基于空间聚类的出租车轨迹数据分析

汇报人：钟代琪

2022.10.12

目录 Content

上下客热点分析

K-means核心思想：通过迭代把数据对象划分到不同的簇中，以求目标函数最小化，从而使生成的簇尽可能地紧凑和独立。

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} (C_i - x)^2$$

步骤：

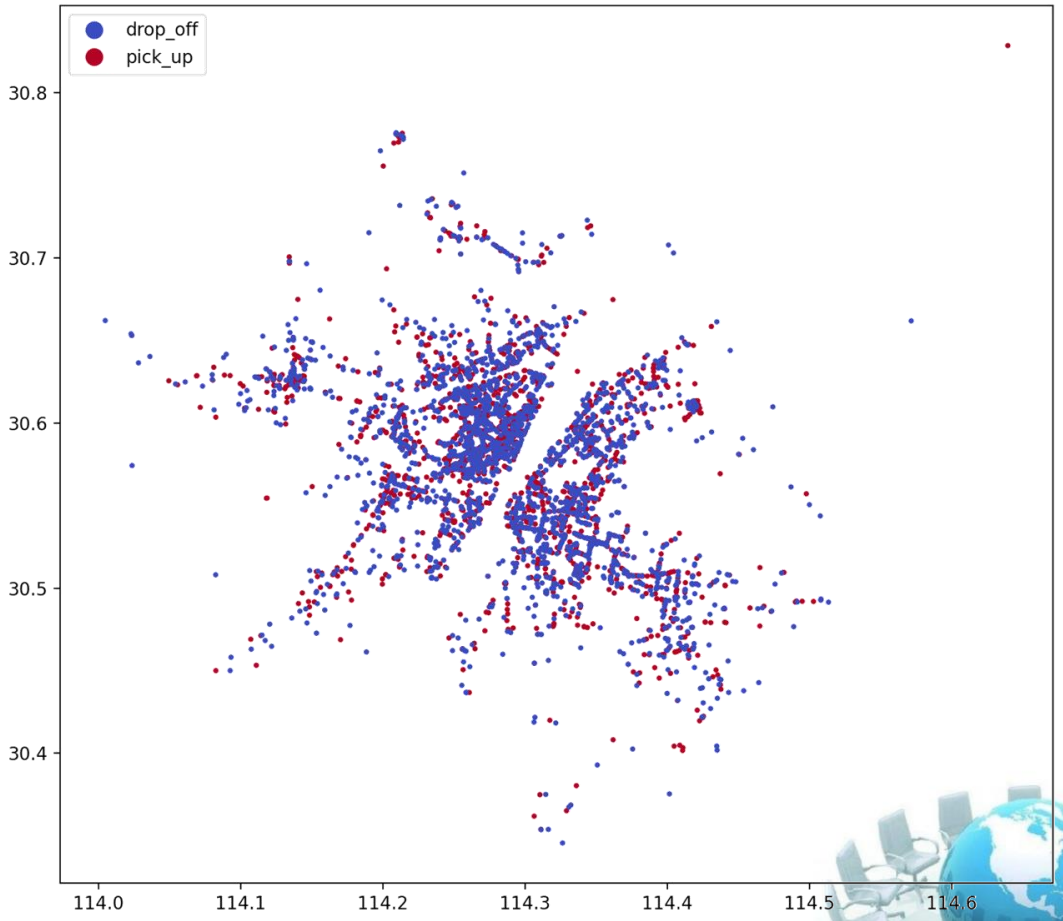
Step1. 随机选取K个对象作为初始的K个簇的质心；

Step2. 将其余对象根据其与各个簇质心的距离分配到最近的簇；

再求新形成的簇的质心；

Step3. 不断重复迭代重定位过程，直到满足终止条件为止。

武汉市2018年11月5日出租车上下客记录

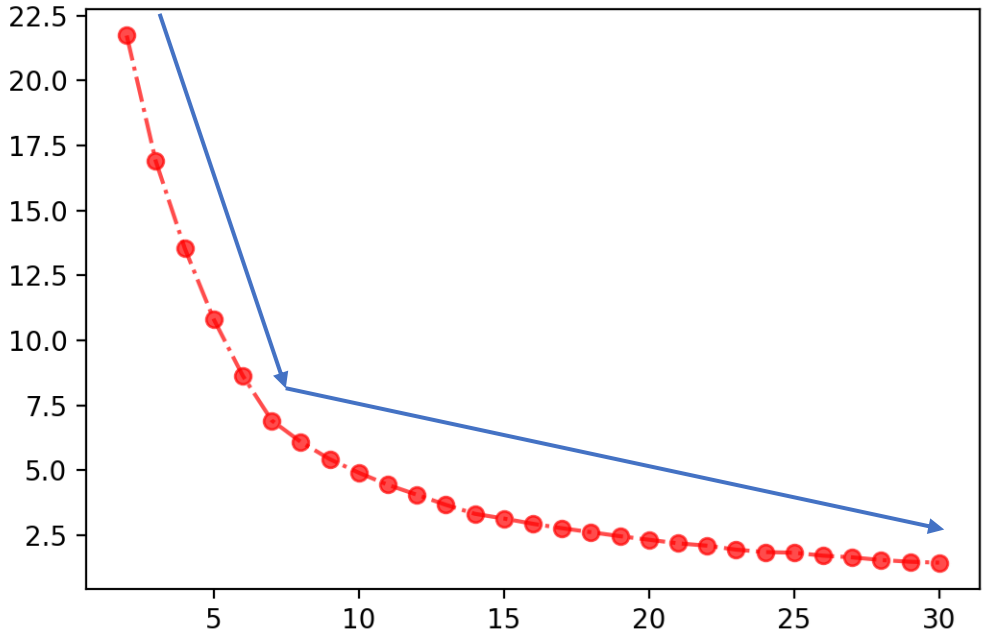


聚类数确定 — 手肘法

手肘法思想：

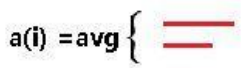
k-means以最小化样本与质心平方误差作为目标函数，将每个簇的质点与簇内样本点的平方距离误差和称为畸变程度。

