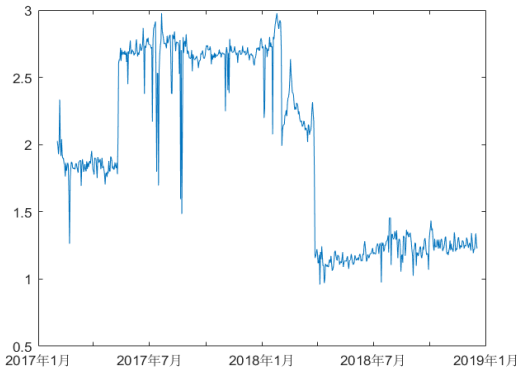
解题思路

一、观察数据

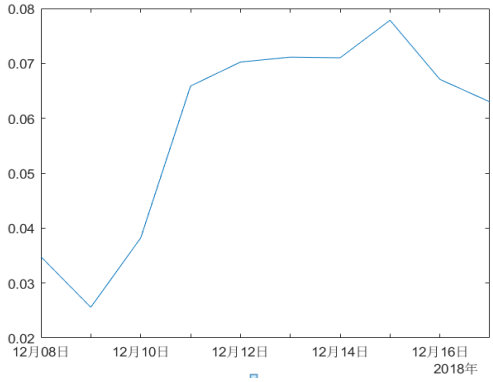
初步观察训练集数据，发现结构为大量传感器检测的人流量数据，每个传感器已知人流量数据为 2017 年 2 月 1 号到 2018 年 12 月 17 号之间全部或者部分数据，但是结束时间都为 2018 年 12 月 17 号，需要预测测试集中每个传感器在 2018 年 12 月 18 号到 2019 年 3 月 18 号之间的人流量数据，但是测试集中有比较多的重复数据。

二、选定模型

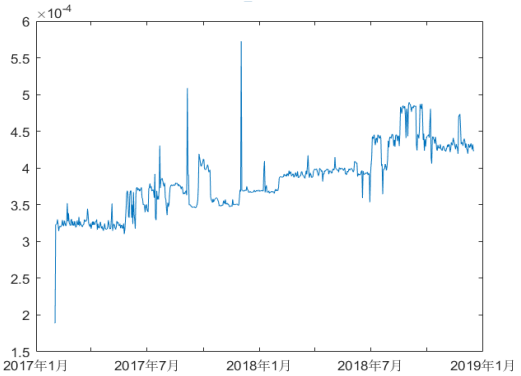
抽取部分传感器数据进行可视化，如下所示：



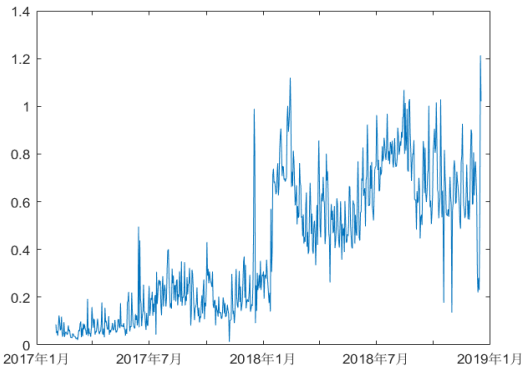
(a)第 3000 个传感器



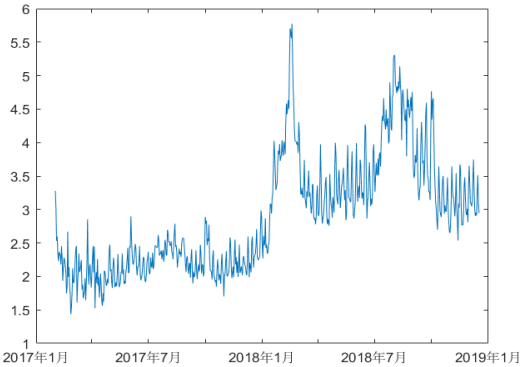
(b)第 6000 个传感器



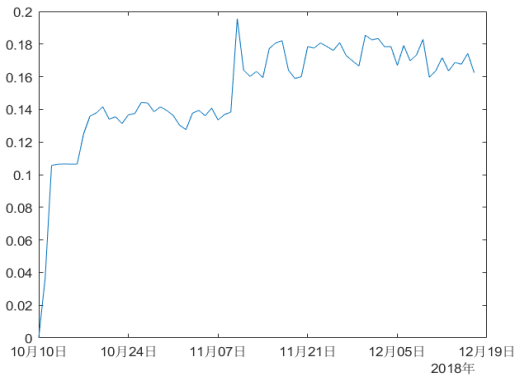
(c)第 9000 个传感器



(d)第 12000 个传感器



(e)第 15000 个传感器



(f)第 18000 个传感器

观察数据分布可以发现数据量有多有少，先分析数据量多的 acde 图，大致观察可以看出 2018 年 1-3 月数据分布有较大改变，但是改变之后围绕一定范围上下波动趋于稳定，因

而可以考虑移动平均法进行预测，可变参数主要有移动平均数、阶数，暂不考虑其它情况，再看数据量不足的传感器，如图 b，若也以移动平均法则会受较大影响，同时若移动平均数高于数据量则无法运行，因此将数据量少于某一阈值的传感器预测值统一设置为已知数据的平均值，可调参数有数据量的阈值。

由于每个传感器的数据都有各自的模型，因而需要独立确定最佳模型，即最佳移动平均数，考虑自动选取，第一个模型先把最佳移动平均数范围设置为 2 到 20，在训练集中每个传感器循环训练模型已知数据，找到使得历史数据平均平方误差最小的移动平均数，设置为该传感器对应移动平均值，再对测试集数据进行预测。

三、编写代码

模型实现采用 MATLAB 编码，首先读取数据，根据测试集传感器 ID 顺序提取训练集中对应传感器数据，对其进行训练找出最优移动平均数，统计测试集对应传感器需要预测的数据量进行预测，然后记录预测数据，如此循环直至所有传感器都预测完成。

因为测试数据中有部分重复数据，可以先筛除重复数据再进行预测，然后回填重复数据，但是观察预测的数据发现再预测若干步之后预测数据趋于平稳甚至某一值不变，再观察测试集中重复数据对每个传感器来说影响有但是可以先忽略，这也是一个优化方向，暂且保留不作优化。

MATLAB 最终得到的预测数据保存在 value.txt 文档中，再利用 python 合并 test_step1.csv 和 value.txt 文档。

然后提交测评，得到反馈有目的优化模型。

四、优化模型

通过上面分析可知优化方向一个移动平均项数下限与上限，经过思考暂且步优化平均项数上限，一个是判定数据量不足以进行移动平均的阈值，经过 11 次调试优化得到较优结果，相关参数如下表所示:

优化代数	移动平均数下限	判定阈值	数据量过少传感器个数	测评得分
1	2	5	32	98.573
2	3	10	70	99.617
3	4	15	109	99.895
4	4	20	127	99.814
5	5	15	109	99.896
6	5	13	101	99.903
7	5	11	81	99.939
8	5	9	59	99.989
9	5	7	53	99.991
10	5	6	48	99.995
11	4	6	48	100

表 1：模型优化过程

从表中优化过程即可看出优化方向以及较优超参及参数值，因为第一阶段有百分之 20 容错率，故是否为最优参数仍需要进一步考量。