这个文档包括三部分（代码和可视化图片在相应文件夹里）：

1 数据预处理部分的整理（基本与期中一样，因为整理了代码，文档里有个图换了一下）

2 聚类可视化

3 预测结果的部分可视化与分析

1. 预处理

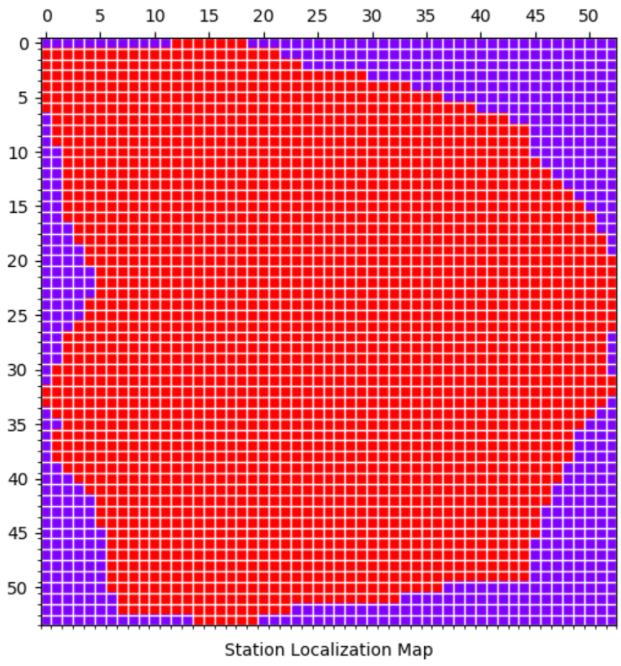
1）确定基站编号与地理位置间的关系

之所以我们需要得到基站与网格分布的对应关系，是因为每个网格的人口数据不仅是时序的，还与空间分布有关。例如某网格人数的增加与周围网格人数减少相关。前期的数据预处理得到网格分布可能为后续预测提供方便。

原始基站编号---wkt 坐标对应形式如下：

```
格网编号,wkt坐标
12,"MULTIPOLYGON (((433943.69714851433 4393540.4441694766, 434943.69714851433 4393540.4441694766, 434943.69714851433 4394540.4441694766, 433943.69714851
13,"MULTIPOLYGON (((434943.69714851433 4393540.4441694766, 435943.69714851433 4393540.4441694766, 435943.69714851433 4394540.4441694766, 434943.69714851
14,"MULTIPOLYGON (((435943.69714851433 4393540.4441694766, 436943.69714851433 4393540.4441694766, 436943.69714851433 4394540.4441694766, 435943.69714851
15,"MULTIPOLYGON (((436943.69714851433 4393540.4441694766, 437943.69714851433 4393540.4441694766, 437943.69714851433 4394540.4441694766, 436943.69714851
16,"MULTIPOLYGON (((437943.69714851433 4393540.4441694766, 438943.69714851433 4393540.4441694766, 438943.69714851433 4394540.4441694766, 437943.69714851
17,"MULTIPOLYGON (((438943.69714851433 4393540.4441694766, 439943.69714851433 4393540.4441694766, 439943.69714851433 4394540.4441694766, 438943.69714851
18,"MULTIPOLYGON (((439943.69714851433 4393540.4441694766, 440943.69714851433 4393540.4441694766, 440943.69714851433 4394540.4441694766, 439943.69714851
53,"MULTIPOLYGON (((421943.69714851433 4394540.4441694766, 422943.69714851433 4394540.4441694766, 422943.69714851433 4395540.4441694766, 421943.69714851
54,"MULTIPOLYGON (((422943.69714851433 4394540.4441694766, 423943.69714851433 4394540.4441694766, 423943.69714851433 4395540.4441694766, 422943.69714851
55,"MULTIPOLYGON (((423943.69714851433 4394540.4441694766, 424943.69714851433 4394540.4441694766, 424943.69714851433 4395540.4441694766, 423943.69714851
56,"MULTIPOLYGON (((424943.69714851433 4394540.4441694766, 425943.69714851433 4394540.4441694766, 425943.69714851433 4395540.4441694766, 424943.69714851
57,"MULTIPOLYGON (((425943.69714851433 4394540.4441694766, 426943.69714851433 4394540.4441694766, 426943.69714851433 4395540.4441694766, 425943.69714851
58,"MULTIPOLYGON (((426943.69714851433 4394540.4441694766, 427943.69714851433 4394540.4441694766, 427943.69714851433 4395540.4441694766, 426943.69714851
59,"MULTIPOLYGON (((427943.69714851433 4394540.4441694766, 428943.69714851433 4394540.4441694766, 428943.69714851433 4395540.4441694766, 427943.69714851
60,"MULTIPOLYGON (((428943.69714851433 4394540.4441694766, 429943.69714851433 4394540.4441694766, 429943.69714851433 4395540.4441694766, 428943.69714851
61,"MULTIPOLYGON (((429943.69714851433 4394540.4441694766, 430943.69714851433 4394540.4441694766, 430943.69714851433 4395540.4441694766, 429943.69714851
62,"MULTIPOLYGON (((430943.69714851433 4394540.4441694766, 431943.69714851433 4394540.4441694766, 431943.69714851433 4395540.4441694766, 430943.69714851
63,"MULTIPOLYGON (((431943.69714851433 4394540.4441694766, 432943.69714851433 4394540.4441694766, 432943.69714851433 4395540.4441694766, 431943.69714851
64,"MULTIPOLYGON (((432943.69714851433 4394540.4441694766, 433943.69714851433 4394540.4441694766, 433943.69714851433 4395540.4441694766, 432943.69714851
65,"MULTIPOLYGON (((433943.69714851433 4394540.4441694766, 434943.69714851433 4394540.4441694766, 434943.69714851433 4395540.4441694766, 433943.69714851
66,"MULTIPOLYGON (((434943.69714851433 4394540.4441694766, 435943.69714851433 4394540.4441694766, 435943.69714851433 4395540.4441694766, 434943.69714851
```

