第九讲: 连环犯罪凶手藏身地预测

数学模型和算法的应用与 MATLAB 实现

周吕文

中国科学院力学研究所

2017 年 7 月 22 日

微信公众号: 超级数学建模

Part I

题目与分析

2010MCM problem B: Criminology

In 1981 Peter Sutcliffe was convicted of thirteen murders and subjecting a number of other people to vicious attacks. One of the methods used to narrow the search for Mr. Sutcliffe was to find a "center of mass" of the locations of the attacks. In the end, the suspect happened to live in the same town predicted by this technique. Since that time, a number of more sophisticated techniques have been developed to determine the **"geographical profile"** of a suspected serial criminal based on the locations of the crimes.

Your team has been asked by a local police agency to develop a method to aid in their investigations of serial criminals. The approach that you develop should make use of at least two different schemes to generate a geographical profile. You should develop a technique to combine the results of the different schemes and generate a useful prediction for law enforcement officers. The prediction should provide some kind of estimate or guidance about possible locations of the next crime based on the time and locations of the past crime scenes. If you make use of any other evidence in your estimate, you must provide specific details about how you incorporate the extra information. Your method should also provide some kind of estimate about how reliable the estimate will be in a given situation, including appropriate warnings.

In addition to the required one-page summary, your report should include an additional two-page executive summary. The executive summary should provide a broad overview of the potential issues. It should provide an overview of your approach and describe situations when it is an appropriate tool and situations in which it is not an appropriate tool. The executive summary will be read by a chief of police and should include technical details appropriate to the intended audience.

2010MCM 问题 B: 犯罪学

在 1981 年, 彼得萨克利夫被判犯有十三起谋杀罪和一系列的恶意伤害罪. 在该案中, 一种用来缩小搜索萨克利夫先生所在范围的方法是找到这些犯罪地点发生的 "重心". 最后, 这个嫌疑犯恰好生活在用这种技术所预测的那个城镇里. 从那时起, 许多更复杂的技术被发展起来, 用来确定系列犯罪的嫌疑人位置的 **"地理轮廓"**.

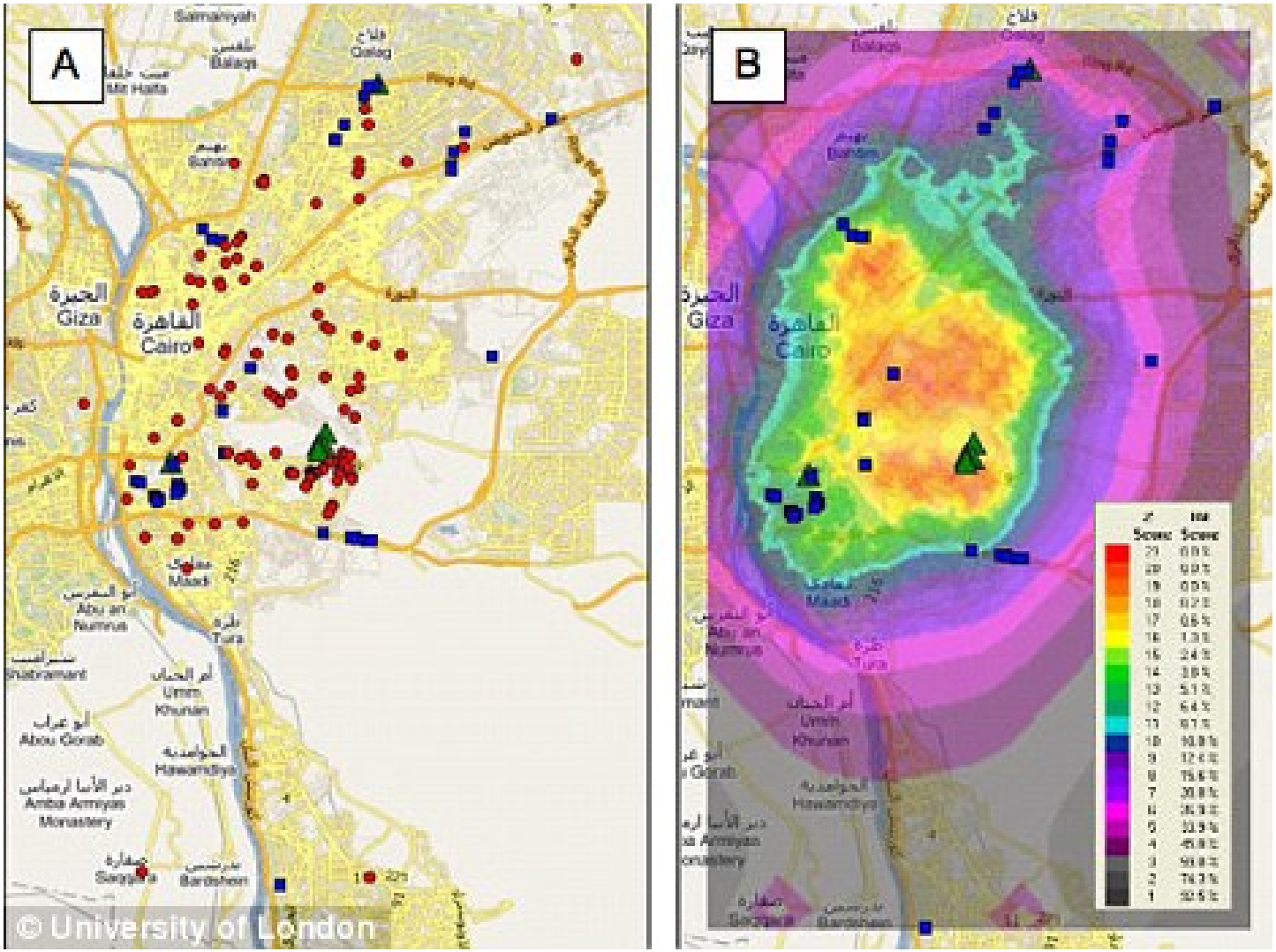
当地的警方要求你的团队开发一种方法来帮助他们调查连环犯罪. 你开发的这种方法, 应至少使用两种不同的方案来产生一个地理轮廓. 你需要发展一种技术, 能综合不同方案的结果为执法人员产生一种有用的预测. 这种预测应基于过去的系列犯罪现场的时间和地点, 提供下次犯罪发生的可能位置. 如果在你们的模型中, 使用了除时间和地点之外的证据, 你必须提供具体的细节, 说明你是如何纳入额外信息的. 你的方法还应提供, 在某一特定情况下方法可靠度的某种形式的估计, 包括适当的警告.

除了要求的一页摘要以外, 你的报告应该包括一个额外的 2 页纸的实施概要. 这个概要应该对潜在问题进行综述. 它要概述你的方法, 描述你的方法适合以及不适合的情况. 概要将会被呈给警方高层阅读, 所以概要中应包括适当的技术细节以适合其读者.

