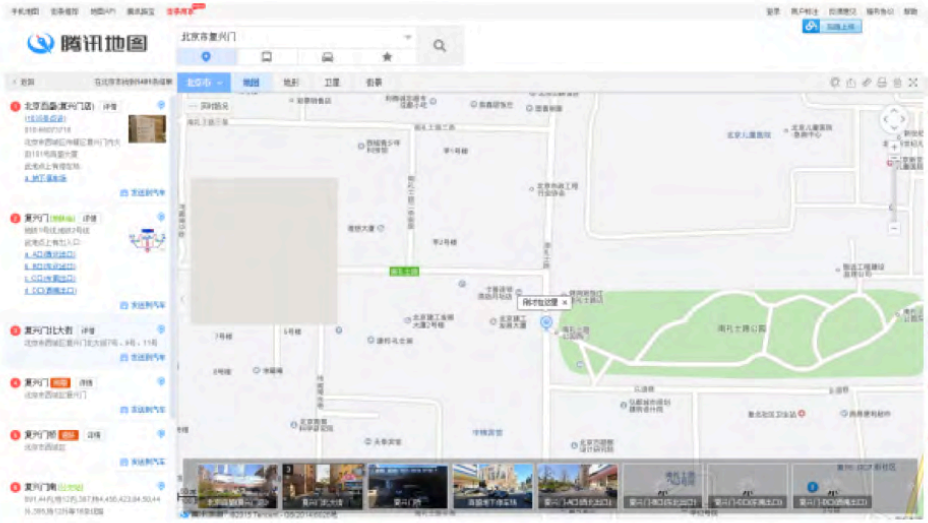
- 基于街景图片对街道绿视率进行评价，主要包括街景点提取（街道50m间隔）、街景图片抓取、街景图片识别以及绿视率统计分析四个环节

腾讯街景地图（中国最大的街景服务提供平台，具有时光机功能）

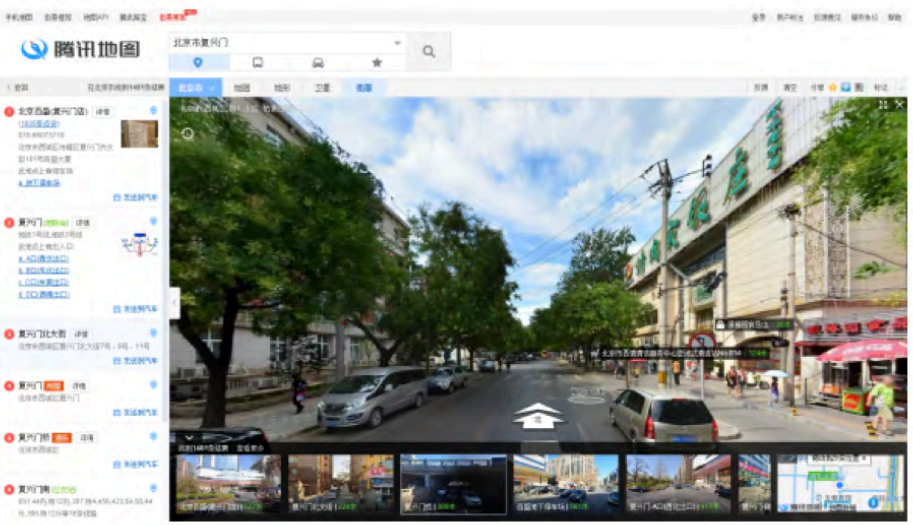
(a)



(b)



(c)



(d)

腾讯提供的街景图片抓取的API

http://apis.map.qq.com/ws/streetview/v1/image?size=600x480&pano=10011022120723095812200&pitch=0&heading=0&key=OB4BZ-D4W3U-B7VVO-4PJWW-6TKDJ-WPB77.

