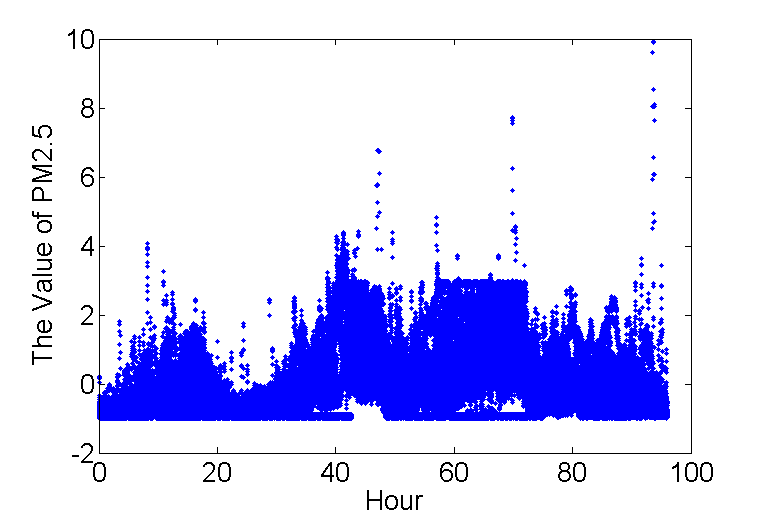
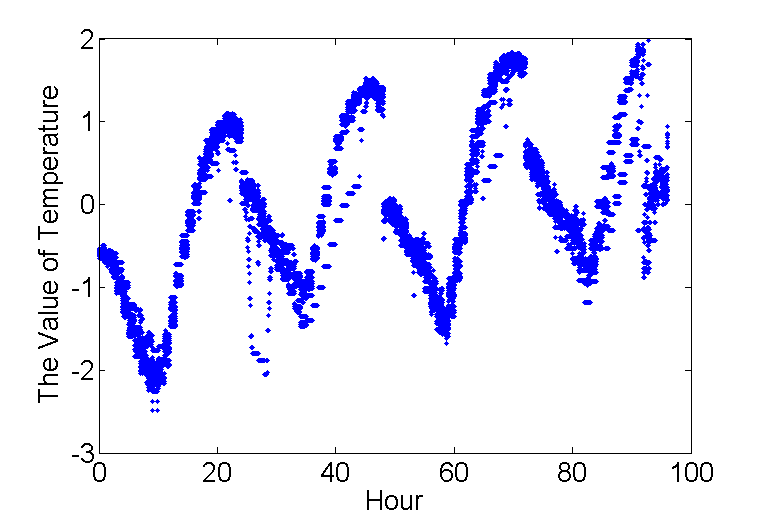
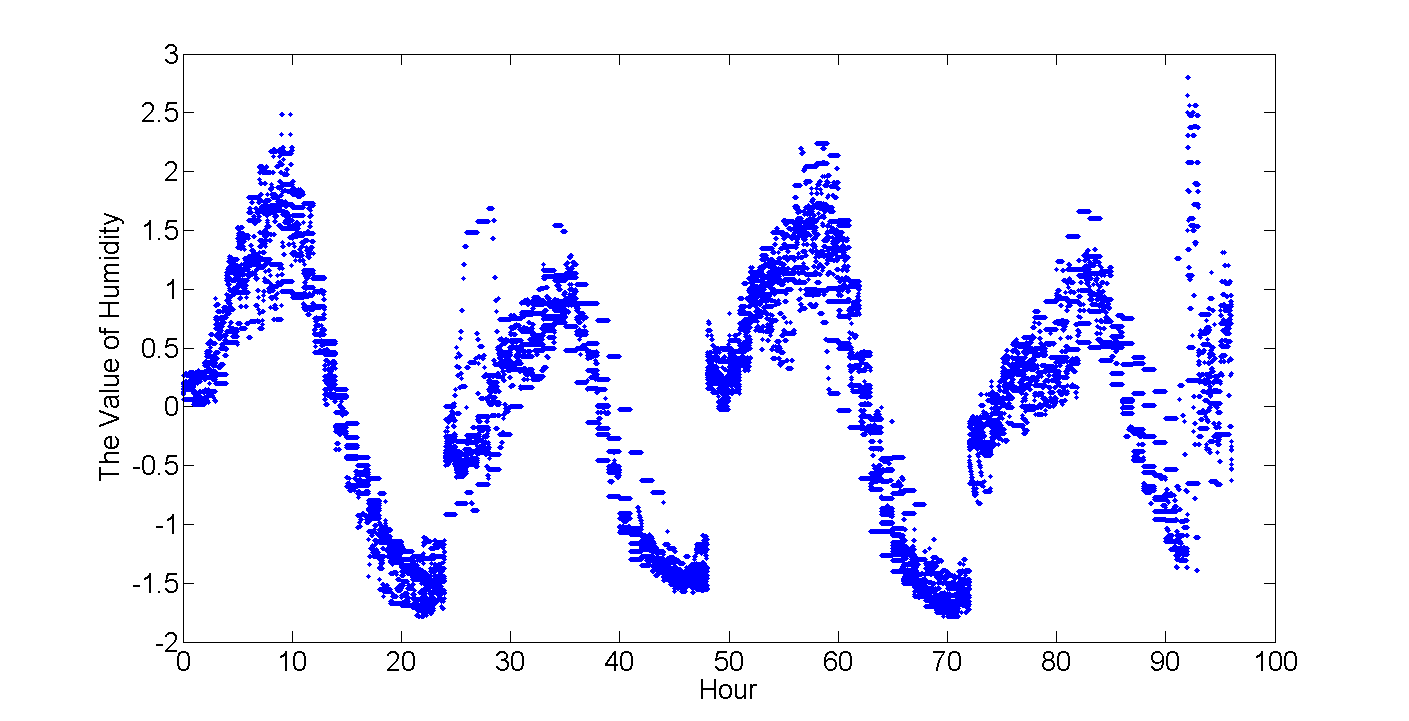
数据处理：