- 原始数据中发现基站编号不是从零开始连续的，说明有些位置如山、



河等没有人流量数据。

- 只给出了基站编号对应的 wkt 坐标数据，需要处理得到基站编号基站位置分布的关系，如上图

图中显示，基站所覆盖区域被基站划分为 54 行 53 列的网格，对应 2862 个基站，且基站编号从网格左上角开始从左到右从上到下依次加一。图中红色网格代表提供人口流动数据的基站，蓝色网格则是无人区域。

基于此处理基站与网格分布关系，得到三个表格待后续使用：

a) 基站编号---网格行列转换表

b) 网格---基站编号转换表

c) 无人区基站编号表

2）数据结构化、添补数据缺失等

原始数据形如：

日期	小时数	网格编号	驻留人数	出发人数	到达人数
20170901	0	13 161 8	3		
20170901	0	15 803 14	12		
20170901	0	17 1613	30	18	
20170901	0	53 126 8	3		
20170901	0	55 581 7	8		
20170901	0	57 914 13	8		
20170901	0	59 73 3	3		
20170901	0	61 147 5	4		
20170901	0	63 820 6	6		
20170901	0	65 39 2	3		
20170901	0	67 133 3	6		
20170901	0	69 1332	21	19	
20170901	0	71 210 5	7		
20170901	0	73 1101	43	23	
20170901	0	107 602 6	6		
20170901	0	109 819 5	6		