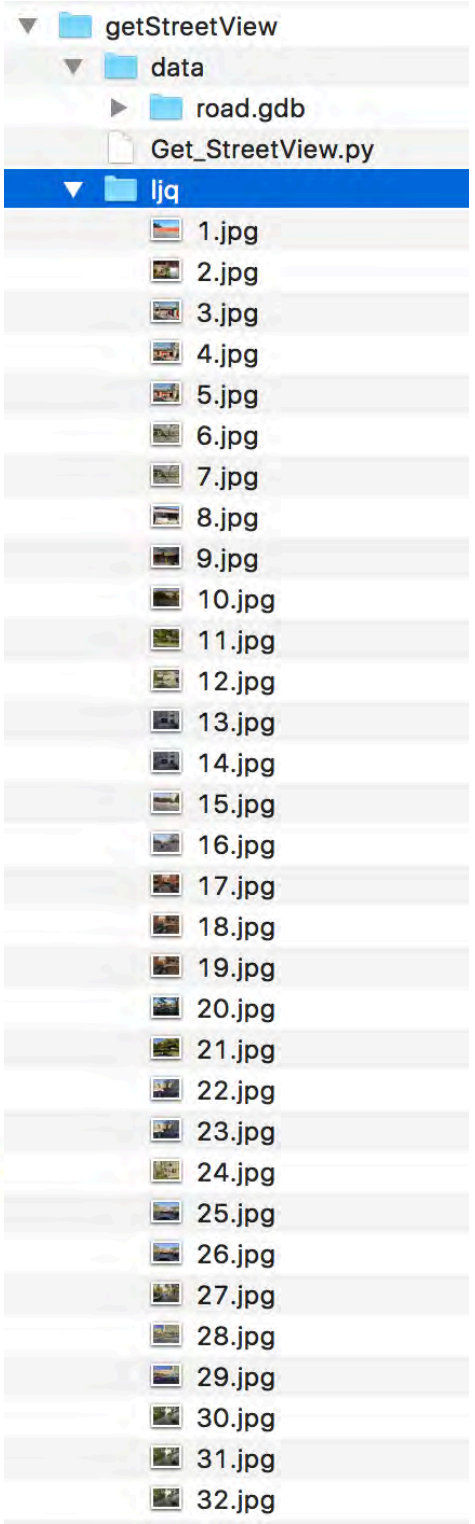
Parameter	Mandatory item or not	Description	Examples
size	Yes	Picture size in pixel, maximum width 960 px and height 640 px	size=138x187
location	One in location or pano	Coordinates or place name for confirming the street view location	location=Tsinghua University or location=39.12,116.83
pano		Street view ID for confirming the street view location	pano=10011022120723095812200
heading	No	The value of heading represents the angle the forward direction making with the north, which is measured in clockwise with a range from 0 to 360 degree (0 as the default value)	North: heading=0 East: heading=90 Sourth: heading=180 West: heading=270
pitch	No	The vertical angle of the camera covers -20 to 90 degree, in which a positive number stands for the level of looking up and vice versa (0 as the default value)	pitch=0
key	Yes	Developer's key (can be retrieved by online application)	key=OB4BZ-D4W3U-7BVVO-4PJWW-6TKDJ-WPB77