Z\_score方法归一化数据

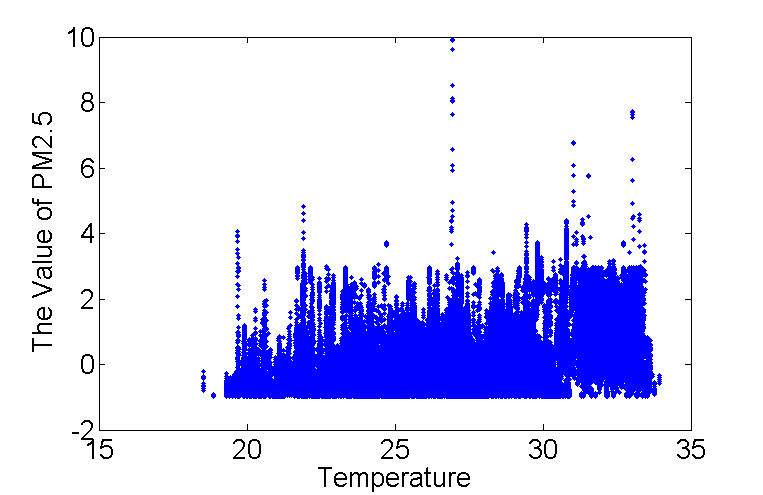
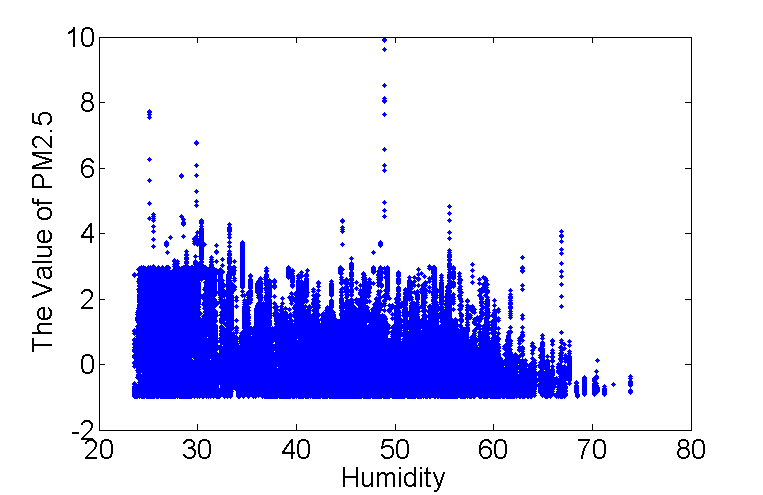
可以看出总体数据集，PM2.5在时间轴上具有周期性，4天的数据，具有四个峰值。同样压强和湿度等属性在时间轴上也具有明显的周期性。