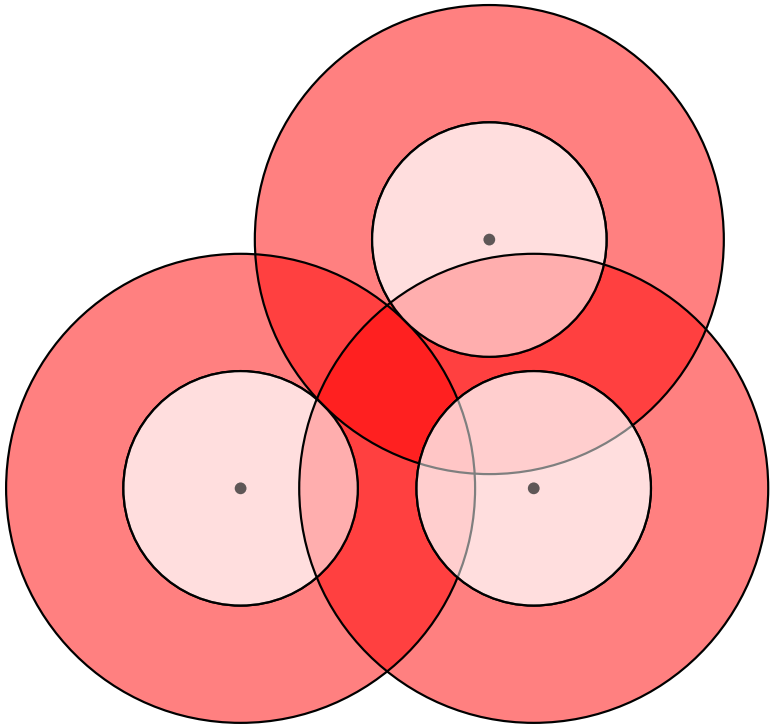
明确问题

- 给定连环犯罪的作案时间和地点及其它潜在的信息, 要求建立模型预测下一次作案的地点.
- 必需建立多种预测模型分别预测下一次作案的地点.
- 必需综合不同预测模型的结果, 并研究其效用.
- 建立一种评估预测模型准确性的方法.

明确概念: 地理轮廓 (geographical profile)



明确概念: 地理轮廓 (geographical profile)

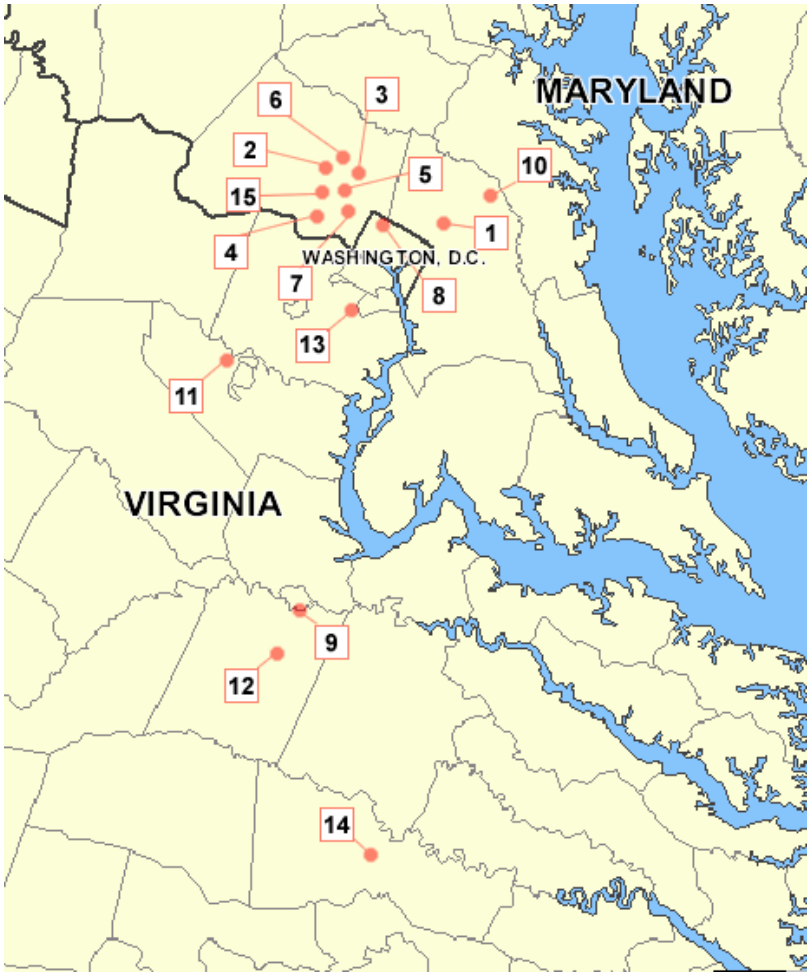


我们先搞个简单的轮廓, 根据犯罪学理论和以前的办案经验

- 作案者不会在离固定活动点 (家, 工作地点等) 很近的地方犯罪.
- 在离作案者的固定工作/生活点越远的地方, 他在那里犯罪的可能性越小.

假设: 罪犯的固定活动地点就是以该现场为圆心的两同心圆间.

可用数据



- 某些著名连环案件的系列案发时间，地点. 获得这些数据的最好方式就是 wiki 百科和 google，例如搜索以下关键词:
  - wiki Peter_Sutcliffe
  - beltway sniper kmz
- 著名连环案件的系列受害人特点，这些 wiki 百科或 google 也有介绍.
- 案发地点周边的其它信息: 人口，地理，交通 (公路，铁路等主干道).

类似问题

- 2011 国赛 A 题: 城市表层土壤重金属污染分析. 已知 8 种主要重金属元素在一些采样点处的浓度, 求金属元素在该城区的空间分布, 确定污染源的位置...
- 传染病源头定位: 已知数位传染病感染者的地理位置和感染时间, 定位传染病源头.

