核密度分析改进算法：Adaptive KDE

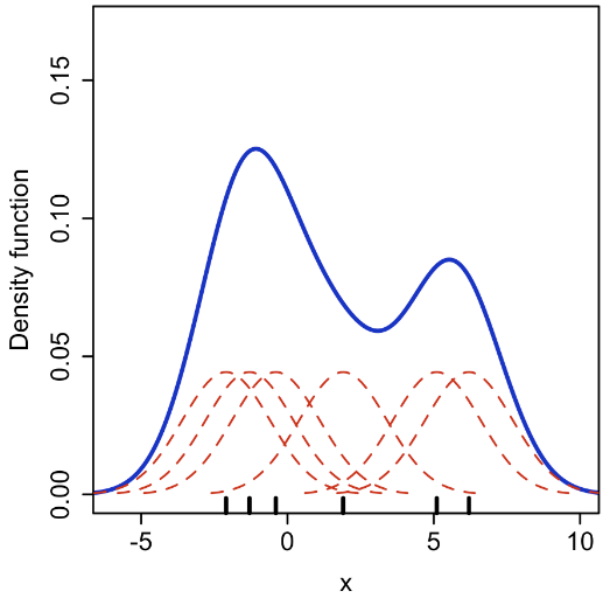
汇报人：钟代琪

2022.11.09

目录 Content

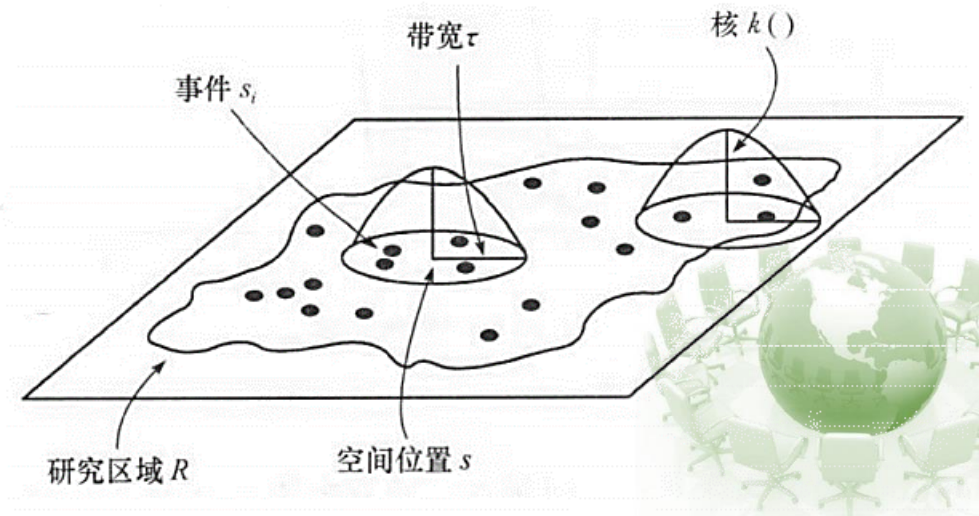
1 核密度分析原理

KDE核心思想：使用事件的空间密度分析表示空间点模式。点密集的区域事件发生的概率高，点稀疏的区域事件发生的概率低。



设$(x_1, x_2, \ldots, x_n)$为一组独立同分布的样本点，它的概率密度是f，采用一个核函数$K(\cdot)$估计

$$f(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

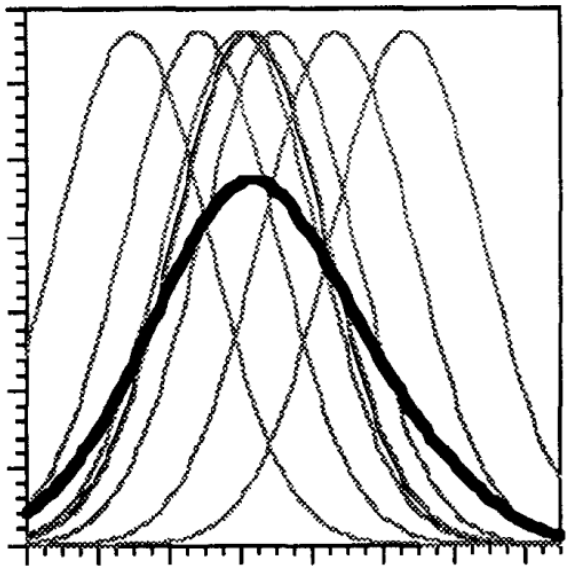


2 Adaptive KDE

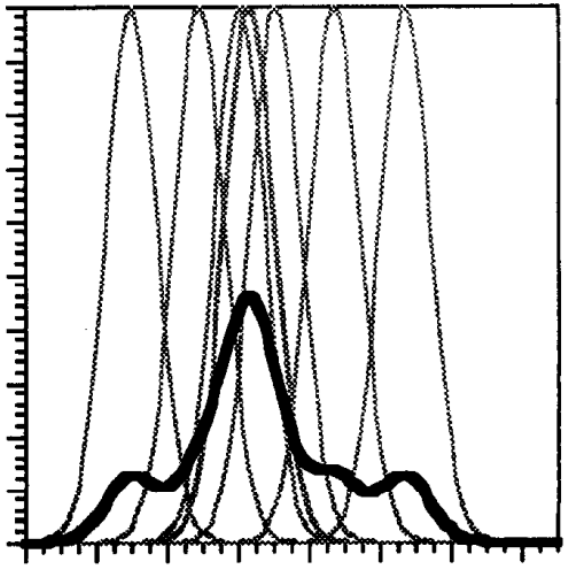
问题提出：Bandwidth?