从对比图可以看出PM2.5的噪声非常多。





而PM25在其他属性上周期性并不明显。



Lab1 : 验证 KD树一层划分结果确实保证了数据在天气和空间上的相邻相似性

通过查看划分结果的PM2.5方差来判定

从数据量5000、10000、20000、50000进行实验

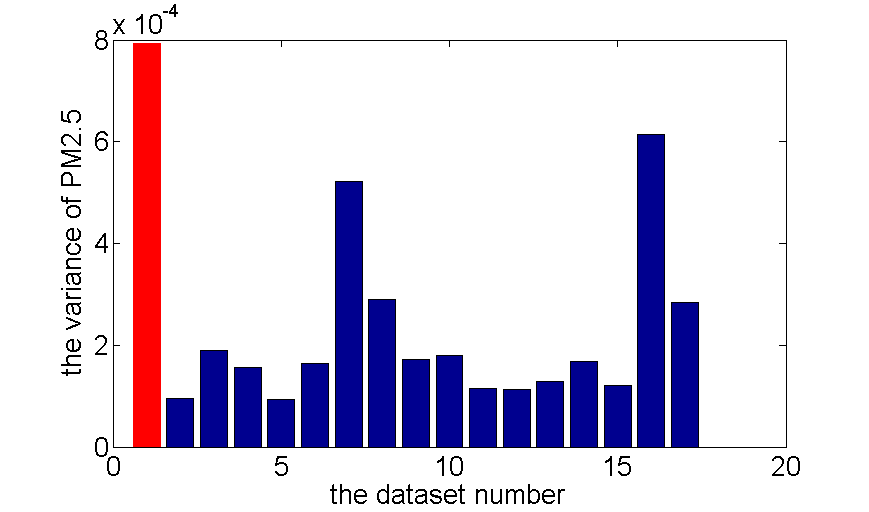
红色柱子表示划分前数据集的方差

蓝色柱子表示划分后每个划分集合的PM2.5方差

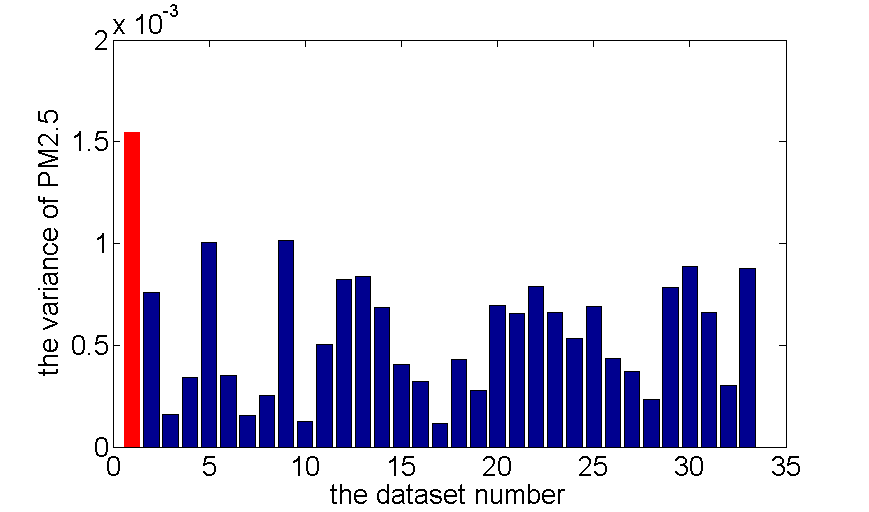
从图可以看出，划分以后数据集的PM2.5方差确实得到明显的下降

