

中国地质大学

实习报告



课程名称 高级空间分析建模课程报告一

教师姓名 晁 怡

本科生姓名 童川博

本科生学号 20221000679

本科生专业 地理与信息科学

所在院系 未来技术学院

日 期 2024 年 5 月 22 日

目录

1. 实习准备.....	1
1.1. 实习任务	1
1.2. 数据来源	1
1.3. 数据说明	1
2. 数据预处理.....	2
1.4. 数据准备	2
1.5. 异常值处理	2
1.6. 缺失值处理	2
1.6.1 删除数据大量缺失的省份.....	2
1.6.2 使用相邻城市填补缺失值.....	3
1.6.3 删除无法填补的数据.....	5
3. 数据统计特征描述.....	5
4. 数据总体特征描述.....	6
5. 实习总结.....	6

1. 实习准备

1.1. 实习任务

- 1) 采集/选择相关数据，根据数据具体情况进行相应预处理，包括缺失数据处理、异常值处理。
- 2) 对与处理后的数据进行平均值、极差、标准差、变异系数等的描述性统计，并对统计结果进行解释。
- 3) 选定一个特征，对该特征进行全局性分析，包括绘制劳伦兹曲线、计算基尼系数或集中化指数、计算熵值，并对每一个全局性分析结果进行解释。

1.2. 数据来源

本次实习使用的所有数据均来自教师提供，未标明出处。

- 2018 年中国各城市降水量年度数据
- 2018 年地级市空间数据
- 2019 年中国城市统计年鉴(精修版)

1.3. 数据说明

本次实习的目的是以城市为单位对全国城市的水资源供需和人均占有和水资源情况进行描述性分析。统计结果为 2018 年全年数据，包括 2018 年全国城市行政区划，户籍人口与常住人口，降水量，水资源总量，售水量和居民售水量数据。其中售水量和居民售水量的统计口径为市辖区，因此对应的常住人口统计口径也为市辖区；水资源总量的统计口径为全市，因此对应的户籍人口统计口径也为全市。

2. 数据预处理

1.4. 数据准备

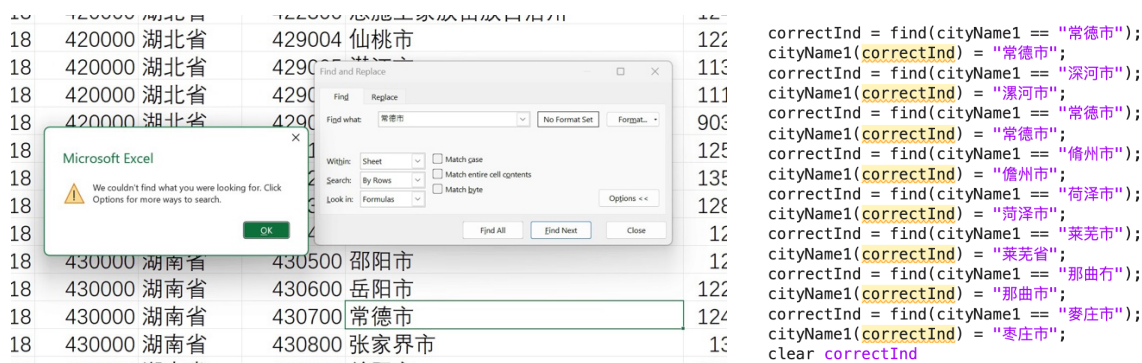
本次实习的所有数据处理均使用 MATLAB 实现，因此先将各数据表导入 MATLAB 并连接为一份表数据。

```
22 cityWater = sortrows(cityWater, 'cityName');
23 provinceInd = intersect(find(~[endsWith(cityName1, "省")]), find(~[endsWith(cityName1, "区")]));
24
25 cityName1 = cityName1(provinceInd);
26 waterSupply = waterSupply(provinceInd);
27 totalWaterResources = totalWaterResources(provinceInd);
28 hostPop = hostPop(provinceInd);
29 stayPop = stayPop(provinceInd);
30 ConsumerSupply = ConsumerSupply(provinceInd);
31
32 [cityName1, index] = sort(cityName1);
33 waterSupply = waterSupply(index);
34 totalWaterResources = totalWaterResources(index);
35 hostPop = hostPop(index);
36 stayPop = stayPop(index);
37 ConsumerSupply = ConsumerSupply(index);
38
39 for i = 1 : length(cityWater.provinceID)
40     if (cityWater(i,:).cityName ~= cityName1(i))
41         cityName1 = [cityName1(1:i-1); "", cityName1(i:end)];
42         waterSupply = [waterSupply(1:i-1); NaN; waterSupply(i:end)];
43         totalWaterResources = [totalWaterResources(1:i-1); NaN; totalWaterResources(i:end)];
44         hostPop = [hostPop(1:i-1); NaN; hostPop(i:end)];
45         stayPop = [stayPop(1:i-1); NaN; stayPop(i:end)];
46         ConsumerSupply = [ConsumerSupply(1:i-1); NaN; ConsumerSupply(i:end)];
47     end
48 end
```