数据范围是 2017 年 9 到 11 月人口流动数据。每天分 24 小时记录数据，每一行给出某天第几个小时第几号网格的，驻留人数、出发及到达人数。

为了便于后续数据处理，将数据形式进行处理。以驻留人数为例，每行表示某天某小时的所有基站数据，共 2862 列。无人区补零。

还有很少数的基站在一些时间点有数据缺失的情况，用该时刻前或后一时刻的数据填补。

A	DD	DE	DF	DG	DH	DI	DJ	DK	DL	DM	DN	DO	DP	DQ	DR	DS
	106	107	108	109	110	111	112	113	114	115	116	117	118	119	120	121
2017/9/16 0:00	595	706	361	1080	1598	694	879	1214	491	237	1206	75	45	295	195	1148
2017/9/16 1:00	583	706	355	1080	1605	694	883	1214	486	237	1224	75	45	295	196	1148
2017/9/16 2:00	583	704	355	1091	1605	702	883	1225	486	239	1224	75	45	301	196	1152
2017/9/16 3:00	581	704	367	1091	1616	702	884	1225	491	239	1248	75	47	301	199	1152
2017/9/16 4:00	581	705	367	1110	1616	712	884	1239	491	249	1248	79	47	303	199	1158
2017/9/16 5:00	632	705	379	1110	1636	712	909	1239	511	249	1282	79	59	303	202	1158
2017/9/16 6:00	632	751	379	1159	1636	731	909	1301	511	303	1282	90	59	290	202	1094
2017/9/16 7:00	701	751	411	1159	1650	731	902	1301	582	303	1309	90	69	290	236	1094
2017/9/16 8:00	701	790	411	1129	1650	760	902	1342	582	344	1309	102	69	252	236	991
2017/9/16 9:00	749	790	425	1129	1536	760	875	1342	573	344	1267	102	73	252	232	991
2017/9/16 10:00	749	771	425	1074	1536	767	875	1348	573	327	1267	106	73	240	232	913
2017/9/16 11:00	695	771	400	1074	1506	767	833	1348	588	327	1293	106	81	240	187	913
2017/9/16 12:00	695	758	400	1103	1506	741	833	913	588	359	1293	112	81	233	187	926
2017/9/16 13:00	698	758	411	1103	1466	741	692	913	588	359	1316	112	81	233	236	926
2017/9/16 14:00	698	737	411	1043	1466	732	692	887	588	359	1316	116	81	231	236	924
2017/9/16 15:00	698	737	400	1043	1503	732	689	887	586	359	1291	116	74	231	203	924
2017/9/16 16:00	698	743	400	1015	1503	746	689	844	586	356	1291	96	74	229	203	965
2017/9/16 17:00	693	743	403	1015	1471	746	703	844	545	356	1307	96	63	229	205	965
2017/9/16 18:00	693	723	403	996	1471	735	703	838	545	336	1307	93	63	251	205	1062

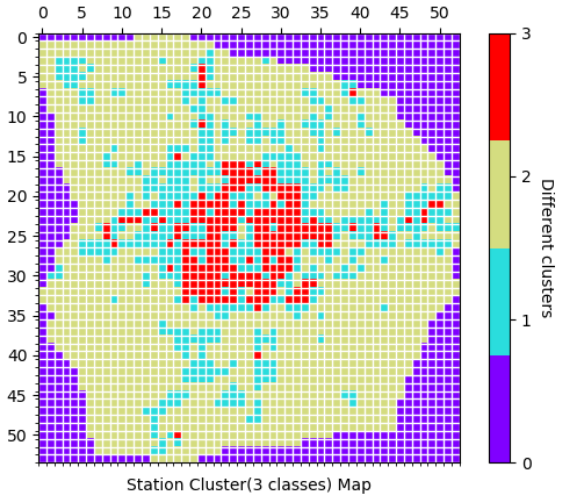
最终得到预处理数据上图。如可以读出 9 月 16 号 0 零点，106 号基站监测到的驻留人数是 595 人。

2. 聚类可视化

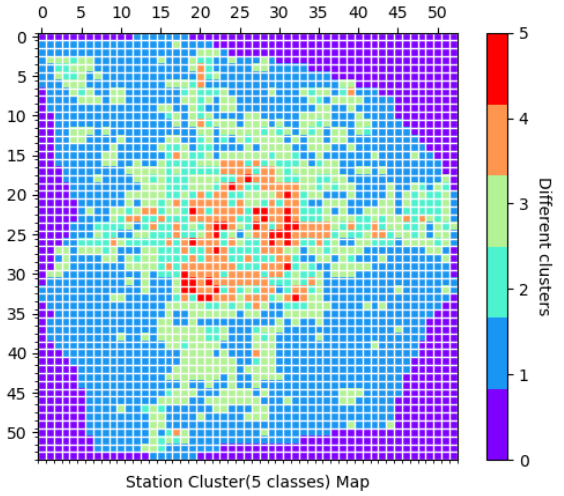
聚类本来是想，分类后对每个类别的网格分别进行建模（后来没用~）但也可以对网格进行分析。