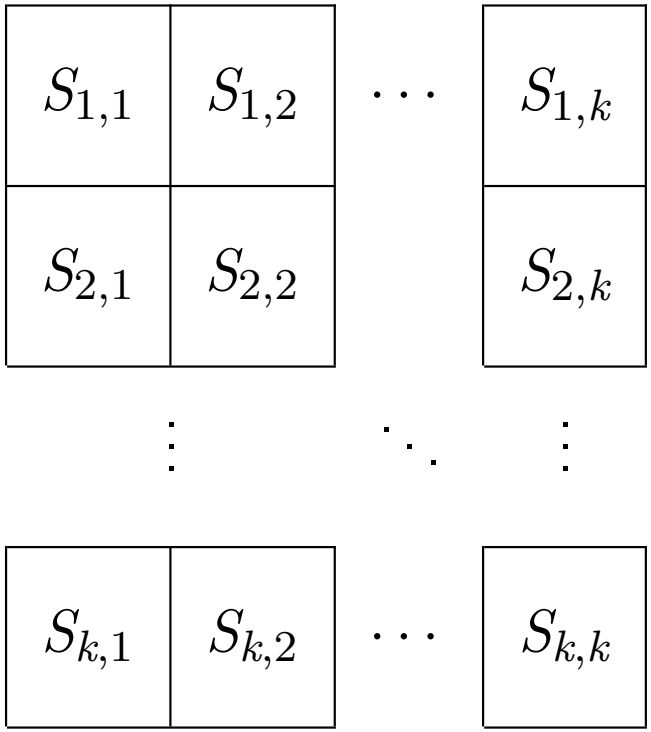
Part II

Predicting a Serial Criminal's Next Crime Location Using Geographical Profiling

文献调研

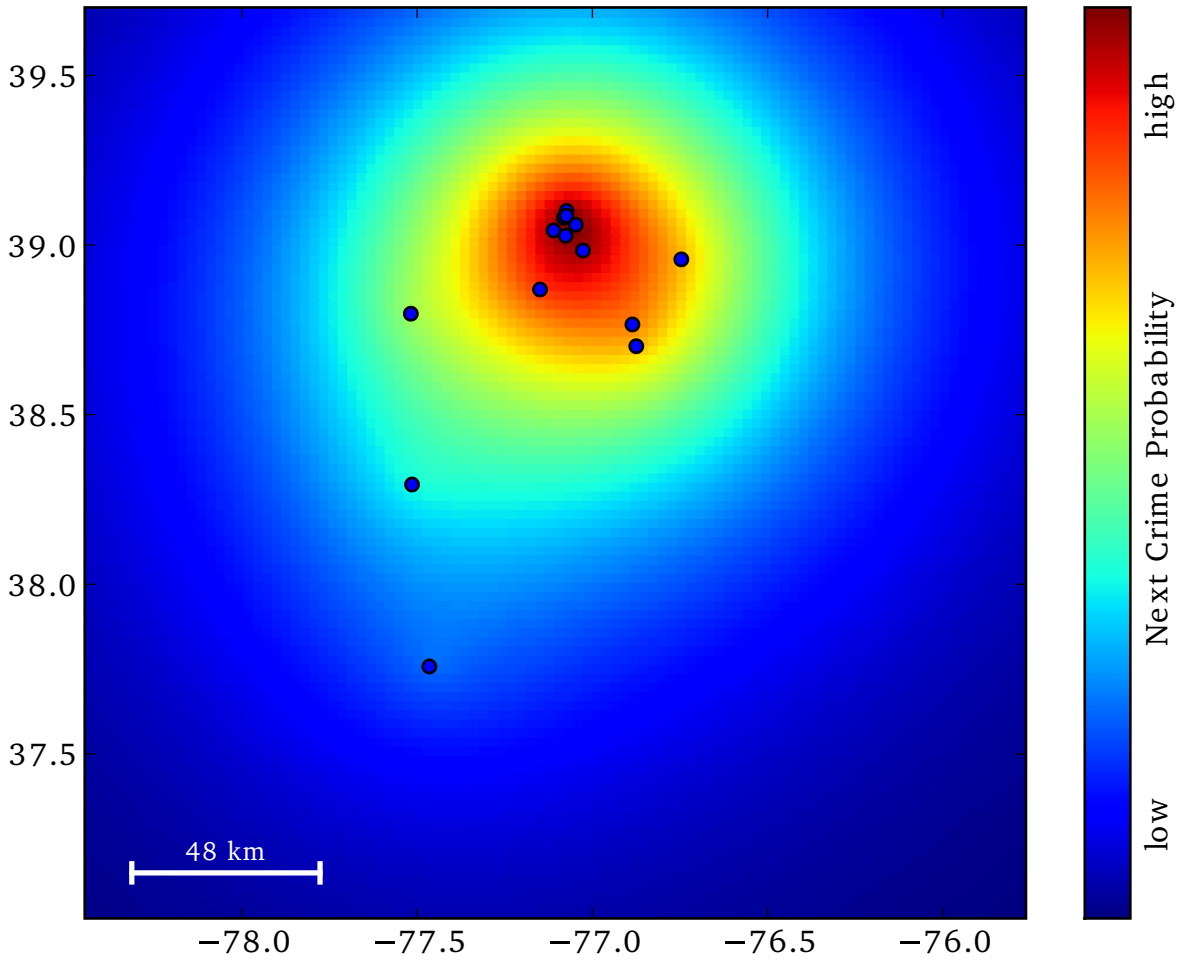
- 绝大多数文献的研究方向是确定连环犯罪作案者的住处位置，而不是下次作案的位置.
- 连环犯罪作案者的行为并不是随机的，而是有潜在原因的.
- 我们很难得到除了时间和地点之外的额外信息.
- 大多研究都是基于犯罪地理学或犯罪心理学，并没有使用严格的数学方法.
- 其中一些研究工作也提出了一些数学方法，这些方法涉及很多经验参数的方程. 我们没有足够的数据来确定这些方程中参数的最优取值.

模型的建立



- 令 $A \subseteq R^2$ 为给定的搜索区域.
- 令 (x_i, t_i) 表示连环犯罪的某个作案地点和相应的时间.
- 定义下次作案地点发生在 x 处的概率 $P(x)$.
- 将区域 A 离散成块 $S_{i,j}$; 问题转化为求 $P(S_{i,j})$, 其中 $P(S_{i,j})$ 为下次为作案地点落在该块中的概率.

实例


39.5
39.0
38.5
38.0
37.5
48 km
−78.0
−77.5
−77.0
−76.5
−76.0
high
Next Crime Probability
low

模型的评价方法或指标

- 需要确定一个数值方法来比较各种不同的犯罪地理学模型.
- 我们可以把某次作案看为当前还未发现, 来研究一个模型的预测精度.
- 命中比分 (hit score) 是在发现下次作案位置之前需要优先搜索的区域所占搜索区域的百分比: 令 S 表示所有块的集合, 令 $L \ni x_{n+1}$ 表示包含下次作案位置的块. 那么 $B = \{S_{i,j} \in S : P(S_{i,j}) > P(L)\}$ 就是概率比下次作案点实际所在块高的所有块的集合. 则命中比分为

