关键词：实时空气质量预测 用浮动车数据

In recent years, air quality has attracted increasingly attention from all over the world. High levels of surface-level air pollutions are responsible for a range of disease. Nowadays, air quality monitor stations have been built to inform people about air quality every hour in many cities. However, it is so expensive to establish and maintain air quality monitor stations which only provide sparse information about the spatial distribution of air pollutants. Due to the sparsity of observation from static stations, many existing methods are good at coarse-grained estimation not fine-grained on time, usually a week, a month or even a year[][]. While, compared to coarse-grained estimation, real-time estimation is much guiding significance to people’s daily life to some extent[].

浮动车数据的特点：

1. 成本低，易维护 (相对于传统静态基站)
2. 采样频率大，导致数据量大，这是相对于静态基站的优点，但也成为应用到实时应用上的局限。
3. 时空分布不均匀

With the prevalence of mobile sensor which is available and affordable compared with traditional monitor station, there is an emerging way to monitor the air quality by the mobile sensor, which can provide abundant and fine-grained monitoring data. In this paper, the experiment dataset is from mobile sensor system of Beijing, China which consists of 15 taxis equipped with mobile sensor to monitor the concentration of PM2.5\ref{}. The monitoring frequency is once every 30 seconds, which leads to the generation of about 35 thousands monitoring data every day. Especially the spatio-temporal distribution of monitoring positions is random depending on where the taxi wants to go which can be seen in Fig.x. The abundant dataset gives us an opportunity to do real-time air quality estimation. In this paper, we propose a real-time estimation model to infer air quality at given query points.

Challenge：

实时空气质量估计对模型的计算效率提出了要求，同时还要求模型具有自适应更新能力，要做实时的空气质量插值with observation from mobile sensor, 一个最突出的矛盾就是浮动车的巨大数据量与实时计算对时间性能的要求之间的矛盾。因此我们的挑战是充分利用数据信息，同时达到实时要求，同时模型还需具有较高适应性。

如图所示，浮动采样具有时空分布不均匀的特性，如果利用实时信息做实时估计，对于监测数据稀少的区域，很难得到可靠的估计结果。因此我们的挑战是如何利用有效的历史信息.

There are some challenges to do real-time estimation with monitoring data from mobile sensors. The one is the contradiction between the huge amount of data and the high demand time performance of real-time application. The second challenge is how to improve the adaptivity of real-time estimation model. The third challenge can be seen in Fig.x, which is the spatio-temporal distribution of monitoring data is random. If we just use latest observation to do real-time estimate, it’s difficult to get reliable estimation on areas with sparse observations. So how to make full use of historical information and diffusion of air pollutant is a challenge.

。

Solution:（在这一段要说明我们提出的模型的名称）

为了解决数据量与计算时间之间的矛盾，我们利用增强版KD树将原始数据集组织成多个同质数据集合，并对每一个同质数据集合训练一个子学习器，引入集成学习的思想对子学习器的计算结果进行集成得到最终的结果。为了加强在线模型的自适应性，我们提出了two-layer framework based on time slices实时更新模型。一个地区的污染物主要来源于本地属性造成和空间传播，我们充分利用污染物这样的空间传播和历史信息。为了解决时空分布不均匀的问题，结合污染物的产生原理，提出一种新的数据采样机制，为目标插值点(query point)的预测筛选出有效的训练样本。

To improve the time performance, we apply enhanced KD-tree to organize the original data into multiple homogeneous data sets and train an individual learner for each homogeneous data set. And then the estimations from these individual learners can be integrated based on ensemble learning. In order to enhance the adaptability of model, we propose a two-layer framework based on time slices. Besides, the pollution of a region depends on its own land use and dispersion from other places. To solve the problem of non-uniform spatio-temporal distribution of data, a new data sampling mechanism is proposed to select more related data for the query point, which combines the principle of pollutants generation.

总结贡献如下

1. 我们提出了一个Real time Ensemble Estimation Model(REEM) based on Gaussian Process Regression, which takes advantage of air quality observations from mobile sensors located on the car.
2. 核函数
3. 筛选机制
4. Two-layer framework
5. 实验

The main contribution of this paper can be summarized as follows:

1. We present a real-time Ensemble Estimation Model (REEM) based on Gaussian Process Regression, which takes advantage of air quality observations from mobile sensors located on the taxis
2. We introduce a kernel function to discovery the pattern of the spatio-temporal, climate and pollutions features.
3. We propose a selection mechanism to obtain more related observations for a given query point.
4. We present a two-layer framework ensemble learning model based on time slice and apply KD-tree to improve adaptivity and computational efficiency.
5. Experiments show that our proposed method outperforms the traditional spatial regression method in both precision and time cost.