图 1 使用城市编号为标识连接不同表格数据

1.5. 异常值处理

由于 1.4 节的连接算法是先对各数据排序并使用线性方式连接以提高效率，从而发现在统计表中出现了不同的地名书写。因此需要制作修正表统一地名书写。



The image shows a Microsoft Excel spreadsheet with a list of cities and their corresponding province codes. A 'Find and Replace' dialog box is open, showing the search for '常德市' and the replacement with '常德市'. The spreadsheet shows a list of cities with their province codes and names, and the dialog box is used to correct the city names.

图 2 使用映射表修正地名书写

1.6. 缺失值处理

1.6.1 删除数据大量缺失的省份

水资源具有很强的地理属性，可以认为相邻地区的水资源状况较相近。然而统计数据中部分省份的水资源和售水量数据存在大量缺失，难以使用相邻地区的水资源状况进行填补。如图 3 所示将缺失值超过 40% 的个省份进行删除，分别

是：海南省、贵州省、云南省、甘肃省、青海省和新疆维吾尔自治区。

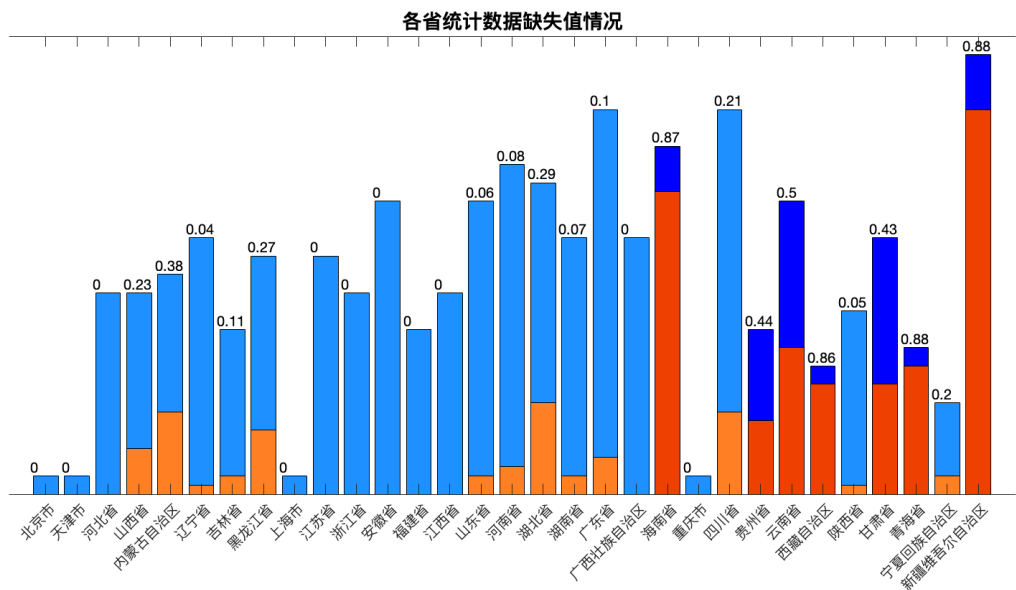


图 3 各省水资源统计数据缺失值占比情况

1.6.2 使用相邻城市填补缺失值

考察城市空间关系以使用相邻城市值填补缺失值。在 QGIS 中打开中国城市 Shapefile 文件，连接属性表，直观上看一下降水数据检查数据是否正确。

全国部分省市2018年降水分布图

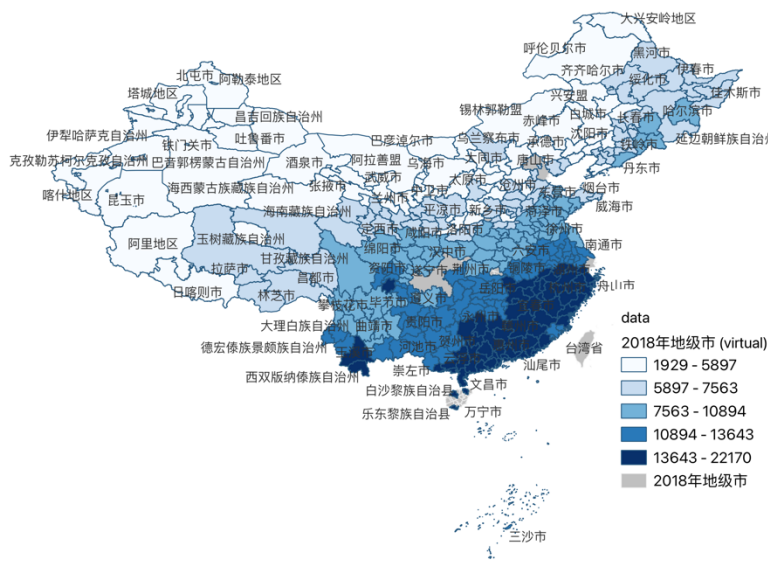


图 4 全国部分省市 2018 年降水分布图

如图 4 数据正确连接，进一步对属性表查询筛选出数据完备的城市，并在剔除不予考虑的省份后标注待填充数据的城市。如图 5 图 6 所示，深灰色为数据完整的城市，蓝色为待填充的城市。可以发现西部城市，海南省和港澳台地区数据大量缺失，部分中东部地区城市也有缺失。