http://lbs.qq.com/panostatic_v1/guide-getImage.html

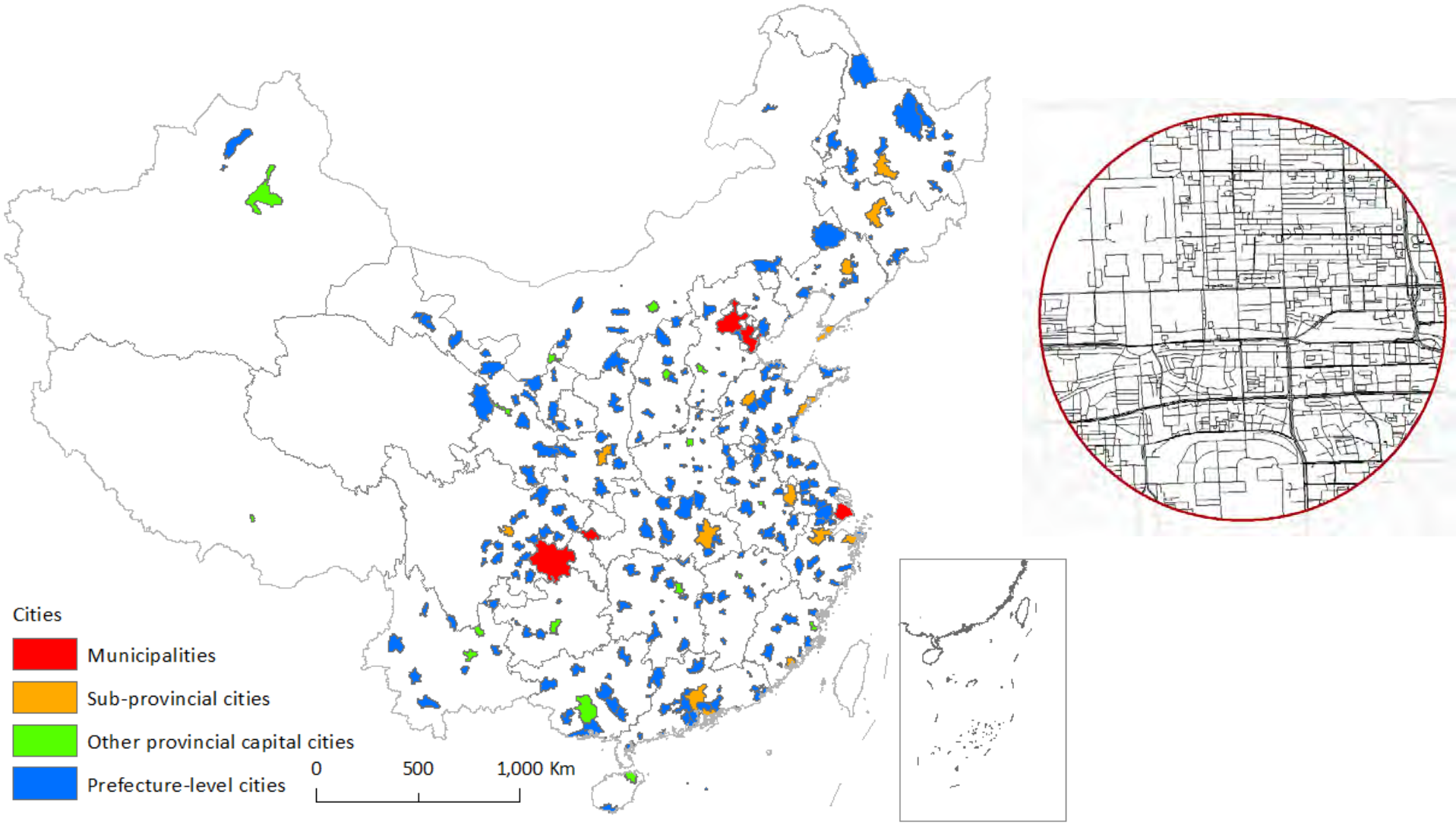
基于API抓取街景的代码示意

```
def save_file(path, file_name, data):
    if data == None:
        return

    mkdir(path)
    if(not path.endswith("/")):
        path=path+"/"
    file=open(path+file_name, "wb")
    file.write(data)
    file.flush()
    file.close()

#读取坐标
#获取经度坐标
def get_Long(shp):
    env.workspace="d:/data/road.gdb"
    #print 'Processing'
    #读取坐标
    cur=arcpy.SearchCursor(shp)
    #经度(x)
    point_Long=[]
    for row in cur:
        point_Long.append(row.POINT_X)
    return point_Long
#获取纬度坐标
def get_Lat(shp):
    env.workspace="d:/data/road.gdb"
    #print 'Processing'
    #读取坐标
    cur=arcpy.SearchCursor(shp)
    #经度(x)
    point_Lat=[]
    for row in cur:
        point_Lat.append(row.POINT_Y)
    return point_Lat

#获取街景所对应的点的ID
def get_FID(shp):
    env.workspace="d:/data/road.gdb"
    cur=arcpy.SearchCursor(shp)
    point_FID=[]
    for row in cur:
        point_FID.append(row.FID)
    return point_FID

#获取ID
def get_id(data):
    match1=re.search(r'id',data)
    if match1 is not None:
        ID=re.findall(r'"id": "(\w*)',data)
        return ID
```

```
#开始，获取街道节点坐标及其FID
print "start"
shp="shp_FeatureVerticesToPoints1.shp"
ponit_Long=get_Long(shp)    #经度
point_Lat=get_Lat(shp)  #纬度
FID=get_FID(shp)    #FID

i=0
for i in range(0,len(point_Lat)):
    print "Print Picture"+str(i+1)
    #print i
    #根据街道节点坐标获取街景ID
    url_point="http://apis.map.qq.com/ws/streetview/v1/getpano?location="+str(point_Lat[i])+","+str(ponit_Long[i])+
    "&radius=200&key=ULTBZ-VZ7WD-YLJ4Y-P4W7P-O2LLS-54F3J"
    ID_1=get_id(get_file(url_point))
    #ID存在，获取该ID对应的街景
    if ID_1 is not None:
        #除去读取出的街景ID前后多余字符
        ID_2=re.sub(r'\']','',str(ID_1))
        ID=re.sub(r'\[\'','',str(ID_2))
        #print ID
        #获取相应街景
        url="http://apis.map.qq.com/ws/streetview/v1/image?size=900x640&pano="+str(ID)+"&pitch=0&heading=0&
        key=ULTBZ-VZ7WD-YLJ4Y-P4W7P-O2LLS-54F3J";
        #print url
        #以该街景对应点的FID命名，保存该街景
        save_file("d:/ljq/",str(FID[i])+".jpg", get_file(url))
    i+=1
```