将基站 3 个月的驻留人数时序数据作为特征向量，对基站通过 Kmeans 算法进行聚类，得到结果可视化如下图。

聚类中心为 3 的情况：



聚类中心为 5 的情况：



分析：聚类结果基本是按照距市中心远近划分的，当然也有一些特殊点。一般的，城市中心人口多、迁移量大，城市外围人口少、迁移量小的。

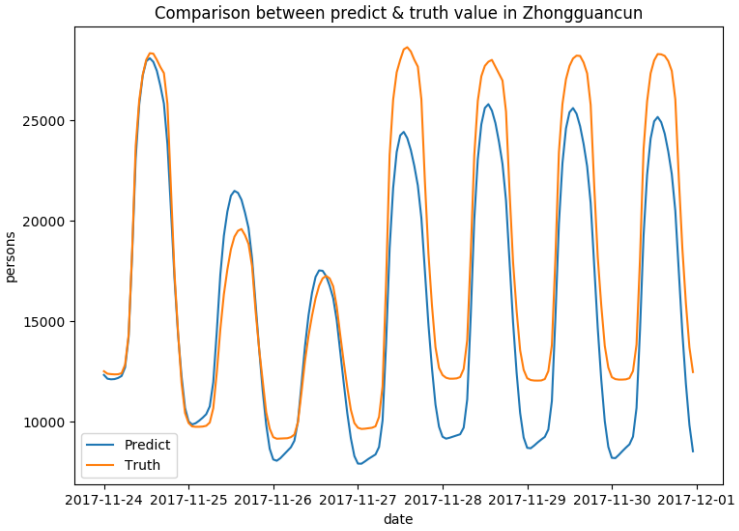
聚类结果与基本常识是相符的。

3. 预测结果部分可视化

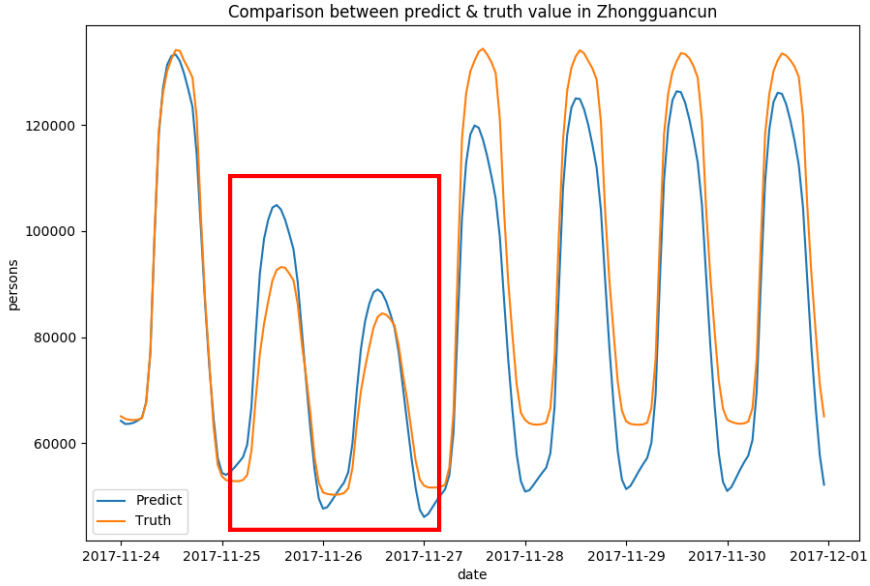
模型利用 2017.9.1~2017.11.23 的驻留人口数据进行训练，预测 2017.11.24~2017.11.30 共 7 天 168 小时的驻留人口数据。

在通过 wkt 坐标获得基站实际经纬度坐标后，可以针对感兴趣的地理位置进行分析。

1. 如查询到中关村经纬度坐标(39.98, 116.31)，对应到基站编号为 1715 网格坐标 (19,32)，绘制 7 天的预测与真是每小时驻留人数曲线进行对比。