$$H = \frac{|B|}{|S|}$$

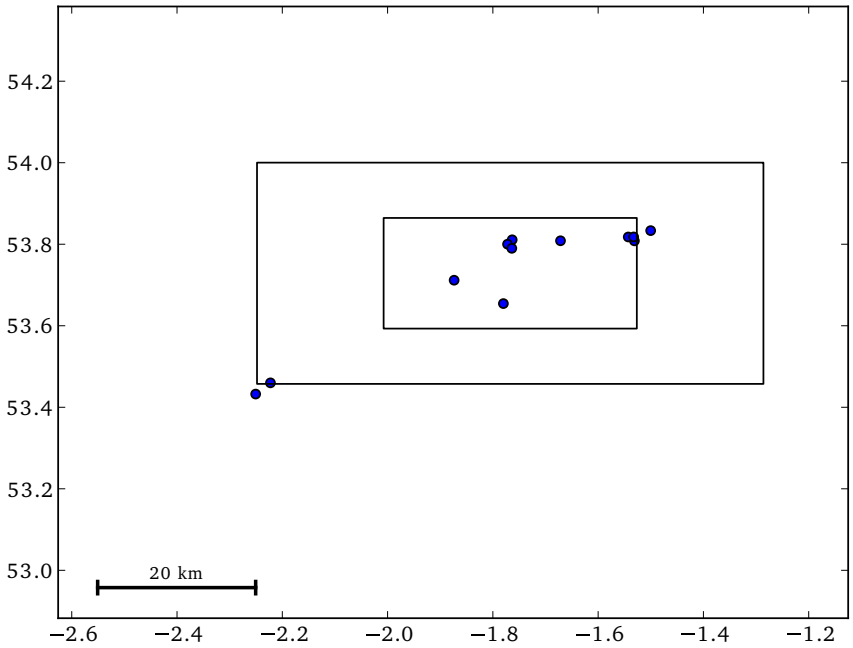
标准差模型

- 标准差矩形和模准差椭圆的定义：　中心由犯罪位置的中心确定，范围由多个作案地点的标准差来确定. 比如标准差矩形

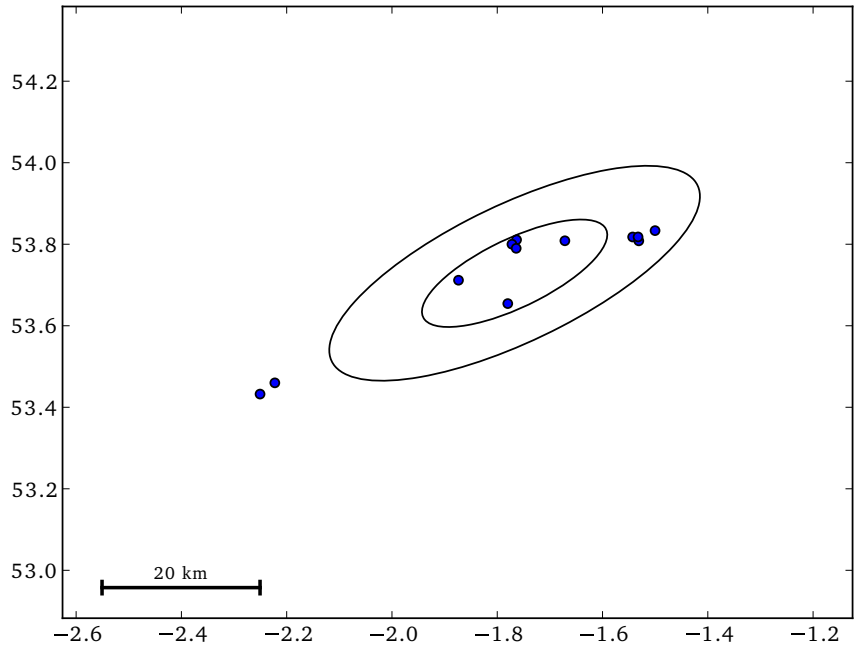
$$
\begin{aligned}
\overline{x} \quad &+ \quad (-c\,\sigma_{lon}, -c\,\sigma_{lat}) \\
\overline{x} \quad &+ \quad (-c\,\sigma_{lon}, +c\,\sigma_{lat}) \\
\overline{x} \quad &+ \quad (+c\,\sigma_{lon}, +c\,\sigma_{lat}) \\
\overline{x} \quad &+ \quad (+c\,\sigma_{lon}, -c\,\sigma_{lat})
\end{aligned}
$$

- 标准差法并不能给出标准差范围内概率分布的有关信息，只是提供了一个缩小潜在的搜索区域的方法.

标准差模型



标准差矩阵



标准差椭圆

中心化方法

中心化方法: 定义犯罪的中心, 各点下次作案概率随着离中心的距离增加而减小.

Definition

令 C 为一个连环犯罪的中心. 对于中心化方法, 区域 $S_{i,j}$ 包函下次作案点的概率为

$$P(S_{i,j}) = \frac{1}{d(S_{i,j}, C)}.$$

有多种计算犯罪中心 C 的方法, 比如平均法, 中位数法.

概率距离法

概率距离法: 基于与以往作案点的距离, 建立一个概率分布函数.

Definition

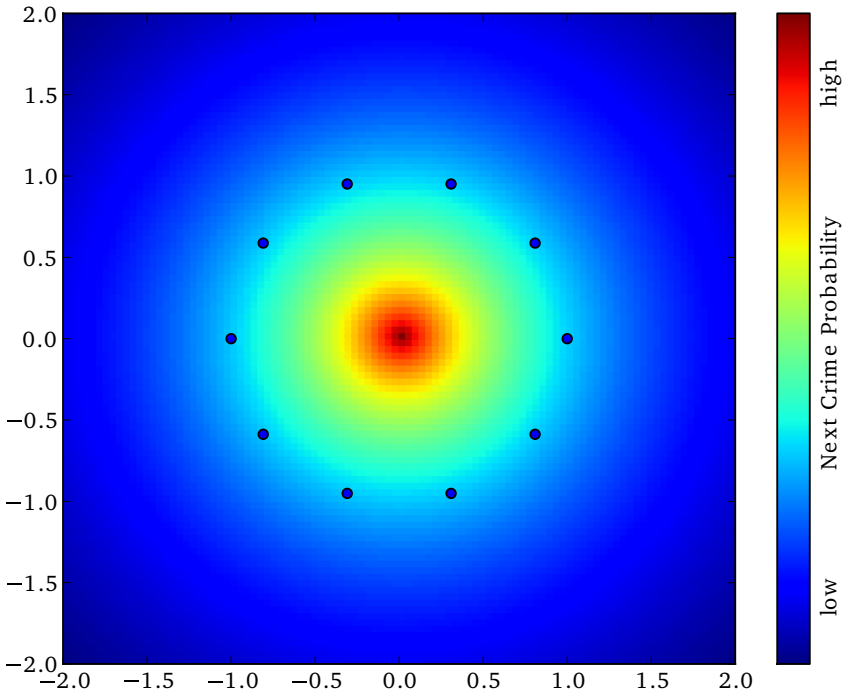
对于一种概率距离法, 区域 $S_{i,j}$ 包函下次作案点的概率为

$$P(S_{i,j}) = \sum_{k=1}^{n} f(d(S_{i,j}, x_k)),$$

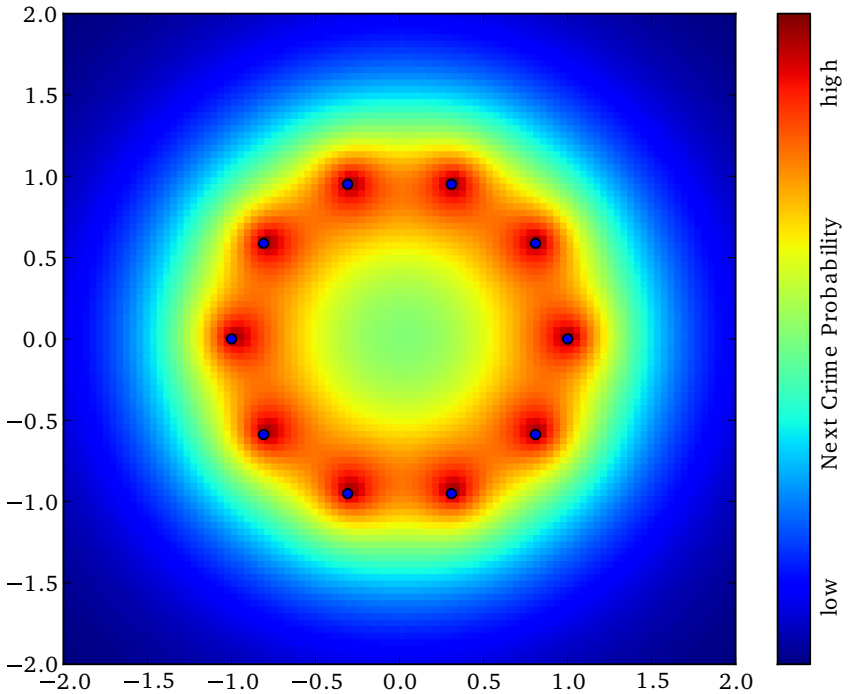
其中 f 距离衰减函数.

- **线性距离衰减:** 概率以线性的方式衰减: $f(d) = \alpha - \beta d$.
- **指数距离衰减:** 概率以指数的方式衰减: $f(d) = e^{-\gamma d}$.

中心化方法与指数距离衰减法举例



中心化方法



指数距离衰减法

时空模型

- 基于本文的空间模型，引入连环犯罪的作案的时间数据，构造时空模型.
- 该模型的构造是受到“时间上最近的作案地点比其它作案地点更加与连环犯罪的空间行为相关”的启发.
- 并不是所有的空间模型都能引入时间数据来构造时空模型的.

纳入时间数据

Definition

本文定义连环犯罪的 **时间权重因子**:

$$w_i = \frac{t_i - t_1}{t_n} + k$$

其中 w_i 表示第 i 次作案的时间权重, k 为一个非零常数.

本文取 $k = 0.1$, 这样一来, 首次作案的时间权重不为零, 同时最后一次作案的时间权重大约为首次作案 10 倍.

加入时间权重的模型

下面举例说明本文如何在空间模型中加入时间权重. 令 W 表示时间权重的总和.