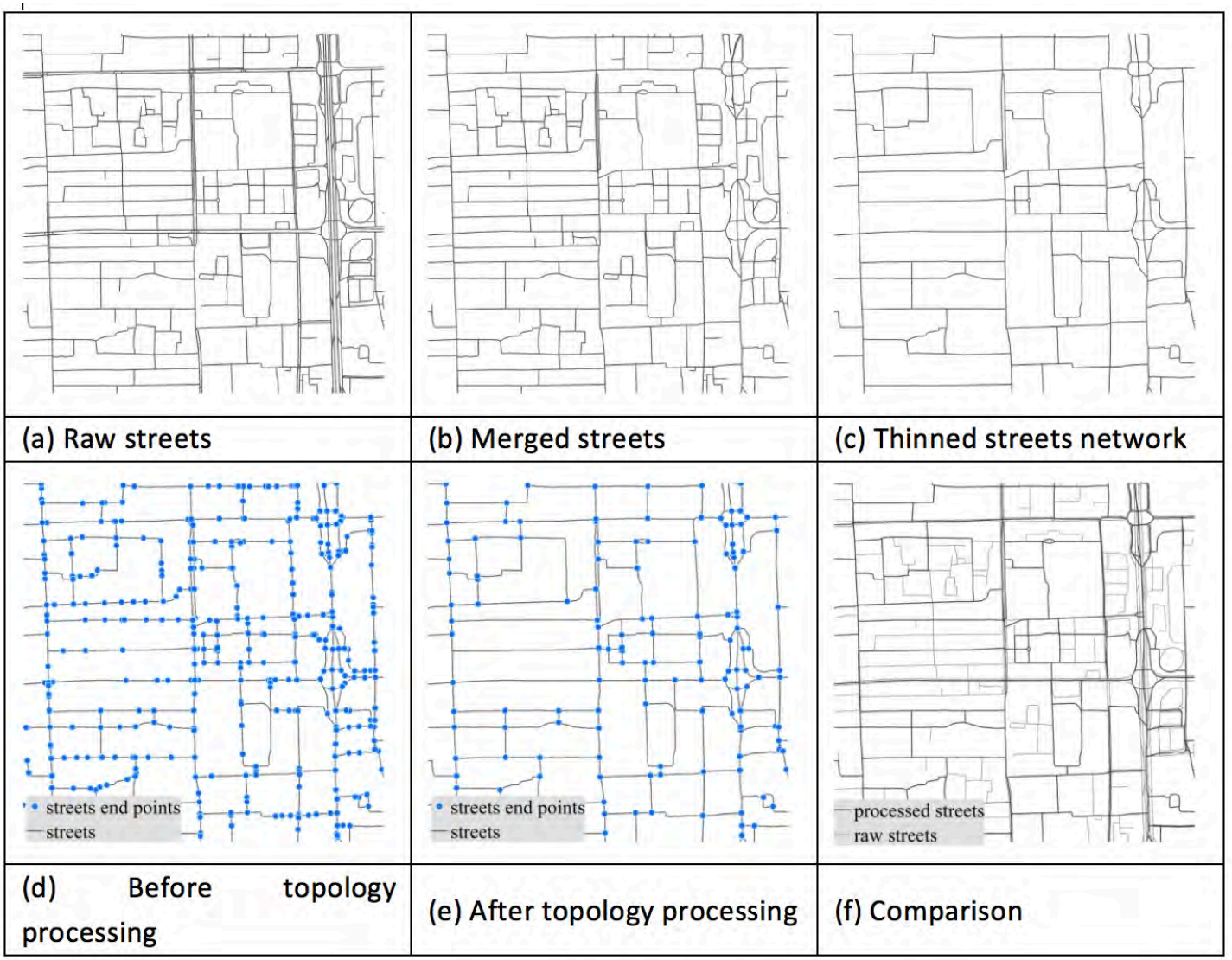


研究范围：中国288个地级及以上城市的中心区



- 4个直辖市，15个副省级城市，17个一般省会城市，252个地级市
- 考虑到有限的抓取时间和计算能力，选取每个城市中心的3km半径范围作为每个城市的研究范围（对应74.8万条街道）