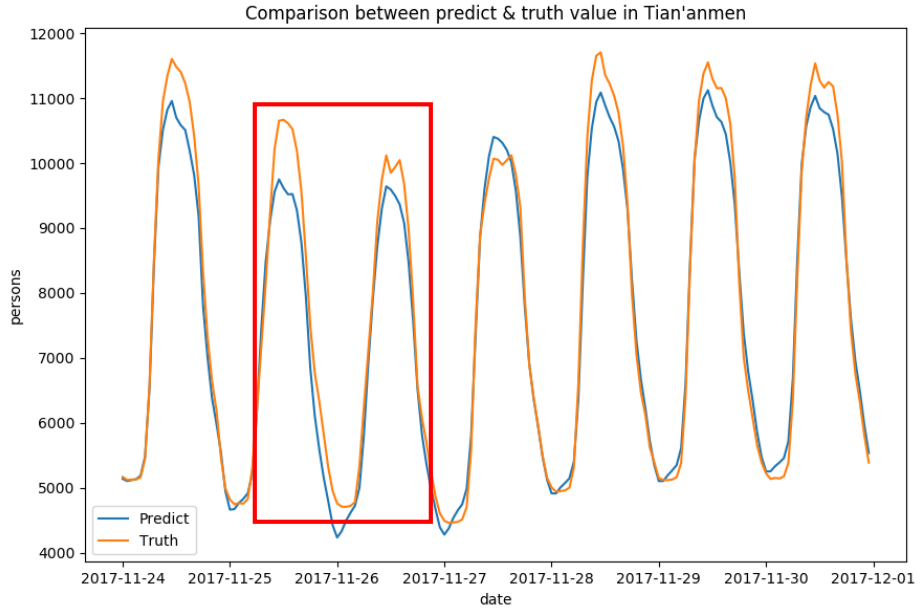


取该地区附近的 9 个基站求和（编号：[1714,1715,1716,1660,1661,1662,1804,1805,1806]）的结果



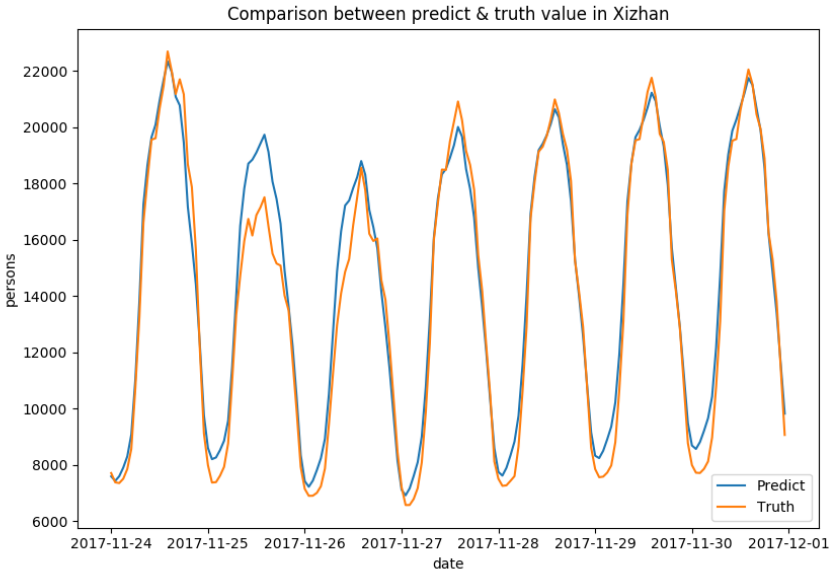
分析：预测结果与真值的符合程度很高。同时可以看出预测时间越靠后误差会增大。

2. 天安门(116.39, 39.90) 对应基站编号 1245(26, 23)