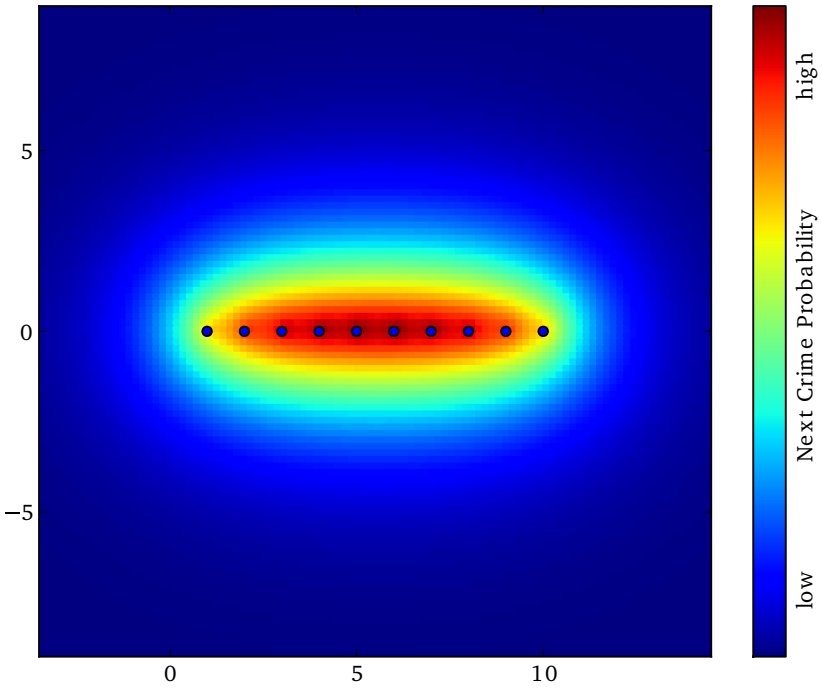
时间中心法

$$C = \frac{1}{W}\sum_{i=1}^{n} x_i w_i$$

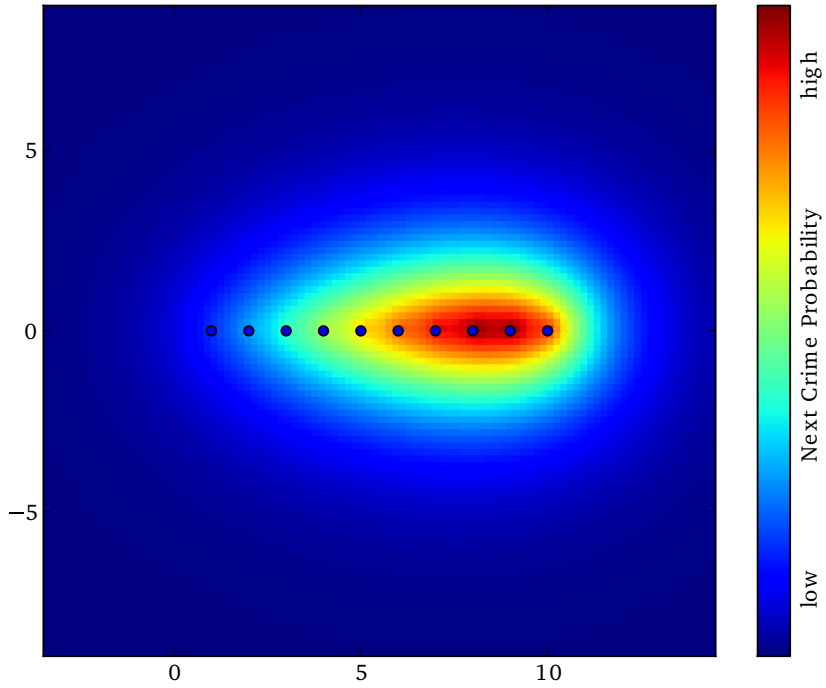
时间概率距离法

$$P(S_{i,j}) = \sum_{k=1}^{n} w_k f(d(S_{i,j}, x_k))$$

空间模型和时空模型对比



空间模型



时空模型

整合模型

对于多个预测模型, 我们希望组合它们的预测结果.

Definition

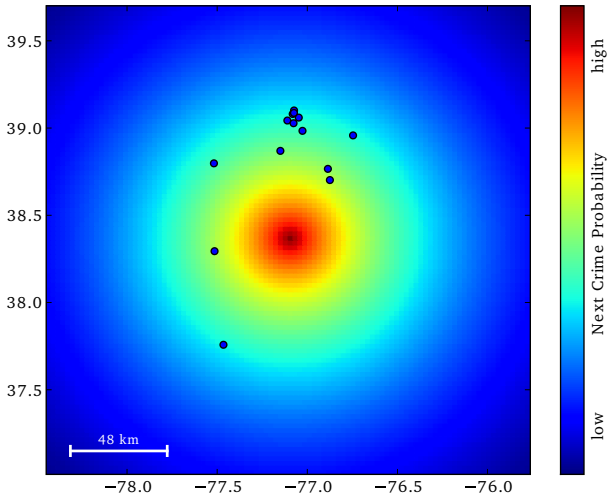
给定预测方法 $P_1, \ldots, P_n$, 以及相应的权重 $W_1, \ldots, W_n$.

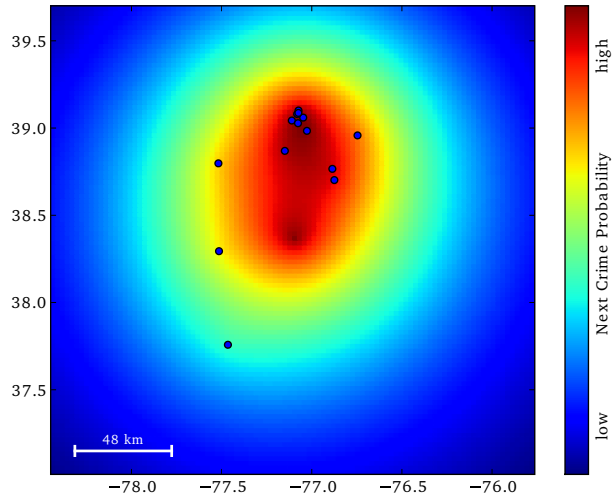
线性组合 方法表示为:

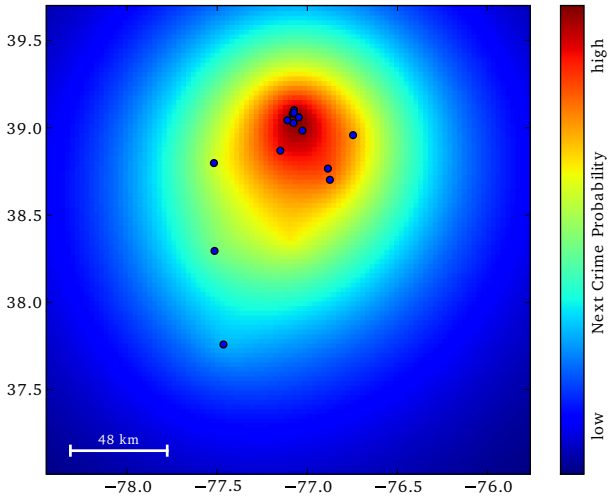
$$P_{S_{i,j}} = \sum_{k=1}^{n} W_k P_k(S_{i,j}).$$