街道数据预处理流程



- 为了更便捷地抓取街道上不同位置的街景图片（每隔50m），需要对街道数据进行预处理，需要合并街道、瘦化街道和拓扑处理三个步骤（ArcGIS中完成）

不同绿视率水平对应的街景图片示意

Figure 6 Street view pictures with various green percentages (only 127 locations/sites with the green ratio greater than 0.8)

not green <=0.2
somehow green (0.2-0.4]
green (0.4-0.5]
very green >0.5

清华大学
BCL

336,990个位置的平均绿视率为0.248

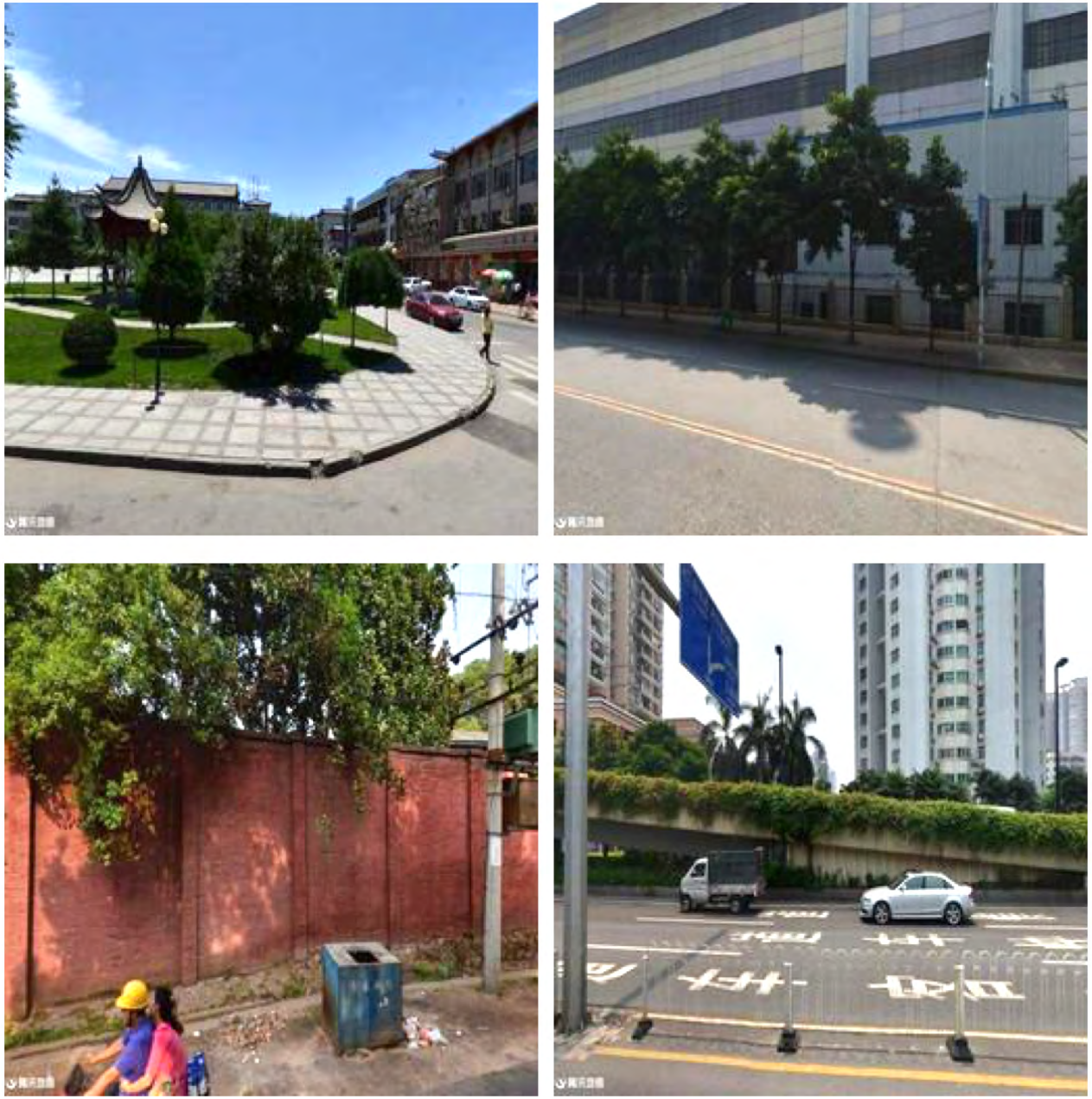


Figure 7 Typical SVPs with a green ratio of 0.248

北京中心城西北地区的街道绿化评价

- 基于各个点的街道绿视率计算结果，对街道层次的绿视率进行评价，较高的街道可以作为步行系统规划的参考

清华大学 BCL

131个有效城市的结果一览

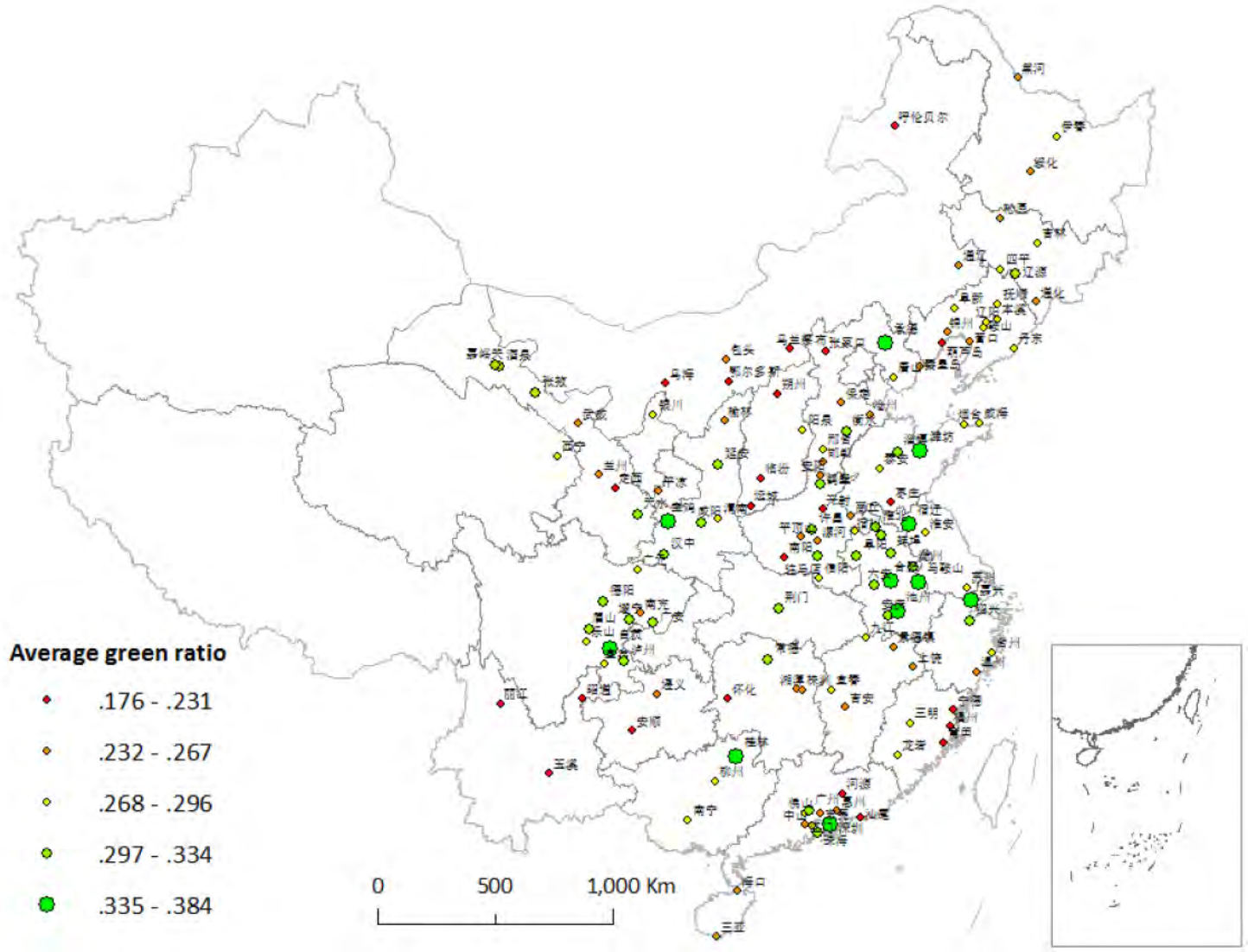
Type	# features	Min	Max	Mean	Green ratio			
					<0.2	0.2-0.4	0.4-0.5	>0.5
Locations	173,425	0.000	0.913	0.277	55,962 (32.3%)	85,702 (49.4%)	21,224 (12.2%)	10,537 (6.1%)
Street segments with over 13 locations per km)*	23,917	0.002	0.840	0.261	8,188 (34.2%)	12,619 (52.8%)	2,258 (9.4%)	852 (3.6%)
Blocks greater than 1 ha and with over 1 location per ha**	9,424	0.002	0.737	0.265	2,583 (27.5%)	5,931 (62.9%)	718 (7.6%)	192 (2.0%)

* “13” is the average value of location density for all street segments.
** “1” is the average value of location density for all blocks greater than 1 ha