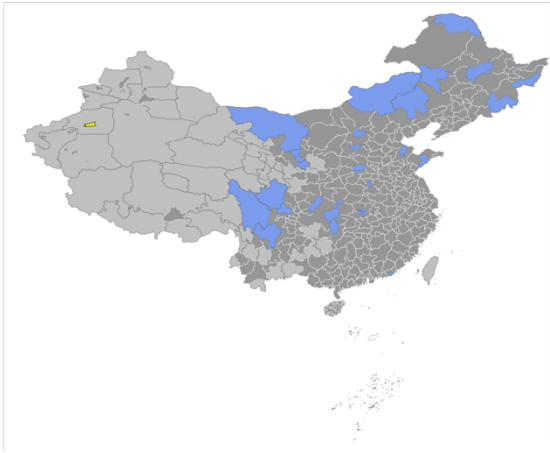


图 5 全国城市水资源总量统计情况调查

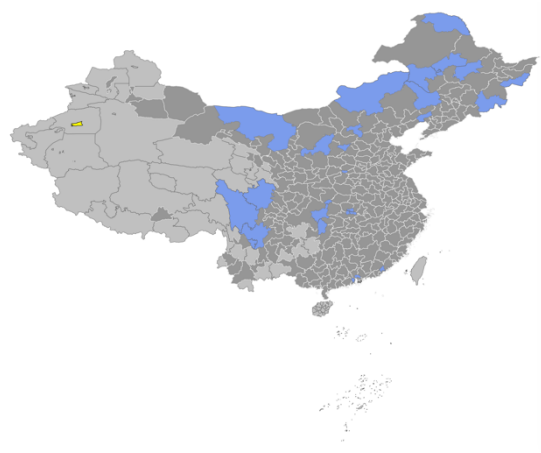


图 6 全国城市售水量统计情况调查

并如图 7 所示在 QGIS 中使用 Python 脚本统计每一个城市的相邻城市并新建字段填入相邻城市代码。得到结果如图 8 所示。

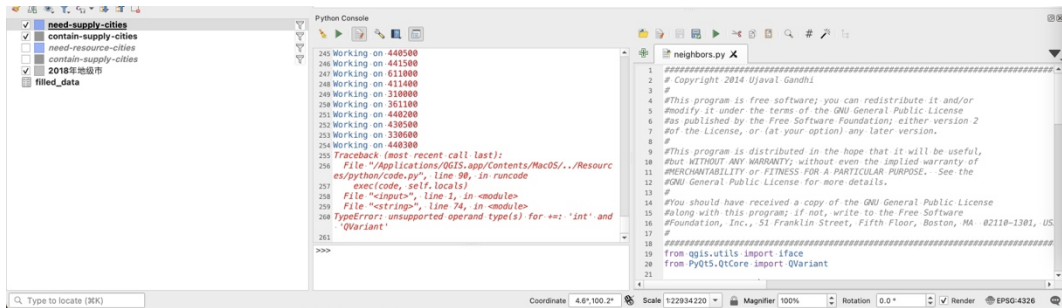


图 7 使用脚本统计相邻城市代码

	W	X	Y	Z	AA	AI
prep	supply	resource	NEIGHBORS	SUM	filled_data	
2	2580.603	-1	-1	653200,653100,659002,653000,654000,652800	-6	
3	1242.994	-1	-1		652900	-1
4	5793.456	-1	-1	150600,150800,620400,640500,620600,620300,640200,640100,150300,640300,620700,620900		41681
5	3484.542	-1	-1	659005,654200,650100,652300		27459
6	5244.644	-1	-1	653200,540200,540600,652800		-4
7	9057.473	1634	880600	511700,500000,610100,611000,420300,610700		232090
8	13333.802	7505	920300	340700,340100,341700,421100,341500,360400		72241
9	12890.855	3436	633600	522700,522300,520200,520100,520500		38204
10	7094.040	5686	119992	140400,410900,410600,130400,410700		32248
11	6613.305	11016	195600	210100,210200,210600,211100,211000,210700,210800		118061
12					440400	36614
13	4283.991	1894	561200	150600,150200,152900		17939
14	2859.013	-1	-1	653200,652900,542500,654000,659006,654200,650400,650100,652300,540600,632700,632800,620900,650500		34478
15	10272.431	4662	716800	511700,510800,511300,610700		21883
16	5402.898	1011	256000	230200,220700,150500,230600,152200		26440
