数据特征：

北京市移动监测系统，51辆出租车，监测PM2.5浓度，每30s监测一次

1. 采集数据量大
2. 分布不均

challenges：

1. 大数据量和实时性之间的矛盾
2. 实时框架的自适应性
3. 监测数据分布的时空不均衡性
4. 如何充分用历史信息

为了改善时间性能，利用KD-tree划分同质数据集，再基于数据集学习每一个子分类器（**实验证明KD树划分出来的数据集就是同质的**）

为了加强模型的自适应性，提出两层时间框架

解决数据分布不均：一个区域的污染物来源有二：

1. 积累效应：由自身的土地使用属性决定
2. 传播效应：由大气环流决定

通过扩大时间窗口可以使数据分布均匀，同时采用数据筛选机制，降低数据量

主要贡献：

1. 提出了一个Real-time Ensemble Estimation Model(REEM) based on Gaussian Process Regression
2. 引入两类核函数去解决时间周期性和相邻相似性
3. 利用数据筛选机制选出最适合当前查询点的数据
4. 提出两层时间框架去解决自适应问题

模型：

1. partition and training based on KD-Tree

利用距离特征、天气特征划分KD-Tree。

\kappa 选择实验

1. individual models training

根据数据相邻相似性和周期性，我们选定了两种核

1. Data filter and estimation

KD树只能保证相邻相似性而不能保证周期性（**实验证明：同一个结点内的数据确实具有周期性）**

根据“距离”＝相位差时，选择出相位差小于某个距离的所有样本

1. 两层自适应框架

T\_l: 设定的阈值

T\_s: 受估计误差的影响

T\_l: 时间窗口用来重新训练超参数和构造KD-Tree

实验：

注意：不可再选择连续的数据进行试验测试

**特征选择实验**

自适应实验窍门：我们的模型始终只预测T\_l 以内的数据，而对比模型需要预测T\_l以外的数据，即预测0~T\_l, T\_l~2T\_l,我们训练两个我们的模型做测试，别的对比模型只用一个。