- **部分城市的街景拍摄日期不适合评价街道绿视率（如秋冬季节）**
- **131个有效城市的平均街道绿视率范围为 0.132-0.384**

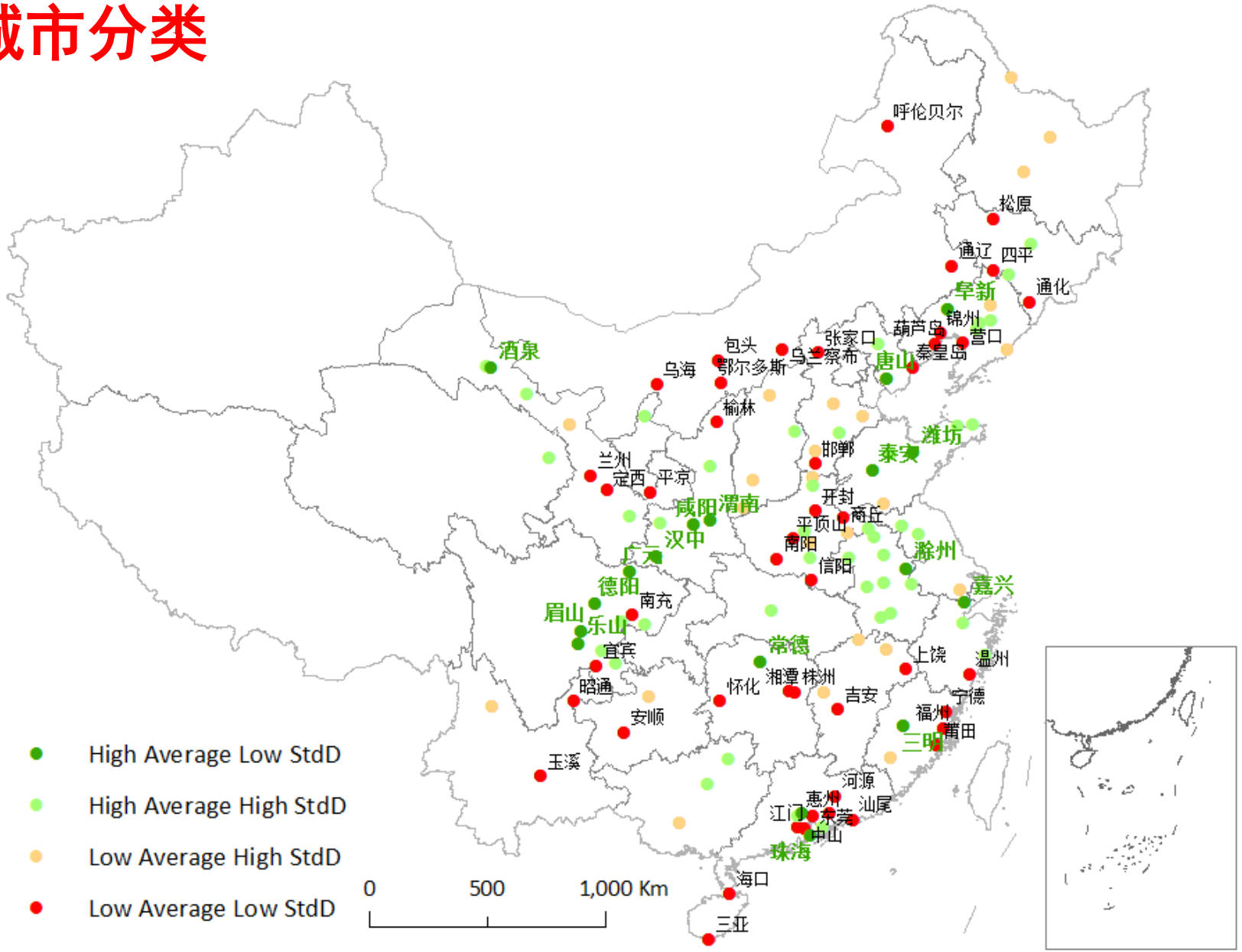
清华大学 BCL

131个有效城市的平均街道绿视率



- 前五城市均为国家园林城市（潍坊、自贡、宝鸡、马鞍山和承德）

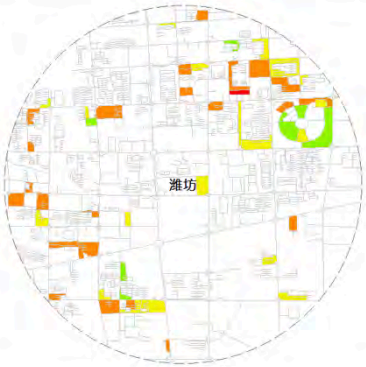
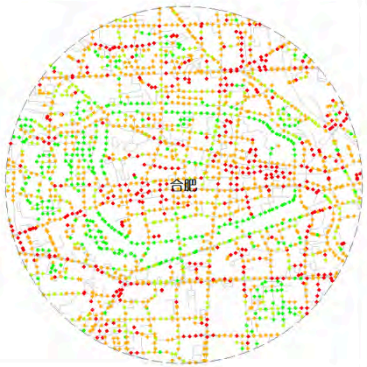
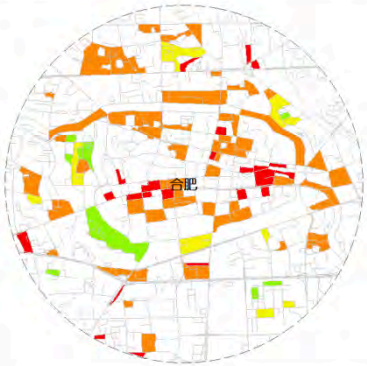
城市分类

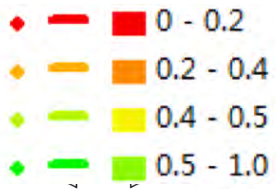


- 将131个有效城市根据街道绿视率的平均值与标准差，以中位数作为界线，分为四类
- 高均值低标准差（整体绿化好）、高均值高标准差（整体好但个别街道低）、低均值低标准差（整体绿化差）和低均值高标准差（整体差但个别街道高）。

典型城市的街道绿视率结果

Table 3 Street greenness for typical cities

City	Locations	Street segments	Blocks
Weifang			
Hefei			
Yanan			



街道绿视率的解释模型

Variables	Model1		Model2		Model3	
	Coefficients	*p* values	Coefficients	*p* values	Coefficients	*p* values
(Constant)						
CENTER	**0.061**	0.000	**0.053**	0.000	**0.052**	0.000
LENGTH	**0.080**	0.000	**0.093**	0.000	**0.093**	0.000
SIZE			-0.014	0.191	**0.025**	0.006
LEVEL			**-0.017**	0.041	-0.010	0.242
DENSITY			**-0.070**	0.000	**-0.088**	0.000
DESIGN			**0.083**	0.000	**0.108**	0.000
ECONOMY			**0.047**	0.000	**0.066**	0.000
MIDDLE					**0.017**	0.019
WEST					**0.060**	0.000
Adjusted R^2	0.010		0.027		0.029	

Note: coefficients in bold indicate being significant at the 0.05 level