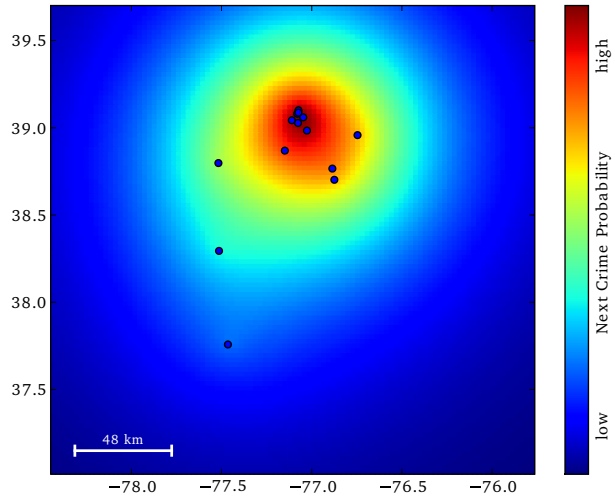
对数组合 方法表示为:

$$P_{S_{i,j}} = \prod_{k=1}^{n} \left(P_k(S_{i,j})\right)^{W_k}.$$





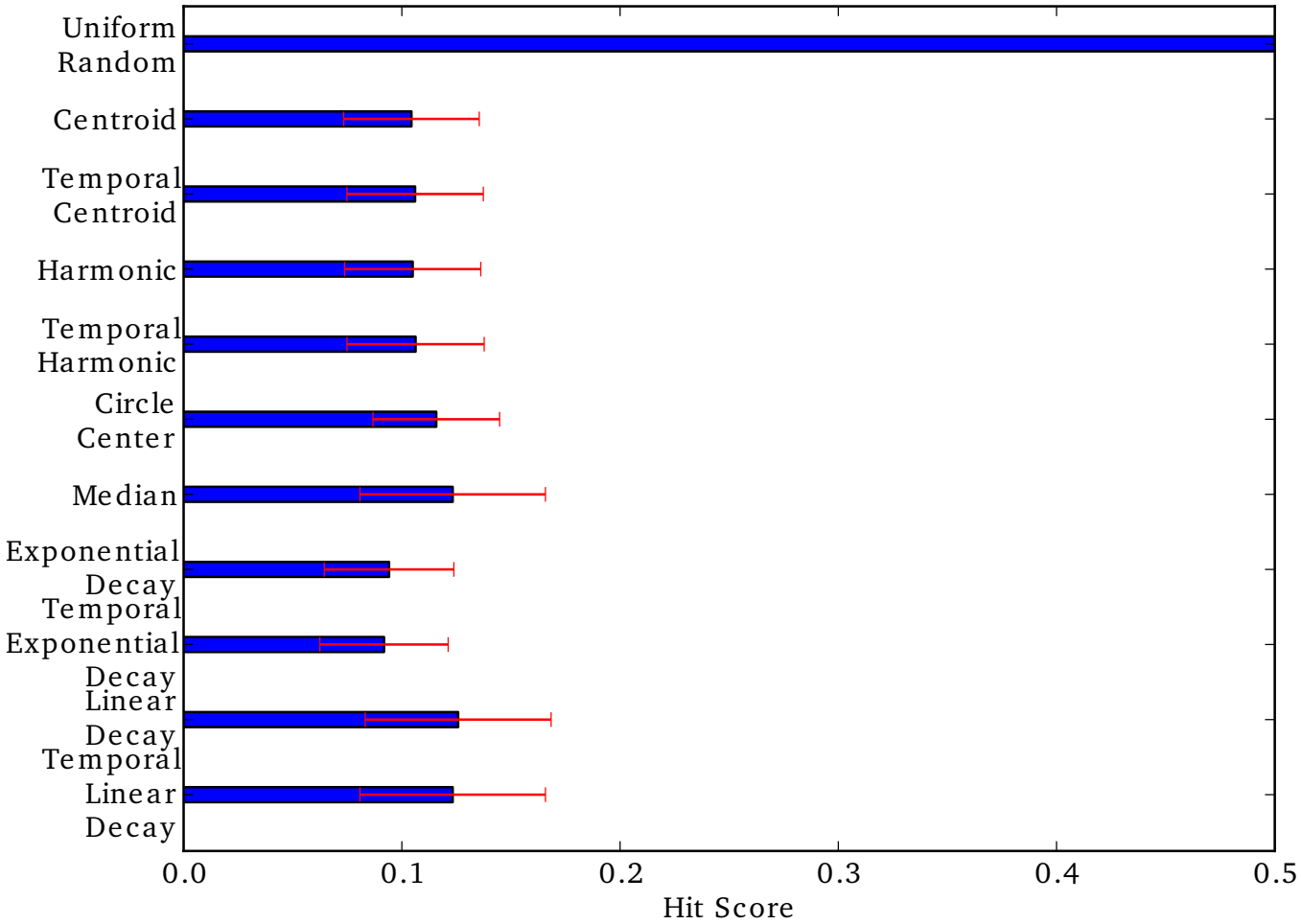




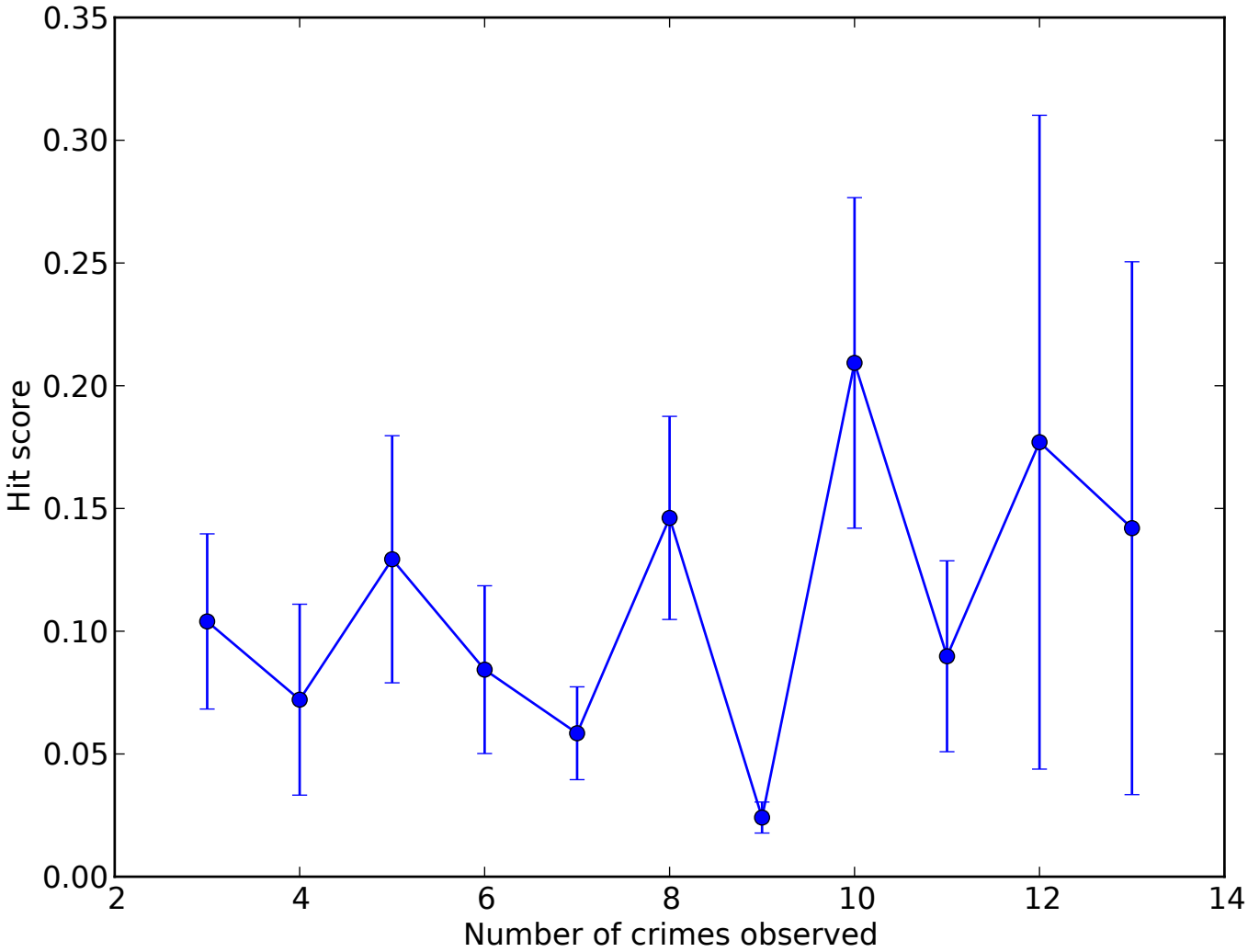
整合模型举例

39.5
39.0
38.5
38.0
37.5
48 km
−78.0
−77.5
−77.0
−76.5
−76.0
high
Next Crime Probability
low