畸变程度会随着K的增加而降低，在达到某个临界点时畸变程度会得到极大改善，之后缓慢下降，其函数形状类似一个手肘。该临界点可以考虑为聚类性能较好的参考点。

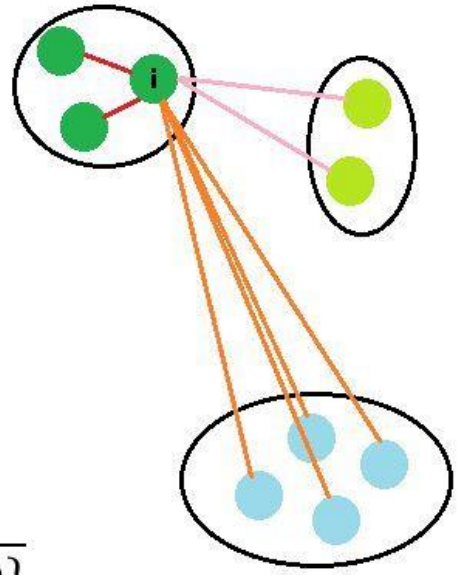


聚类数确定 — 轮廓系数

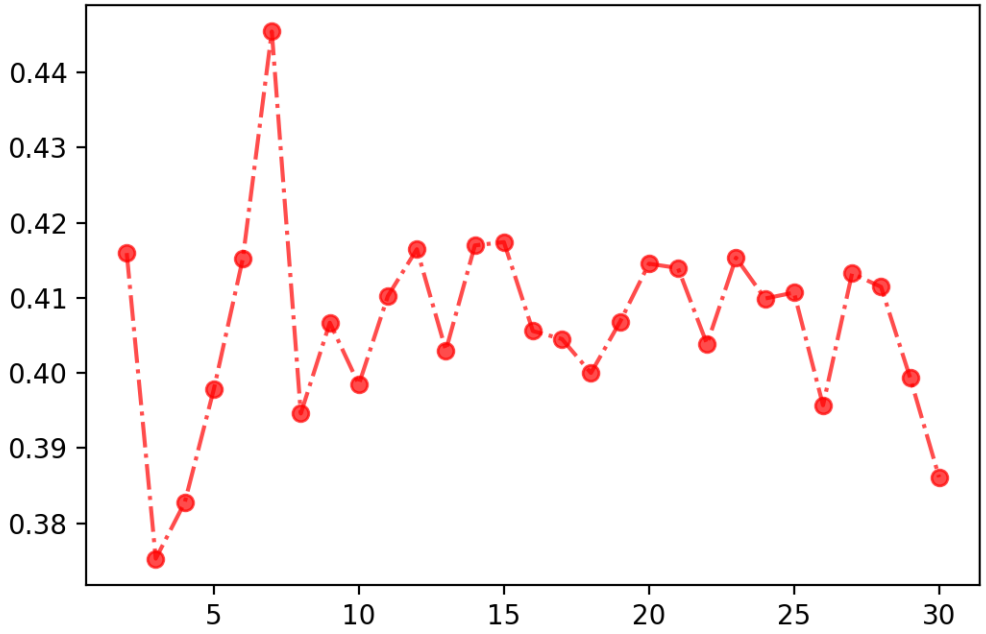
轮廓系数 $S = \sum_i S(i)$