- **在点的层面，街道绿化率与多个因素的回归模型显示，距离城市中心越远，街道倾向于更绿；街道长度更长，城市经济越发达、等级越高、人口密度越低，街道倾向于更绿；**

关于做研究的小窍门

- 第一讲/第二讲：参考文献的重要性
  - 外国人的姓名写法、认真与否、文献等级

- 第三讲：
  - 论文与报告的区别（是否有科学问题）
  - 问题：Problem vs Question

- **第四讲：两类论文**
  - **方法：证明方法优于已有的其他方法（效率/科学性、规划师/公众/同行评价？）**
  - **实证：证明发现，与其他人发现的异同，对理论的贡献**

课后安排

- 邀请外界人士介绍数据抓取的操作
  - 时间？
- 阅读材料：
  - Liu and Long 2016 EPB（地块）
  - 龙瀛和周垠 2016 新建筑（街道）
  - http://www.beijingcitylab.com/big-data-and-urban-planning/
- 答疑
  - ylong@tsinghua.edu.cn
  - 建筑学院新501办公室（默认每周五上午10:00-11:30）
    - 今天10-11AM

龙瀛 周垠 Long Ying Zhou Yin

街道活力的量化评价及影响因素分析

——以成都为例

Quantitative Evaluation on Street Vibrancy and Its Impact Factors: A Case Study of Chengdu

Article

B Planning and Design

Automated identification and characterization of parcels with OpenStreetMap and points of interest

Environment and Planning B:
Planning and Design
2016, Vol. 43(2) 341–360

DOI: 10.1177/0265813515604767
epb.sagepub.com
SAGE

Xingjian Liu
The University of Hong Kong, Hong Kong

Ying Long
Tsinghua University and Beijing Institute of City Planning, China

龙瀛, ylong@tsinghua.edu.cn, 新建筑馆501, 13661386623

BCL 北京城市实验室 Beijing City Lab

http://www.beijingcitylab.com

新浪微博：龙瀛a1_b2 北京城市实验室BCL

微信公众号：beijingcitylab

清华大学