17	20295.189	-1	-1	460400,469030,469026,469027,469001		-5
18	8784.343	1596	1012000	220500,220200,222400		10209
19	5018.972	2354	23200	640400,620800,620100,621100,640500,152900,620600		29005
20	12355.758	3279	2510000	451400,451200,450100,522700,532600,530300,522300		58463
21	12578.448	8512	315943	340600,341100,321300,340400,341600,341300		27695
22	4938.071	15018	72600	150600,150100,150900,150800		18467
23	7198.025	4948	1316400	620800,610400,610100,610700,621200,620500		101967
24	5340.209	9142	254300	110000,130700,140900,140200,131100,130900,131000,130100		153483
25	11853.184	1802	1473300	533300,530900,533100,532900		1159
26	16679.129	-1	-1	469030,469028,469027,460200,469001		11263
27	14640.566	7211	266000	450700,450900,440800		24147
28	5590.117	115852	354573	130800,130700,130600,131000,120000		96819
29	3568.409	-1	-1		654300	-1
30	7354.687	5624	281700	220500,210400,210100,210600,211000		73617
31	11286.861	3785	1285430	510500,520300,530300,520200,520400,520100,530600		51023
32	7051.622	9685	-1	371400,370500,130900,370300,370100		70616
33	9556.659	2597	269775	340600,340300,340400,341200,411400,411600		33838
34	2156.479	-1	-1	654000,659007,654200		-3
35	6344.313	4147	87000	371400,371600,131100,130600,131000,120000		104646
36	7303.500	-1	-1	540400,513300,540600,632700,533400		-5