过大：无论数据是何种模式，生成的核密度估计都与核函数一致，属于欠拟合

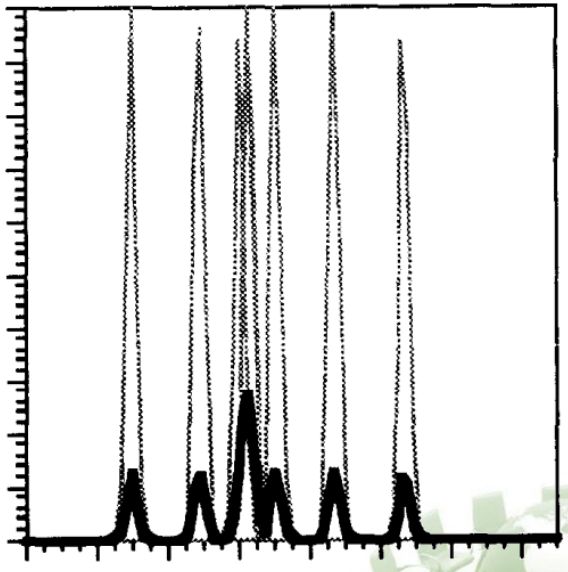
过小：生成的核密度估计是以数据点为中心的尖峰，属于过拟合



Too Large



Just Right



Too Small

2 Adaptive KDE

改进思路1：核密度 → 概率密度 → 最大似然估计

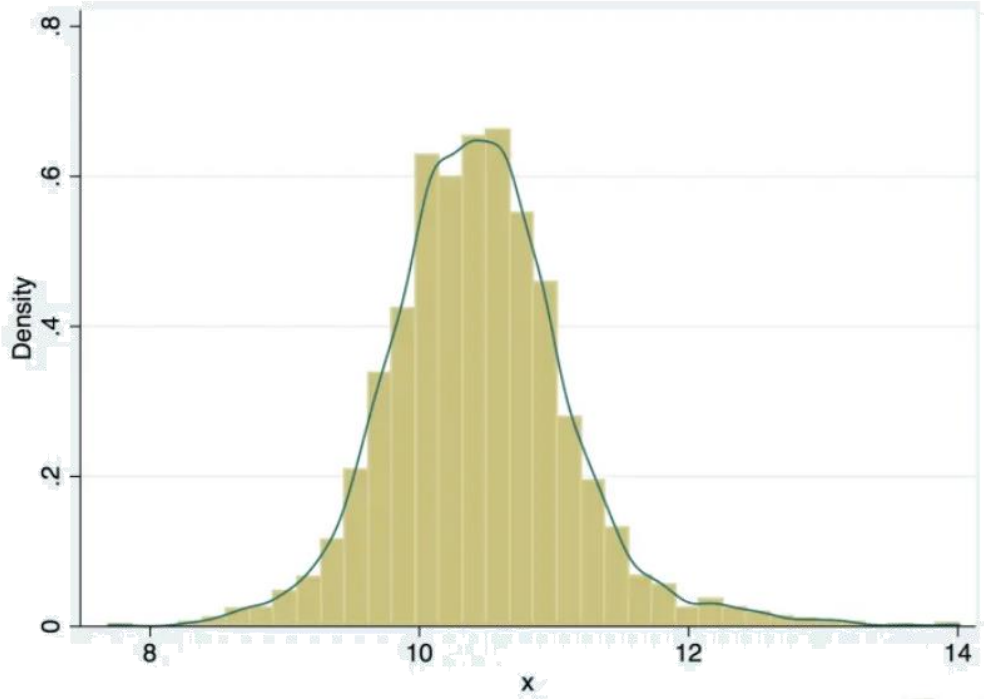
假设x服从参数为$\boldsymbol{a}$的某种分布，即

$$x \sim f(x;\boldsymbol{a})$$

同时，考虑h和a独立

$$x \sim f(x;h)$$

即**x服从概率密度为$f(x;h)$的分布**。



优化目标：

$$\max L(x;\boldsymbol{a}) = \max \prod_{i=1}^{n} f(x_i;h)$$

2 Adaptive KDE

改进思路1：核密度 → 概率密度 → 最大似然估计

$$f(x_i;h)=\underbrace{\frac{1}{nh}K(0)}_{x_i's\ kernel}+\underbrace{\sum_{j=1;j\neq i}^{n}\frac{1}{nh}K\left(\frac{x_i-x_j}{h}\right)}_{remaining\ kernels}$$

当$h\to 0, f\to\infty$，可以认为是由于数据点的自身映射导致似然函数在h=0时最大

一种可行的方法：去掉第一项，使用其余数据点估计该点的密度

$$f^{\{i\}}(x;h)=\sum_{j=1;j\neq i}^{n}K\left(\frac{x_i-x_j}{h}\right)$$

优化目标： $\max L(x;\boldsymbol{a})=\max\prod_{i=1}^{n}f^{\{i\}}(x;h)$

2 Adaptive KDE

改进思路2：固定带宽 → 可变带宽

对于密度高的区域，采用较小的h；对于密度低的区域，采用较大的h

$$f(x;h)=\frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{x-x_i}{h}\right)$$

对每一个x_i采用一个对应的h_i：

$$f(x;\boldsymbol{h})=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_i}K\left(\frac{x-x_i}{h_i}\right)$$

2 Adaptive KDE

核心问题：h_i如何确定?

① 计算合理的估计值h:

$$h = 1.06 \times \sigma(x) \times N^{-\frac{1}{5}}$$

Rule of Thumb

该式由假设数据服从正态分布推导而来。

Reference:
Silverman, B., 1986.Density estimation for statistic and data analysis.

② 根据估计值获得精确的h_i:

