

---

# 数据分析与数据挖掘第三次作业第一大题 – d问分析文档

数据分析与数据挖掘

---

DAM COURSE, SPRING 2018

BY

1552674 李 源



同济大学  
TONGJI UNIVERSITY

*Tongji University*  
*School of Software Engineering*

## 1 代码运行结果

这里我参照了文档要求，选择了预测结果转为绝对经纬度后的中位误差作为评价指标。对于 top k-各自的 MS 主基站的具体结果，在修正前后的情况如下：

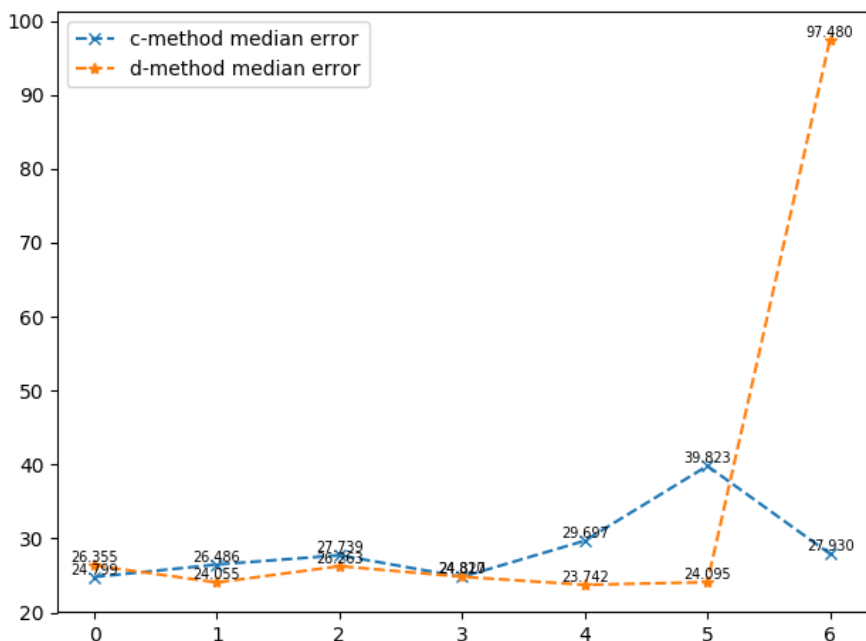


Figure 1.1: 不同的 MS 主基站修正前后中位误差分布

我将所有的 top k- 的 MS 基站的点重新合并到一起，做出了总体的 CDF 曲线，以及与未修正的 CDF 曲线比较如下：

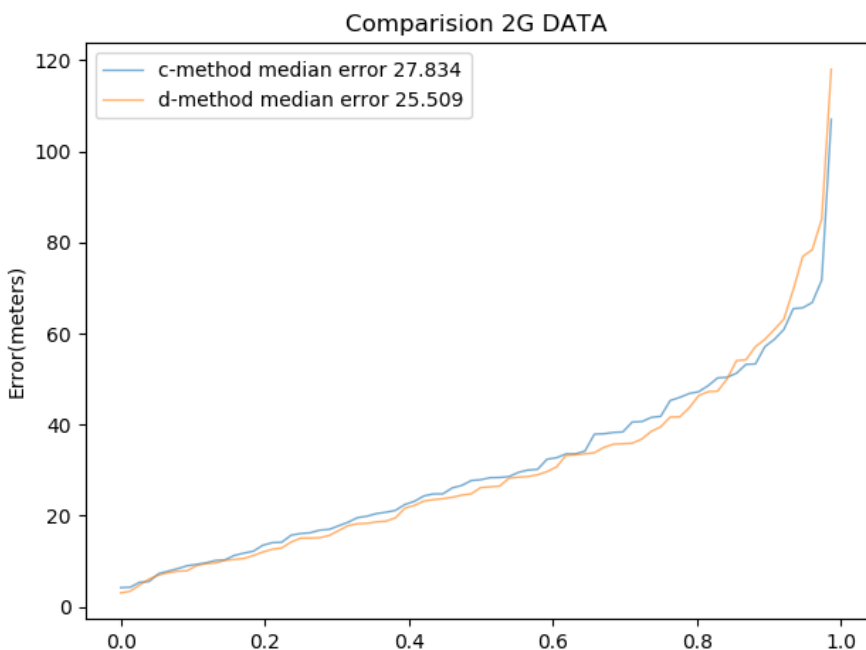


Figure 1.2: 修正与未修正的 CDF 曲线

## 2 分析讨论

这里主要对 d 问的修正方法和 c 问和 d 问的差异进行分析。

### 2.1 d 问的修正方法

d 问的修正办法在于，首先将所有的 MS 基站按照 c 问得到的中位误差结果进行排序，取得中位误差最小的前 top-k 分组（记为 topk+）和中位误差最大的后 top-k 分组（记为 topk-），尝试利用 topk+ 分组中的 MR 数据融入到 topk- 分组中。（这一部分是文档的说明）

这里主要需要考虑的是，如何将数据融入到 topk- 的分组当中。因为我在 c 问当中发现，训练效果较好的部分 MS 基站，其数据集具有较高的重复性，因此我没有简单地将所有数据直接与 topk- 分组的数据合并，而是对 topk+ 中的数据做了一定程度上的筛选，删除了过度重复的数据。