内聚度 $a(i) = \text{avg}\{$ 

分离度 $b(i) = \min\{\ \text{avg}\{$  $,\ \text{avg}\{$ 



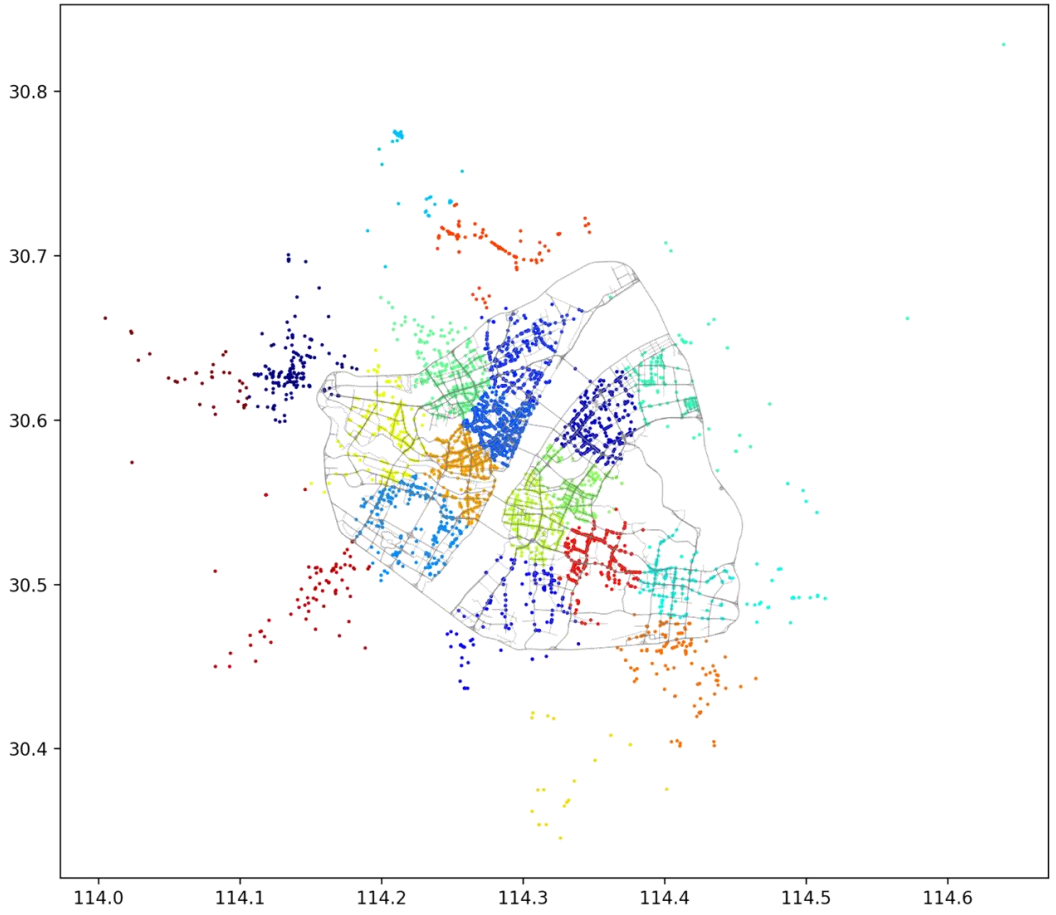
$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$