$$h_i = \lambda_i \cdot h = \left(\frac{f^*(x_i)}{g}\right)^{-\alpha} \cdot h$$

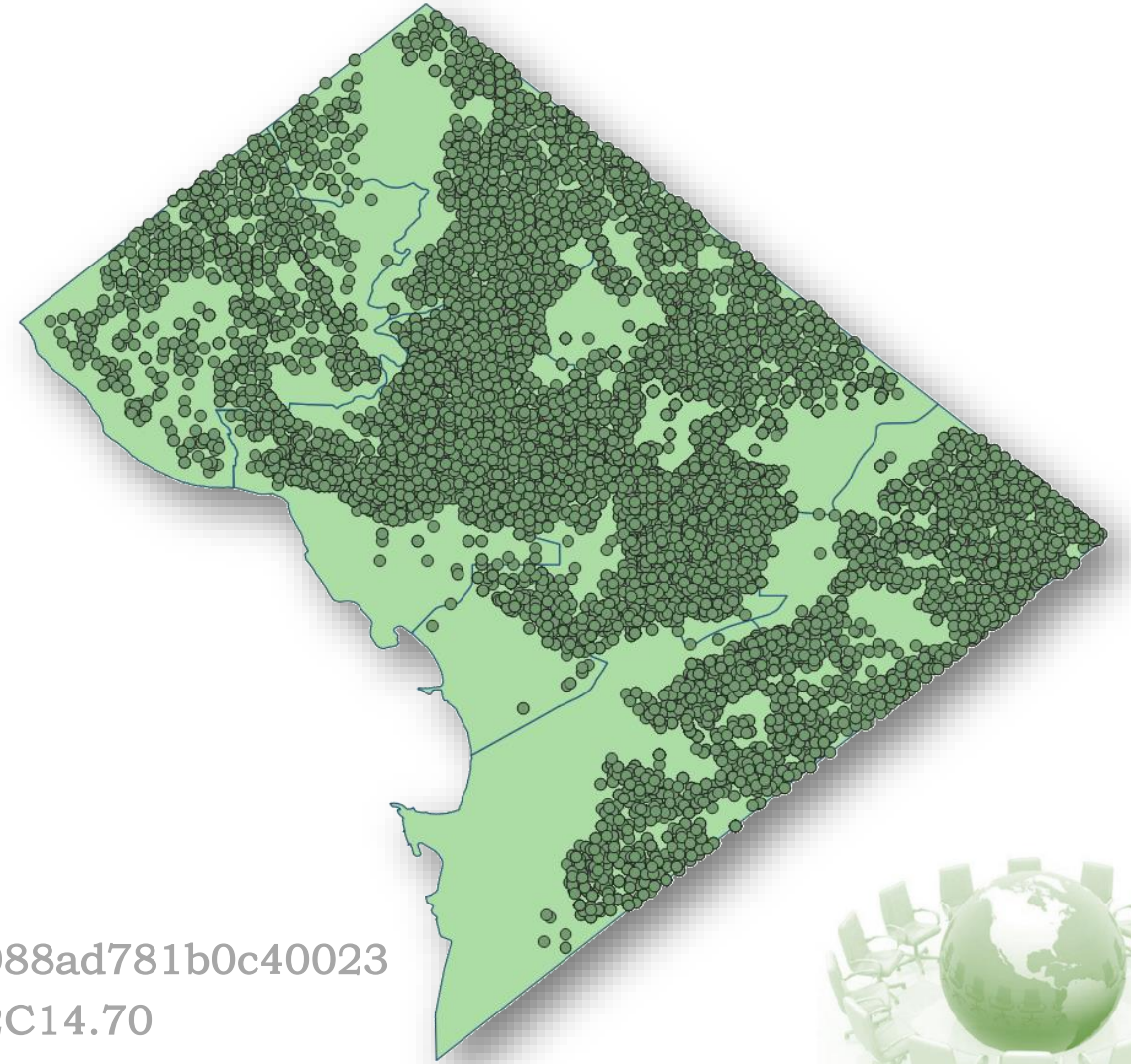
式中，λ_i为带宽因子，$f^*(x_i)$是根据估计值h计算得到的密度估计，g是$f^*(x_i)$的几何平均，$\alpha \in [0,1]$为灵敏因子。

3 实验与分析

数据集：DC crime incident 2020

CCN	PSA
REPORT_DATE	NEIGHBORHOOD_CLUSTER
SHIFT	BLOCK_GROUP
METHOD	CENSUS_TRACT
OFFENSE	VOTING_PRECINCT
BLOCK	LATITUDE
XBLOCK	LONGITUDE
YBLOCK	BID
WARD	START_DATE
ANC	END_DATE
DISTRICT	

https://opendata.dc.gov/datasets/f516e0dd7b614b088ad781b0c4002331_2/explore?location=38.910865%2C-77.020651%2C14.70