### 2.2 c 问和 d 问的差异

参考合并了所有点之后的总体中位误差，d 问的中位误差为 25.509m，稍好于 27.834m。但是如果参考到了每一个 MS 基站的情况，则会发现有一个基站的误差出现了极端的异常情况。因此我将 topk- 分组中的 MS 基站的原始数据分布和融入了的 topk+ 分组中的数据分布画了出来，进行分析。

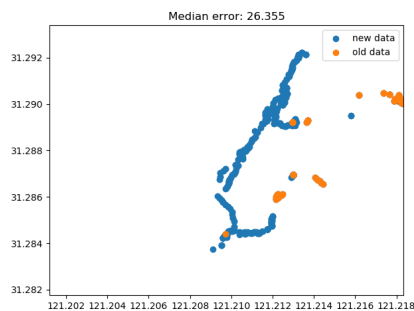


Figure 2.1: 基站1

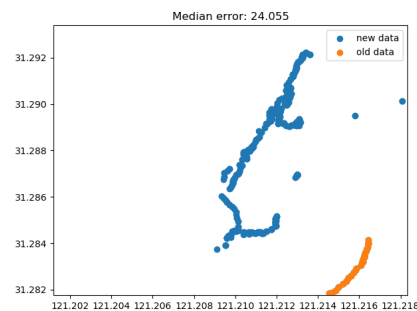


Figure 2.2: 基站2

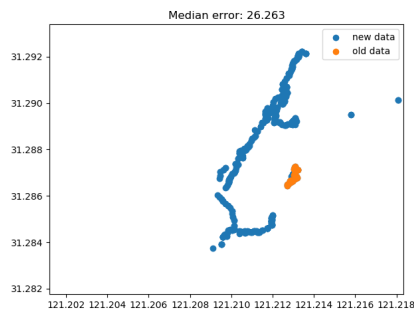


Figure 2.3: 基站3

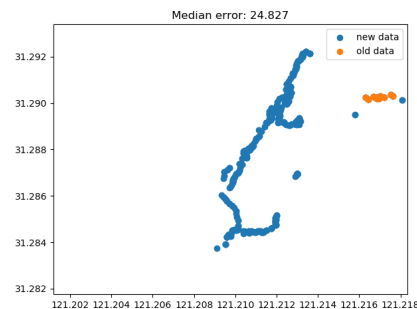


Figure 2.4: 基站4

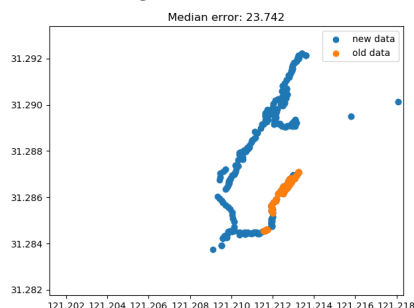


Figure 2.5: 基站5

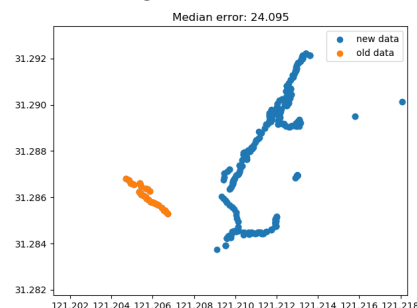


Figure 2.6: 基站6

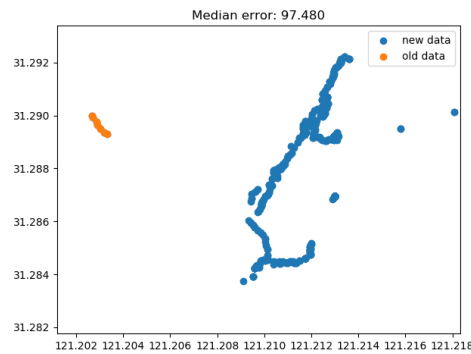


Figure 2.7: 基站7

可以很明显地发现，d 问在合并了数据集后面面临的一个问题，还是数据集的分布是否合理。形如基站 5 和基站 6 的情况，其合并后的数据集和原始数据集分布情况较为相近，因此取得了较好的结果（基站 5 原始的中位误差为  $26.697m$ ，而基站 6 为  $39.823m$ ）。

而对于基站 1 这种情况，由于其原始点分布较分散，虽然加入了新的数据集，但是与原始数据集的分布仍然不是十分相近，所以修正效果不好。再看基站 7 的情况，由于原始数据集和新的数据集两者之间相差太大，所以导致了中位误差从  $27.930m$  上升到了  $97.480m$ 。

由此可见，d 问虽然可能由于提升了数据集的数量，在一定程度上能够提高预测的准确性，但是由于新旧数据集的相似性不一定相近，所以总的来说提升效果并不会特别大。当然还有可能存在两者分布差异太大而导致中位误差不降反升的情况。

总的来说，d 问能够一定程度提高预测的准确性，但仍然需要考虑数据集的分布情况，还不够好。e 问应该就是考虑解决这个问题。

### 3 性能比较

两种方法的耗时如下，d 问因为需要再次预测，所以时间会相较于 c 问更久，但是 d 问并没有取得较大的改进，因此时间的增加并没有取得好的效果。

方法	时间
d 问	15.37s
c 问	9.42s