单一模型的结果


Uniform Random
Centroid
Temporal Centroid
Harmonic
Temporal Harmonic
Circle Center
Median
Exponential Decay
Temporal Exponential Decay
Linear Decay
Temporal Linear Decay
0.0
0.1
0.2
0.3
0.4
0.5
Hit Score

输入不同作案次数，指数衰减模型的准确度得分



结论

- 各种方法的表现没有明显的统计学差异.
- 单一的模型的精度信赖于所研究的连环犯罪.
- 结果对于数据的选用很敏感.

Part III

程序实现

正则表达式相关函数 regexp

Command Window

```
>> data = '9/5/2002,38.766031439525,-76.885793752519'

data =
     9/5/2002,38.7660314395259,-76.8857937525191

>> data = regexp(data, ',', 'split')

data =
     '9/5/2002'    '38.766031439525'    '-76.885793752519'

>> [date, lat, long] = data{:}

date =
     9/5/2002
lat =
     38.7660314395259
long =
     76.8857937525191

>>
```

绝对时间转换函数 datenum

Command Window

```
>> date = '9/11/2012'

data =
     9/11/2012

>> t = datenum(date1, 'dd/mm/yyyy')

t =
     735182

>> v = datevec(735182)

v =
     2012    11     9     0     0     0
```

>> 假设这张照片是今年娃他爹拍的，三个问题：

1. 照片上的婴儿多大?
2. 照片是哪一天拍的?
3. 孩子是不是亲生的?

网格 meshgrid

Command Window

```
>> [x, y] = meshgrid(1:3, 1:3)
x =
     1     2     3
     1     2     3
     1     2     3
y =
     1     1     1
     2     2     2
     3     3     3
>> rsq = (x-2).^2 + (x-2).^2
rsq =
     2     1     2
     1     0     1
     2     1     2
>> r = sqrt(rsq)
r =
    1.4142    1.0000    1.4142
    1.0000         0    1.0000
    1.4142    1.0000    1.4142
```

$(1,1)$	$(2,1)$	$(3,1)$
$(1,2)$	$(2,2)$	$(3,2)$
$(1,3)$	$(2,3)$	$(3,3)$

2	1	2
1	0	1
2	1	2

$\sqrt{2}$	1	$\sqrt{2}$
1	0	1
$\sqrt{2}$	1	$\sqrt{2}$

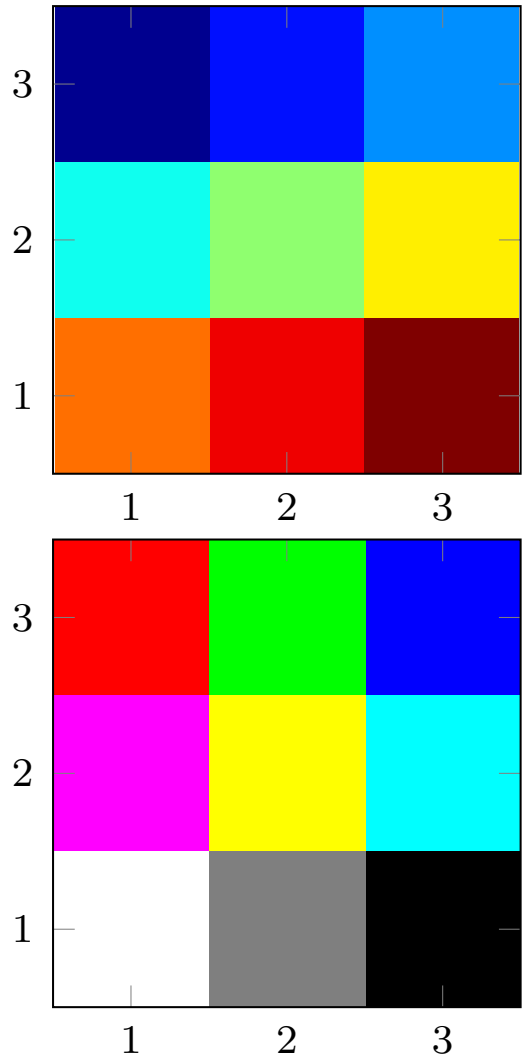
数据可视化

Command Window

```
fx >> x = [1 2 3; 4 5 6; 7 8 9];
fx >> imagesc(x);
fx >> R = [1 0 0; 1 1 0; 1 0.5 0];
fx >> G = [0 1 0; 0 1 1; 1 0.5 0];
fx >> B = [0 0 1; 1 0 1; 1 0.5 0];
fx >> RGB = cat(3,R,G,B)

  RGB(:,:,1) =
     1.00        0        0
     1.00     1.00        0
     1.00     0.50        0
  RGB(:,:,2) =
        0     1.00        0
        0     1.00     1.00
     1.00     0.50        0
  RGB(:,:,3) =
        0        0     1.00
     1.00        0     1.00
     1.00     0.50        0

fx >> image(RGB);
fx >>
```