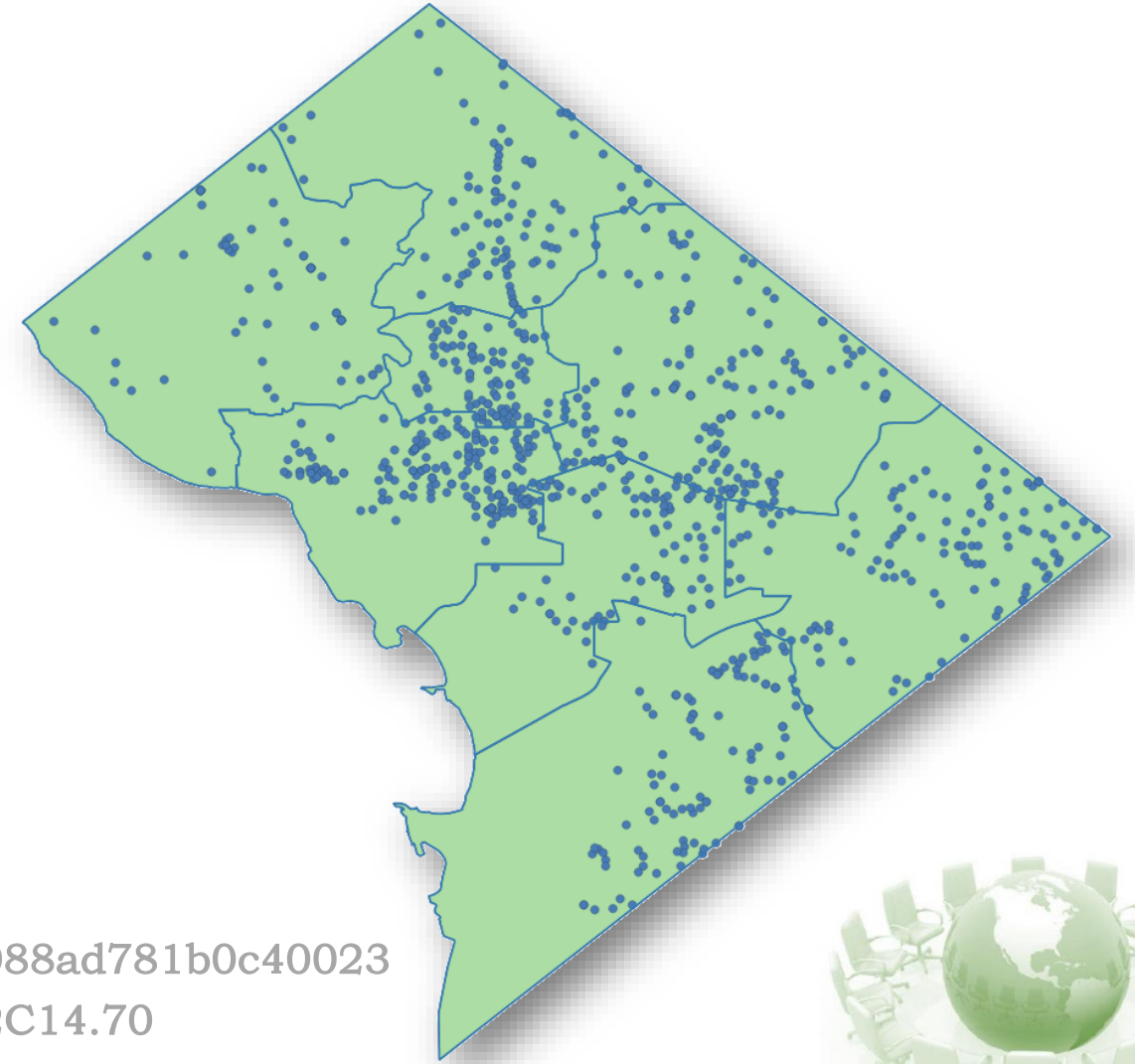
共27882条数据

3 实验与分析

数据集：DC crime incident 2020

CCN	PSA
REPORT_DATE	NEIGHBORHOOD_CLUSTER
SHIFT	BLOCK_GROUP
METHOD	CENSUS_TRACT
OFFENSE	VOTING_PRECINCT
BLOCK	LATITUDE
XBLOCK	LONGITUDE
YBLOCK	BID
WARD	START_DATE
ANC	END_DATE
DISTRICT	

https://opendata.dc.gov/datasets/f516e0dd7b614b088ad781b0c4002331_2/explore?location=38.910865%2C-77.020651%2C14.70