The rest of the paper is organized as follows: we next conclude related work. In section of preliminaries, we define our problems and review the Gaussian Process Regression. We describe the technical details of our approach in section 4. In section 5, the results of experiments on real-world data can be found, and we also evaluate it against state-of-art approaches. Finally, we conclude with a brief summary and discuss future work in section 6.

Data Filtering Mechanism (数据筛选机制) filtering rules

1. 什么是Data Filtering Mechanis

我们通常基于大数据去挖掘数据本身的模式，但是大数据量通常意味着数据存在一定程度的冗余。并且由于the drawback in GPR is the training computational cost with cubical scale of the number of training points. 因此我们定义一种data filtering mechanism to remove data from a huge amount of training data that is less correlated to the query points，从而为每个query point构造有个最优子训练集，从而提升高斯过程回归的精度。

1. 结合图说明 Filtering Mechanism对我们问题的适用性

各个采样时段的移动监测数据具有的时空分布不均匀性 which can been seen in Fig.x，所以如果只用latest data 去做real-time eatimation, 对于位于监测数据sparse areas的query points不能得到a reliable estimation。 为了克服数据时空分布不均匀性给模型带来的不稳定性，我们充分利用历史信息以扩大监测数据的时空监测范围，学习历史信息中的pattern以适应监测数据的各种分布。为了提高计算效率，我们introduce a filtering mechanism to refine training data set for each query point.

以上通过enhanced KD-tree的划分使得空间距离和天气属性相近的数据被划分到了同一individual learners，但是时间属性并没有被纳入到划分标准内，这是因为KD树保证的相邻相似性不能刻画时间feature的周期性。

因此我们在KD-tree的划分结果上应用filtering mechanism挖掘时间的周期性，this can be achieved by defining the filtering rule as follow:

(公式 1)

(公式 2)

K表示与时间相关的核，而k(t\_latest, k\_old)能反映出两个采样数据的在时间属性上的相关性， 由我们的核函数（公式x）可知，K作为一个乘子，如果K小，则最终的核函数值也会非常小，即两个样本之间的相关性弱。因此我们set up a threshold and remove all data which fall below this threshold.

Ensembling and prediction.

介绍污染物的生成方式：累积效应（cumulative effect）+传播效应(propogation effect)， 累积效应通常是由一个region的土地属性决定的，因此存在自身的规律性，通常反应在时间的周期性上；而传播效应对应了同一时间维度上污染物的传播，通常污染物的传播是天气驱动的同时也与空间距离有关，即具有极强的相邻相似性。

The extent of pollution in a region is determined by the cumulative effect and the propagation effect. Cumulative effect

基于KD树构建的estimation model保证了每个individual learner对应的数据集在天气和空间属性上的相邻相似性， 所以对于一个新的查询点，基于KD-tree查询到离它最近的individual learner。对该individual learner对应的数据集应用Data filtering mechanism to refine一个most correlated的数据集 as estimation data set, 这反应了积累效应。对于模型的其他individual learner, 我们选取latest monitoring data作为estimation data set，这反应了传播效应. (见Algorithm. x)

Section 4: OUTLINE

In this section, we presents a Real-time Ensemble Estimation Model(REEM) for air quality estimation. 首先，我们提出了基于KD树的数据organize方法，以及基于gaussian process的individual learner（子学习器） 训练方法。然后，我们提出了基于频率分析的data filter mechanism 和ensemble estimation(集成学习) 方法。最后我们给出了在实时环境下的two-layer self-adaptivity framework.

实验部分：

After training, 在estimation阶段我们首先对数据进行采样，然后集成学习得到最终的estimation

* K selection

在K的选择实验中，我们选取10000训练集， 3000交叉验证集（CV）。实验结果显示， 随着K的增大，训练时间与K正相关，这是由于K决定了每个子学习器的样本量，训练时间是O（n^3）with n的训练集大小。而RMSE先减小后增加，这是因为子学习器需要保证一定的训练样本量才有效果，但当样本量过大时，划分到每个子学习器中的样本集同质性减弱，导致学习效果下降。根据实验结果，Kappa选择9或10时效果较好。