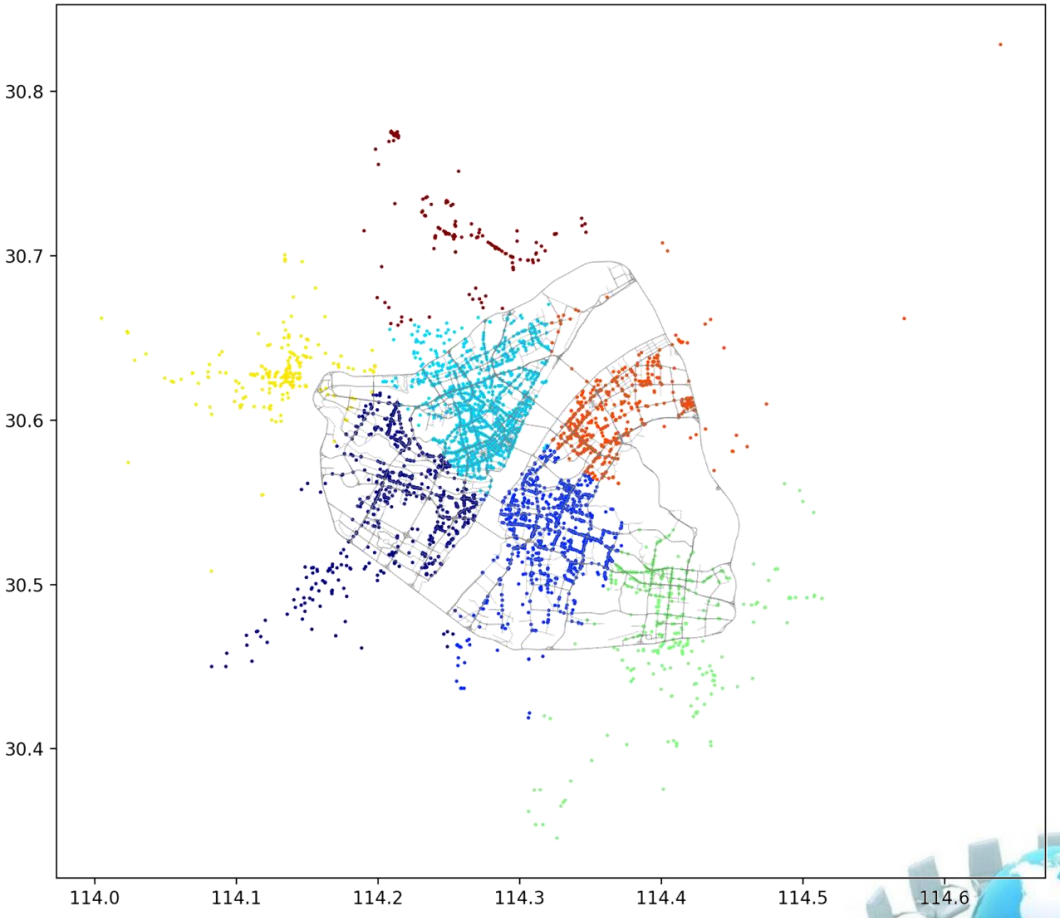
$a(i) =$ avg{i到所属簇内所有其他点的距离}

$b(i) =$ min{i到其他某一簇内所有点的平均距离}

聚类结果



$n = 20$



$\boldsymbol{n = 7}$

DBSCAN

1个核心思想：基于密度

2个算法参数：邻域半径eps和最少点数目MinPts

3种点的类别：核心点，边界点、噪声点

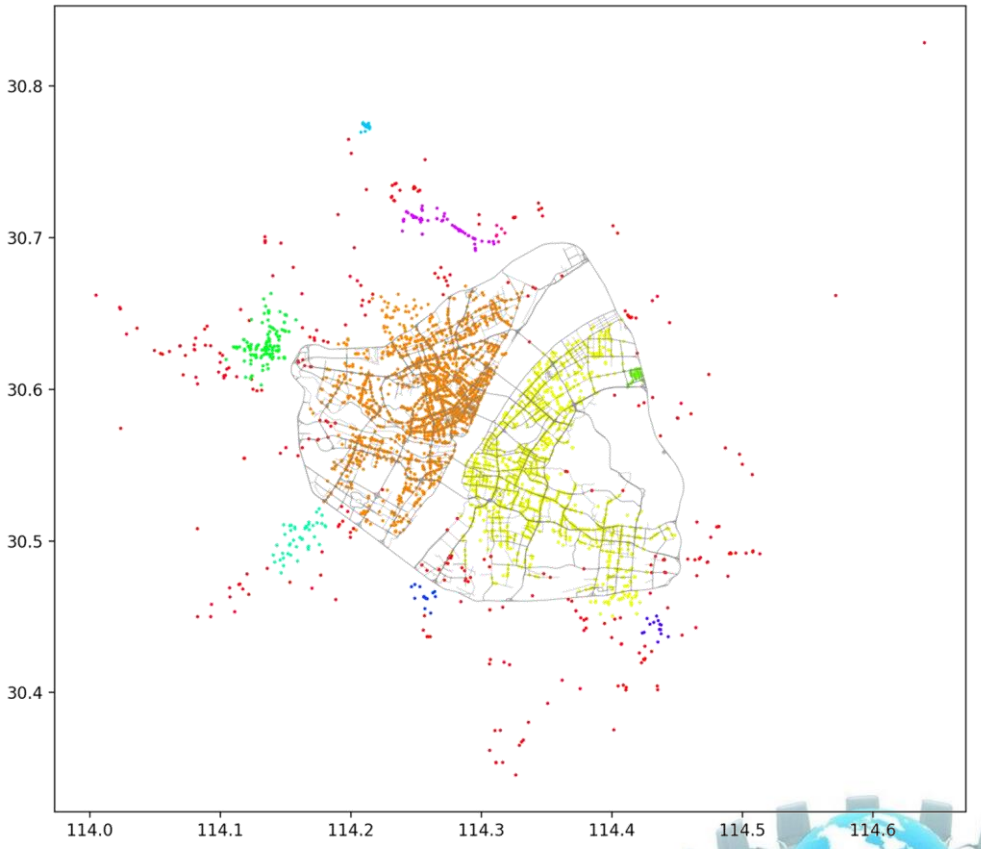
4种点的关系：密度直达，密度可达，密度相连，非密度相连