方差越小，训练误差越小

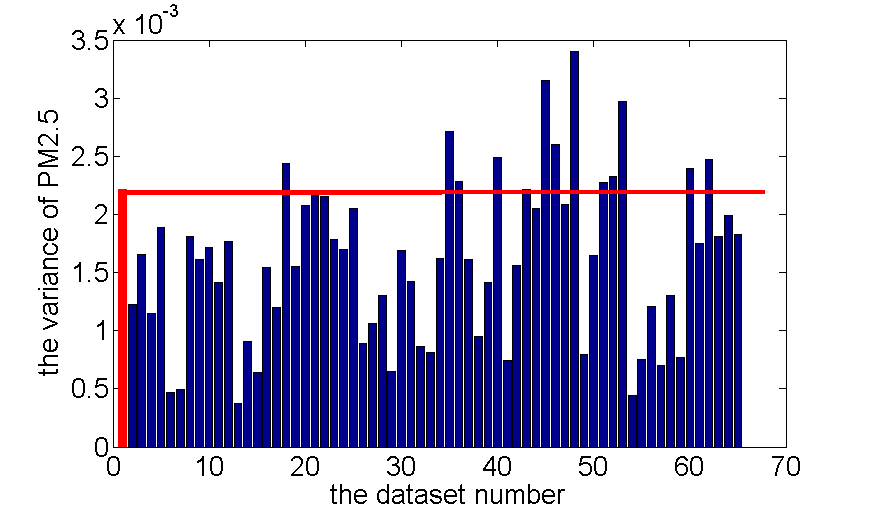
1. 5000



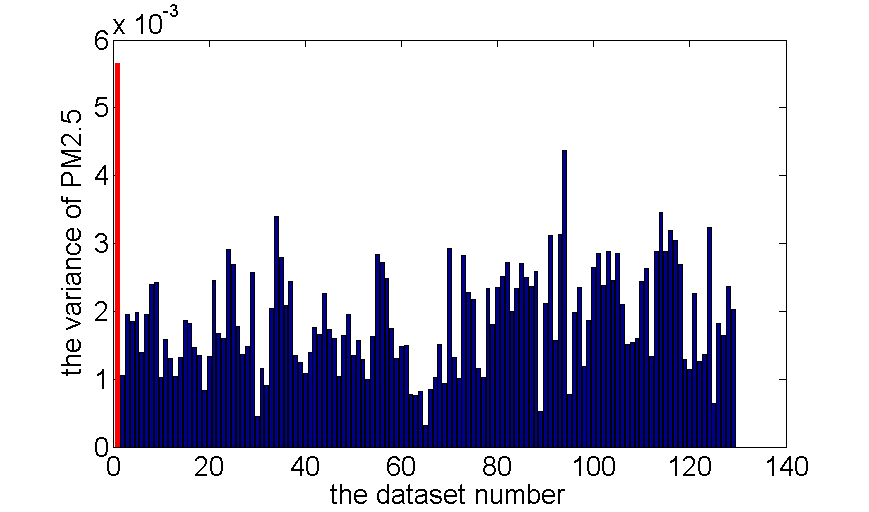
1. 10000



1. 20000



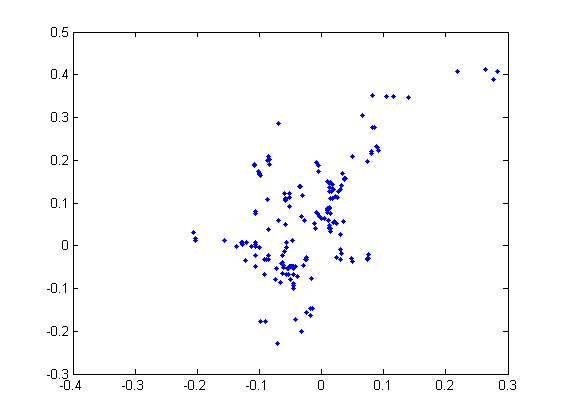
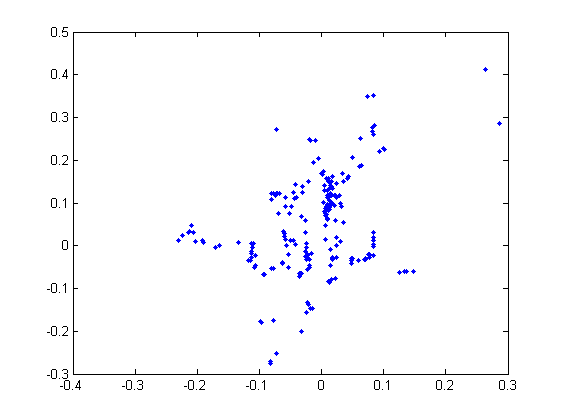
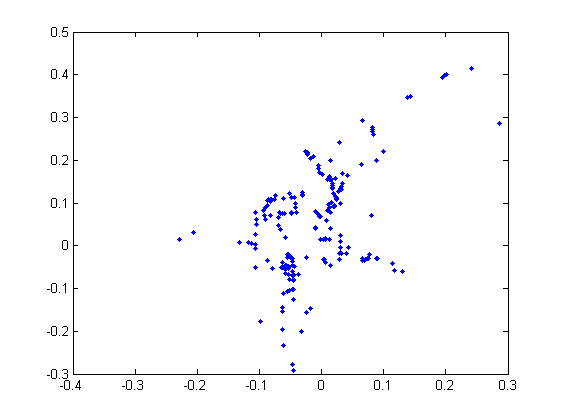
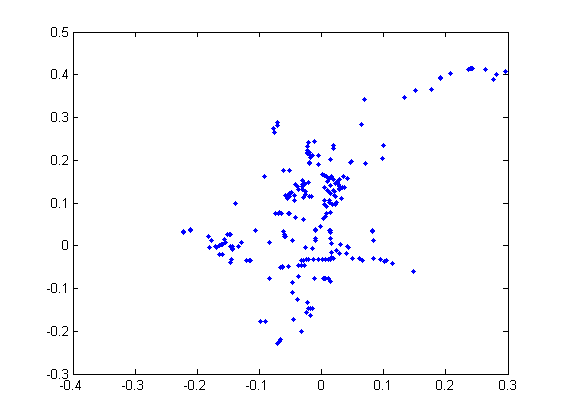
1. 50000