数据的读入

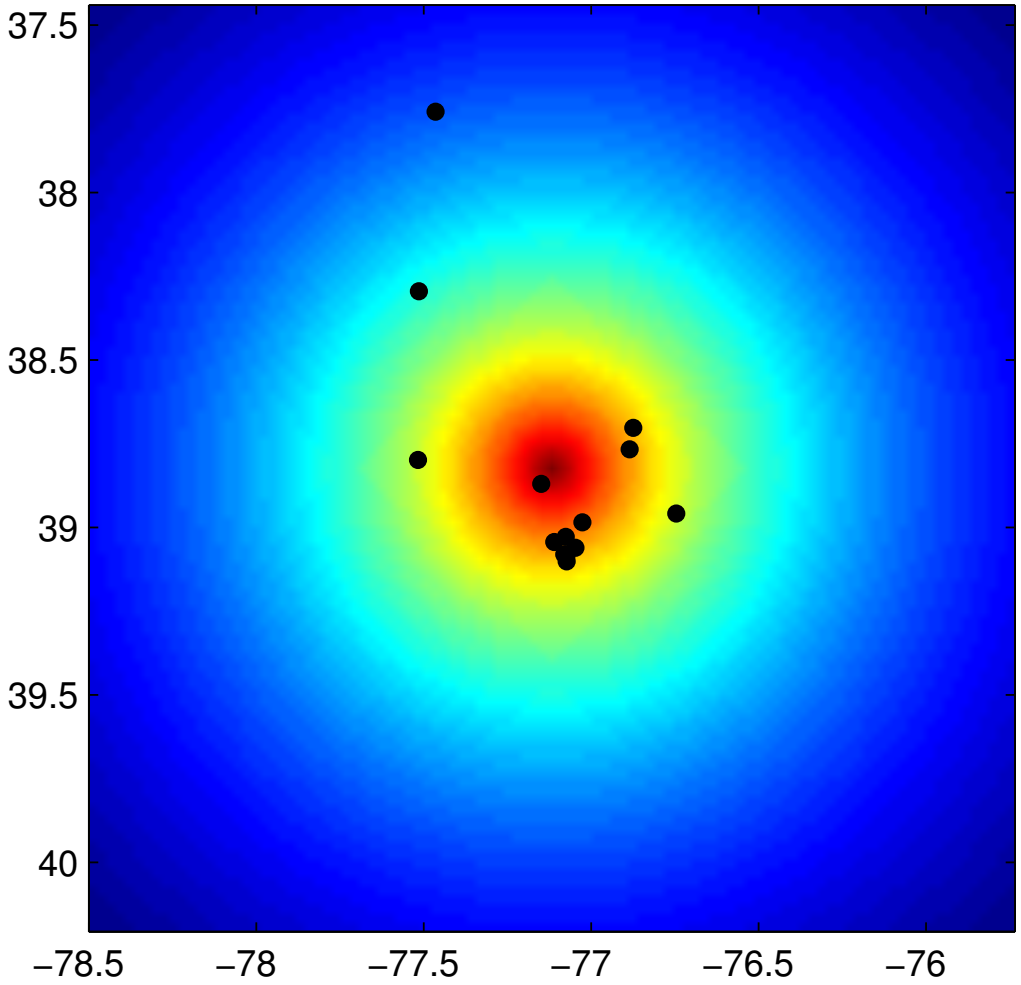
beltway.dat	文件格式说明
Beltway sniper data. 9/5/2002,38.7660314395259,-76.8857937525191 ⋮	数据描述, 计算时忽略 月/日/年, 经度, 纬度

```
01 filename = 'beltway.dat';
02 fid = fopen(filename,'r');
03 tline = fgetl(fid); tline = fgetl(fid);
04 while ischar(tline)
05     data = regexp(tline, ',', 'split');
06     [date,lat,long] = data{:};
07     time = datenum(date, 'dd/mm/yyyy');
08     lat  = str2num(lat); long = str2num(long);
09     tline = fgetl(fid);
10 end
11 fclose(fid);
```

中心化方法

```
01 weights = ones(n,1);
02 x = long;                          y = lat;
03 xbar = x*weights;                  ybar = y*weights;
04 xbar = xbar/n;                     ybar = ybar/n;
05 xran = max(long)-min(long); yran = max(lat)-min(lat);
06 ran = max(xran,yran);
07 DELTA = 100;
08 dx = ((xbar + ran) - (xbar - ran))/DELTA;
09 dy = ((ybar + ran) - (ybar - ran))/DELTA;
10
11 x = (xbar - ran)-dx : dx : (xbar + ran)+dx;
12 y = (ybar - ran)-dy : dx : (ybar + ran)+dy;
13 [X,Y] = meshgrid(x,y);
14 Z = sqrt((X-xbar).^2 + (Y-ybar).^2);
15 P = 1./(Z+1);
16 P = P/sum(sum(P));
17
18 imagesc(x, y ,P)
19 hold on; axis image
20 plot(longs, lats, 'k.', 'markersize', 15);
```

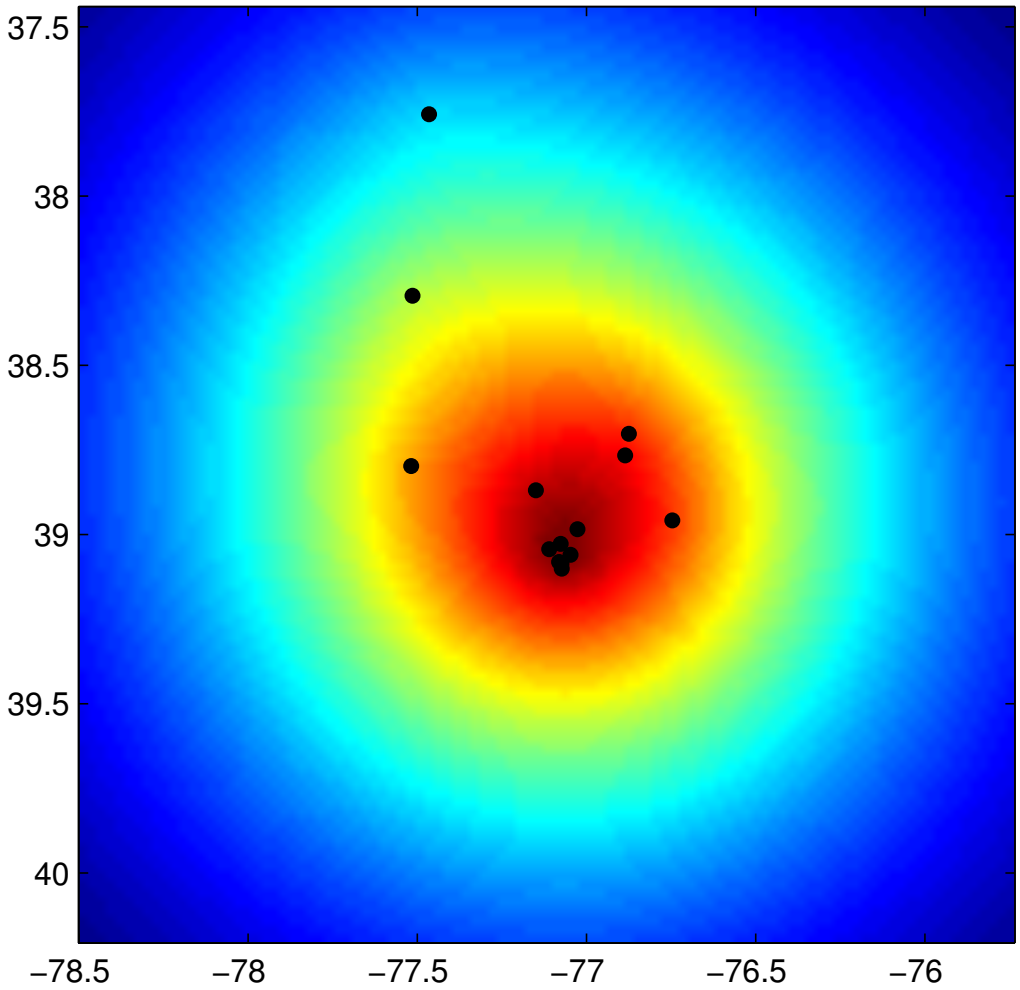
中心化方法



概率距离法

```
01 weights = ones(n,1);
02 x = long;                          y = lat;
03 xbar = x*weights;                  ybar = y*weights;
04 xbar = xbar/n;                     ybar = ybar/n;
05 xran = max(long)-min(long); yran = max(lat)-min(lat);
06 ran = max(xran,yran);
07 DELTA = 100;
08 dx = ((xbar + ran) - (xbar - ran))/DELTA;
09 dy = ((ybar + ran) - (ybar - ran))/DELTA;
10 x = (xbar - ran)-dx : dx : (xbar + ran)+dx;
11 y = (ybar - ran)-dy : dx : (ybar + ran)+dy;
12 [X,Y] = meshgrid(x,y);
13 P = zeros(size(X));
14 for j = 1:length(long)
15      R = sqrt((X-long(j)).^2 + (Y-lat(j)).^2)
16      P = P+exp(-R);
17 end
18 imagesc(x, y, P)
19 hold on; axis image
20 plot(longs, lats, 'k.', 'markersize', 15);
```

概率距离法



Thank You!!!