Step1. 从数据集中任意选取一个数据对象点 p；

Step2. 如果对于参数 Eps 和 MinPts，所选取的数据对象点 p 为核心点，则找出所有从 p 密度可达的数据对象点，形成一个簇；

Step3. 如果选取的数据对象点 p 是边缘点，选取另一个数据对象点；

Step4. 重复（2）、（3）步，直到所有点被处理。



$eps = 0.01, MinPts = 10$

DBSCAN

优点：

1) 可以自主计算聚类的数目，不需要人为指定

2) 不要求类的形状是凸的，可以是任意形状的

3) 对噪音不敏感

4) 算法应用参数少，只需要两个

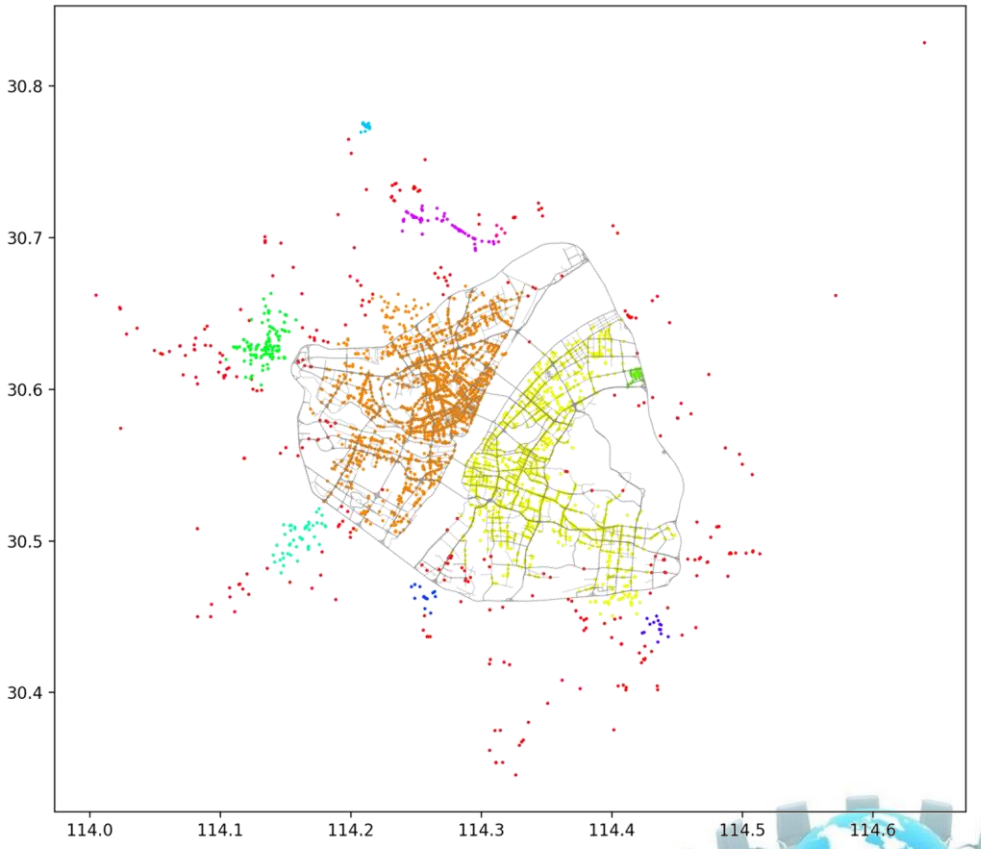
5) 聚类结果几乎不依赖于节点的遍历顺序

缺点：

1) 如果样本集较大时，聚类收敛时间较长

2) 聚类效果依赖于距离公式的选取

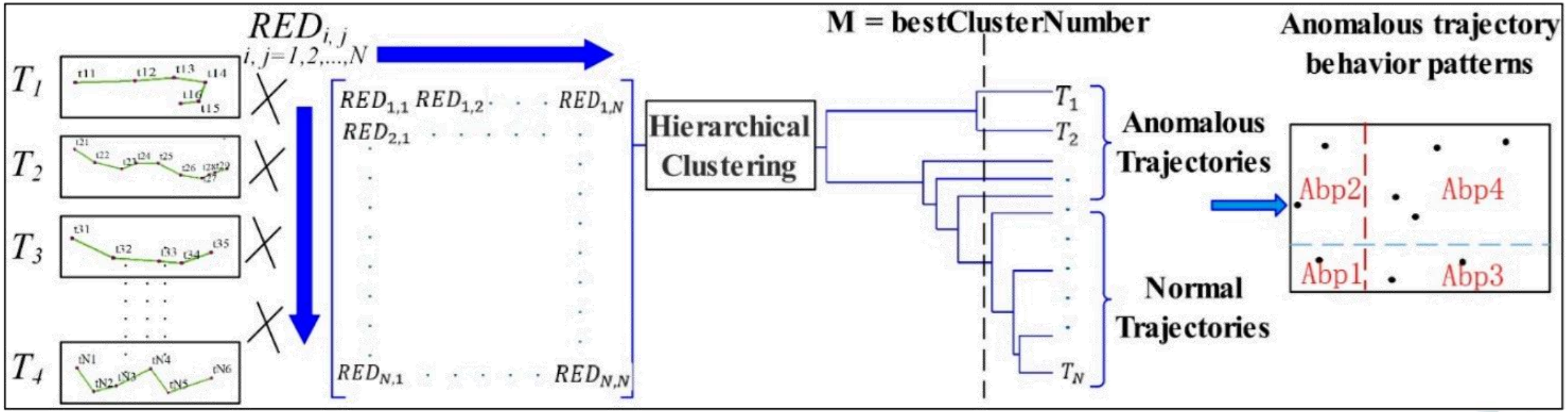
3) 不适合数据集中密度差异很大的情形



$eps = 0.01, MinPts = 10$

异常轨迹分析

1.计算轨迹距离/相似性矩阵
2.进行层次聚类
3.通过统计指标确定最佳聚类个数
4.根据统计指标识别异常轨迹