Lab2: 按照经纬度和天气划分得到的相邻相似数据集，在空间分布上并不完全相邻，而是散乱分布的。

所以

以下图为经纬度分布图



Lab2：

理想状况下，对数据进行一层聚类以后，就可以开始二层聚类，按照频率分析的思路，用多个正弦波去拟合子数据集（时间-PM2.5），但由于数据分布呈现如下状态：及在时间上分布不均匀，数据缺失严重，难以学出波形。

模型代码：

1. hyp\_model = initial\_model()

初始化：Q, size\_limit

1. [model, train\_time, train\_error] = train\_model (hyp\_model. trainingSet\_x, trainingSet\_y)

Model里面的参数：

流程：

tree = kd\_buildtree(trainingSet, 0);

nodeset\_cell = generateSubset(tree, 1, layer\_std); % 划分集合

预设超参数

Hyp\_subset(i) = minimize(); % 为每一个集合训练子模型

存时间、存超参数

返回

1. [test\_result, test\_time] = test\_model(model, testingSet\_x)

流程：

For 每一个测试点

Nearest\_id = findNearest() % 可能需要修改一下代码，因为已经用了7个属性划分

For 每一个自学习器

计算预测结果和权重;

End

加权和；

End

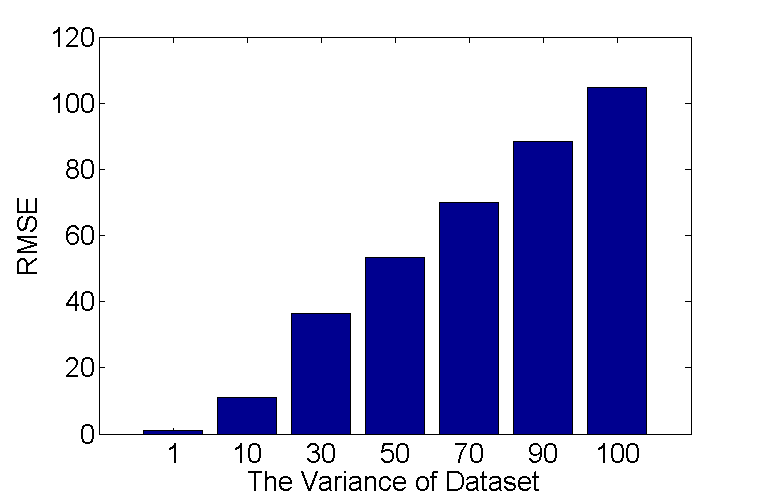
【实验】数据集方差越小，数据集质量越高

位置：lab1\var\_vs\_quality.m

对比实验数据：N=500，D=7（数据维度）

Imean=20

生成均值为20，方差不同的数据集



【实验】两次划分，方差减小

【实验】特征选择实验

用单核跑，输入不同特征

|  |  |
| --- | --- |
| 特征 | RMSE |
| 时间 | 0.4290 |
| 经纬度 | 0.3096 |
| 天气 | 0.3535 |
| 时间+经纬度 | 0.3513 |
| 时间+天气 | 0.3660 |
| 天气+经纬度 | 0.3226 |
| 时间+天气+经纬度 | 0.3491 |

【实验】模型对比

【实验】训练时间

【实验】测试时间对比

【实验】自适应

**实验部分：**

实验之前的部分--【数据分析论证】：

Lab1 ： PM2.5数据在时间上具有明显的周期性

Lab2 ：天气模式和经纬度越近的数据具有相邻相似性

Lab3 ：K值选择实验

【模型性能对比】

Lab4：不同模型RMSE对比

Lab5：不同模型训练时间对比

Lab6：不同模型测试时间对比

Lab7：不同模型自适性对比

【模型内部实验】

--证明核有效

Lab8：用不同核作为模型的输入

Lab9：不同划分的RMSE值，和训练时间对比