抽取前1000条数据

部分代码展示：

1、RoT及其变种 $h = 1.06 \times \min\{\sigma, \frac{IQR}{1.34}\} \times N^{-\frac{1}{5}}$

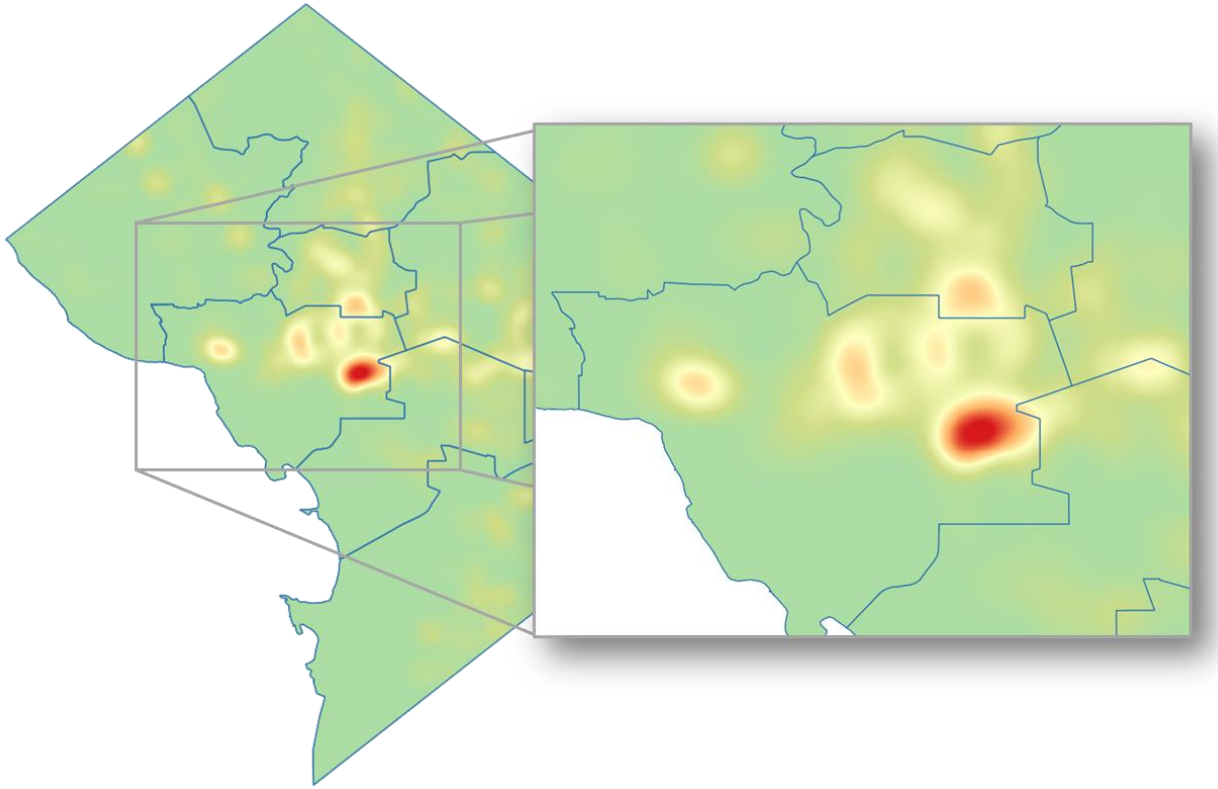
```
delta=min(dis_std,(q3-q1)/1.34)
h=1.06*delta*pow(len(x),-0.2)
print("thumb-h:",h)
```