图 8 统计得到相邻城市代码

计算完成后导出字段，回到 MATLAB 中如所示将该城市的水资源总量和售水量按照图 9 方式相邻城市数据填充结果。

```

10 len = length(data.cityName);
11 deleteInd = 0;
12 deleteCities = [];
13 for i = 1 : len
14     if (data.hostPop(i) == -1 || data.stayPop(i) == -1)
15         deleteInd = deleteInd + 1;
16         deleteCities(deleteInd) = i;
17         continue;
18     end
19     if (data.supply(i) == -1)
20         if (data.neighbors(i) == "")
21             deleteInd = deleteInd + 1;
22             deleteCities(deleteInd) = i;
23             continue;
24         end
25         neighbor = split(data.neighbors(i), ',');
26         count = 0;
27         value = 0;
28         for j = 1 : length(neighbor)
29             orient = find(data.cityID == str2num(neighbor(j)));
30             if (data.supply(orient) ~= -1)
31                 count = count + 1;
32                 value = value + data.supply(orient);
33             end
34         end
35         if (count == 0)
36             deleteInd = deleteInd + 1;
37             deleteCities(deleteInd) = i;
38             continue;
39         end
40         data.supply(i) = round(value / count);
41     end
42     clear count value
43

```

图 9 使用相邻城市数据填充数据核心代码

1.6.3 删除无法填补的数据

在以 1.6.2 节方式填充数据后依然有部分城市数据存在缺失(如部分沿海岛屿在几何上没有相邻城市)和少量城市的人口数据存在缺失，因此将这小部分数据删除。最终从 367 个地级市中得到 213 组数据。