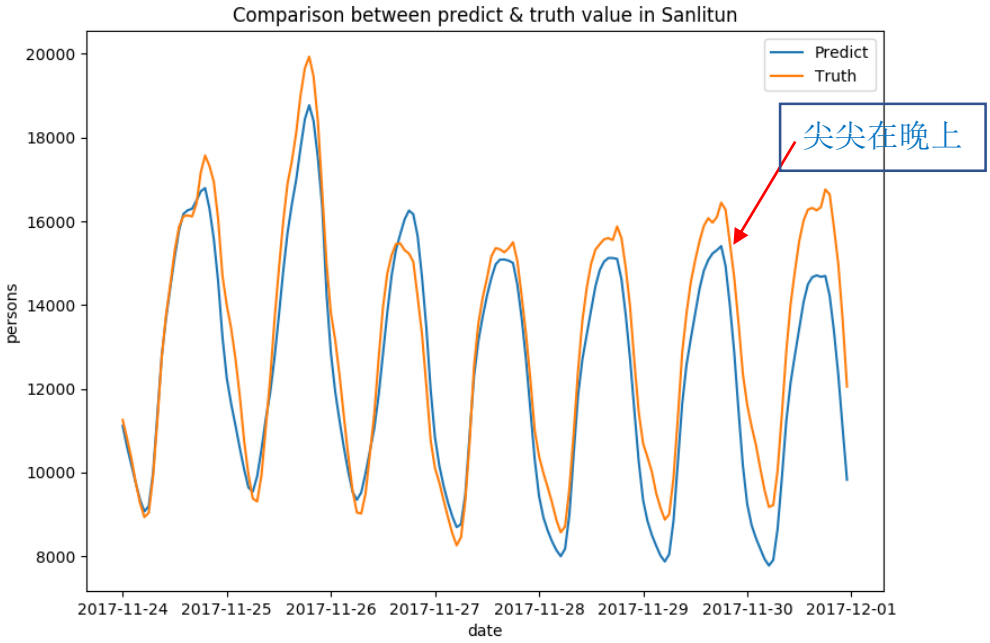


分析：25 和 26 号是周末，对比中关村和天安门的人数有较大的区别。中关村人数较周中有显著降低，预测值较周中有所降低，但仍略高于真实值。天安门在周末的人数仍维持在较高水平，预测值略低于真实值。这种在整体趋势中发生一定突变的情况对于模型预测是较难的。

3. 西站 经纬度(39.892 116.315) 基站编号 1185 (19,22)



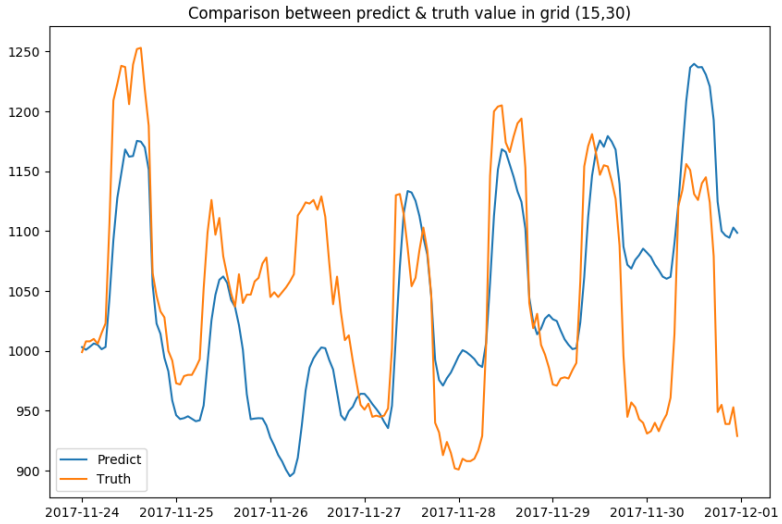
4. 三里屯 经纬度(39.929 116.443) 基站编号 1408 (30,26)



分析：三里屯最特别的一点是周五周六人数多于其他时间。

（还有一个不是很明显的分析，中关村、天安门一天里的高峰偏上午，尖尖在早上，三里屯的尖尖偏晚上，然后预测值对这种细节还不是很贴近，恩~）

5. 最后找了一个预测不太贴合的格子，在南坞桥附近。经纬度(39.9642 116.2675)　编号 1605(15,30)



分析：大概原因是这个地方离城市中心较远，而且没有较明显的人口变化规律，所以预测结果误差较大。