2、Cross-Validation

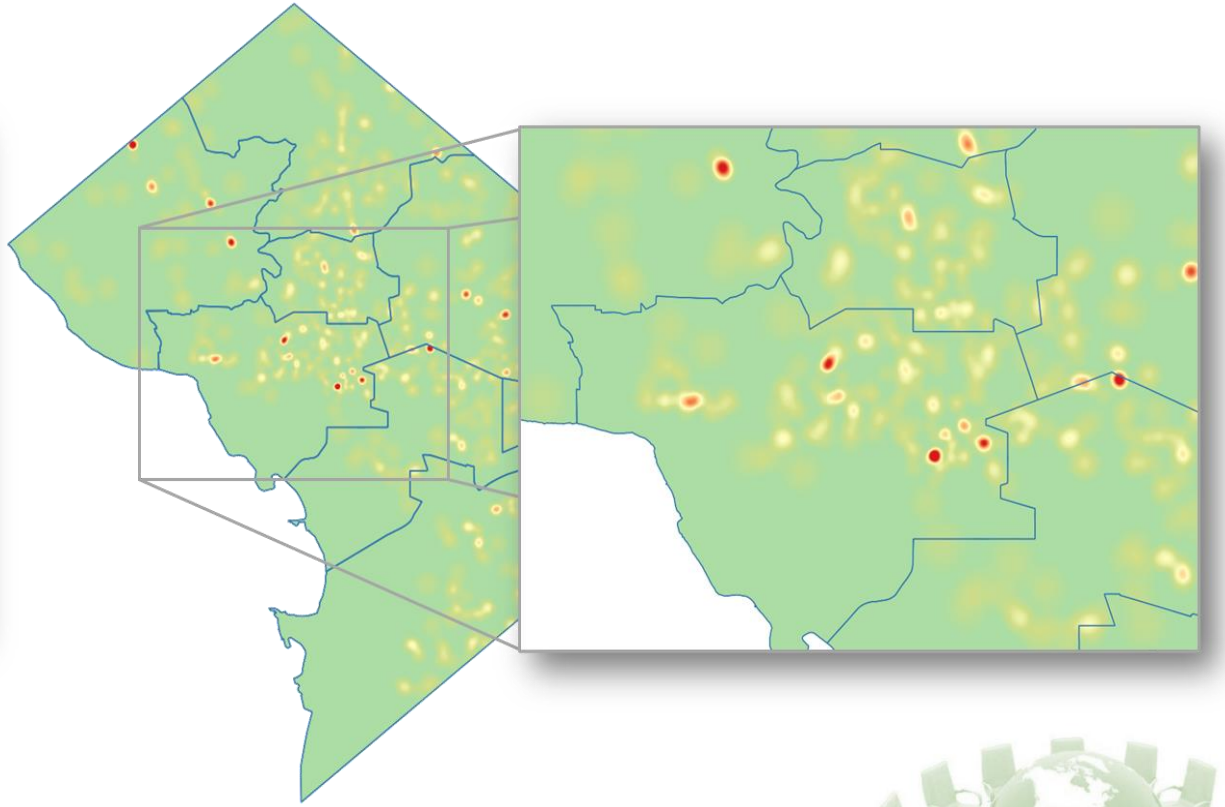
```
h_list=[h-delta_h,h,h+delta_h]
a_list=[a-delta_a,a,a+delta_a]
```

```
if  max_L==L[2]:
        delta_a=delta_a/2
        delta_h=delta_h/2
else:
        idx=L.index(max_L)
        if  idx==0:
                a=a-delta_a
                h=h-delta_h
        elif  idx==1:
                a=a-delta_a
        elif  idx==3:
                a=a+delta_a
        elif  idx==4:
                a  =  a  +  delta_a
                h  =  h  +  delta_h
```

3 实验与分析



Rule of Thumb



Adaptive bandwidth

3 实验与分析

分析：

1、如果数据并非服从正态分布，而是服从某种多峰分布，RoT会造成核密度估计的过分平滑，实际分布的细节会被掩盖。

2、Adaptive KDE利用了可变带宽克服了数据本身的分布带来的局部分布与全局分布不统一的问题。尤其是对于地理事件数据分布与地理位置有着较强的相关性，往往并不服从简单的高斯分布。

3、 Adaptive KDE时间复杂度$O(n^2)$，对于大型数据的实时处理应用不便。

Thanks for listening☺