3. 数据统计特征描述

对中国各城市各水资源指标进行统计性描述，得到图 10 结果。

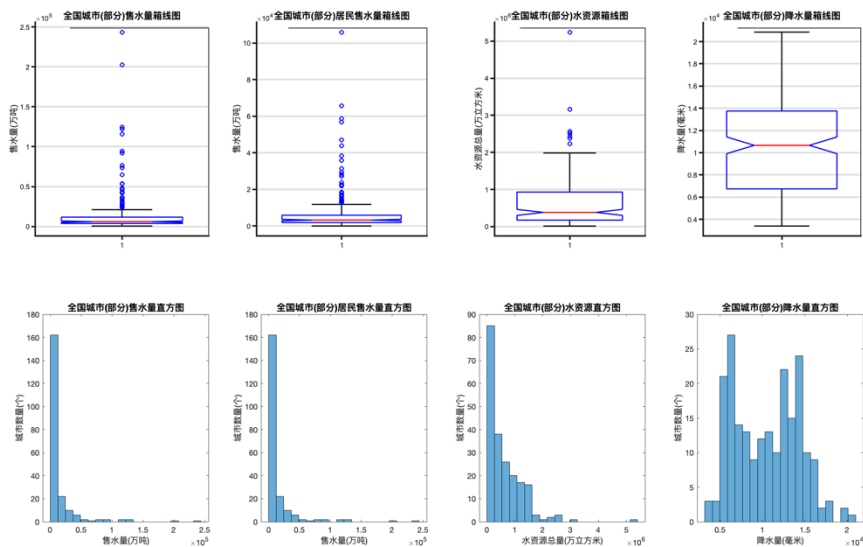


图 10 城市各水资源指标箱线图和直方图

由于西部城市数据存在大量缺失，因此本样本反映的结果贴近中东部地区水资源情况。中东部地区降水量分布呈双峰分布，总体数据差异较小，箱体较大即数据在上下四分之一中位数内分布较多，极差为 177456 毫米，标准差为 3882.1 毫米。而售水量，居民售水量和水资源总量数据峰度和偏度极大，大部分城市统计值较小而少量城市统计值较大。水资源总量数据箱体大于售水量数据，峰度小于售水量数据，表明在中东部地区资源禀赋差距小于社会生产生活实际用水差距。同时生活用水与生产生活用水总体趋势一致。在标准化后两组数据的均方根均为 0.9976。然而，由于各城市自然条件差距大，城市面积与人口也相差很大，进一步对人均占有量的研究具有现实意义。

4. 数据总体特征描述

表 1 为全国(部分)城市数据变异系数。prep 为降水量，supply 为售水量，resource 为水资源总量，contain 为人均水资源持有量，consume 为人均居民售水量。

可见尽管各城市售水量，水资源总量和人均水资源持有量差距较大，但是与生活息息相关的人均居民售水量差距明显减小。进一步地可以使用洛伦兹曲线对数据描述。

`struct` with fields:
prep: 0.0632
supply: 0.5306
resource: 0.3848
contain: 0.5307
consume: 0.0866

表 1 各数据信息熵数据

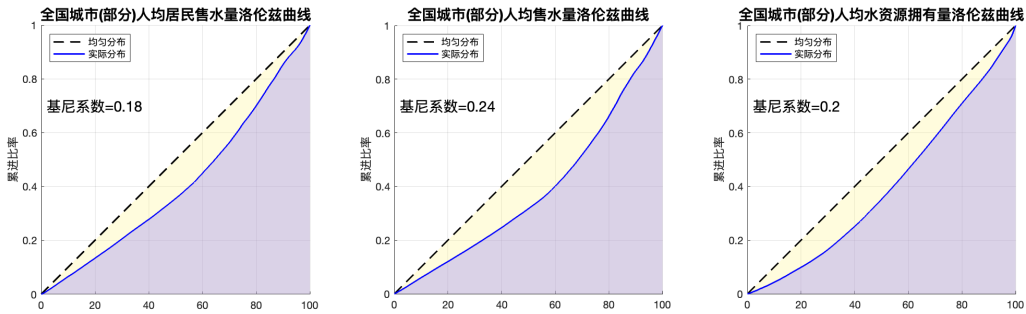


图 11 全国城市(部分)各人均水资源指标洛伦兹曲线

可以发现虽然各城市各水资源指标差异很大，但是在人均指标上反映并不明显，三个指标的基尼系数为 0.18, 0.24 和 0.2。

5. 实习总结

比较小清新的一次实习，图画得很爽(?)。第一次在常规的统计中加入了运用 GIS 技术进行分析，体验感觉不错。

手搓了洛伦兹曲线和基尼系数的计算，对二者的统计学含义与计算方法的理

解有了进一步加深。

```
geni = round((1 - trapz(xx,consumeFit(xx)) * 2)*100)/100;  
hold on  
area(xx,'FaceColor',colorplus(285),'FaceAlpha',0.3,'EdgeColor','none');  
plot([0,1],[0,1],'LineStyle','--','LineWidth',1.5,'Color',[0,0,0]);  
area(xx,consumeFit(xx),'FaceColor',colorplus(40),'FaceAlpha',0.3,'EdgeColor','none');  
plot(xx,consumeFit(xx),'LineStyle','-','LineWidth',1.5,'Color','blue');  
legend('','均匀分布','','实际分布','Location','northwest');  
text(0.02,0.7, strcat("基尼系数=", num2str(geni)),FontSize=14);  
hold off
```

图 12 手搓洛伦兹曲线核心代码