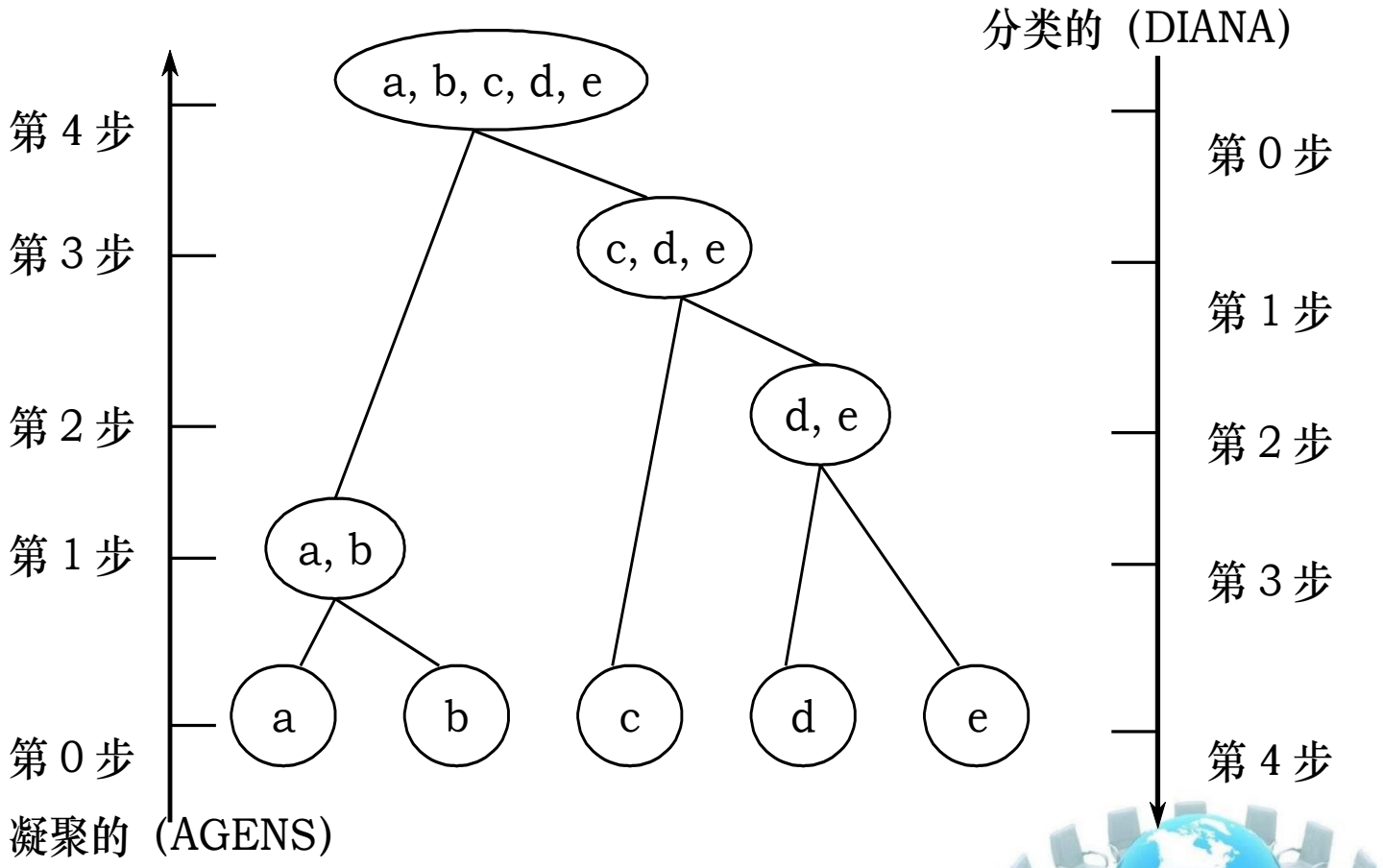
Detecting anomalous trajectories and behavior patterns using hierarchical clustering from taxi GPS data (Wang,Y.,Qin,K.,Chen,Y.,&Zhao,P.)

层次聚类

层次聚类按数据分层建立簇，形成一棵以簇为节点的树，称为聚类图。

按自底向上层次分解，则称为凝聚的层次聚类。

按自顶向下层次分解，就称为分裂的层次聚类。



层次聚类

以凝聚的层次聚类为例：

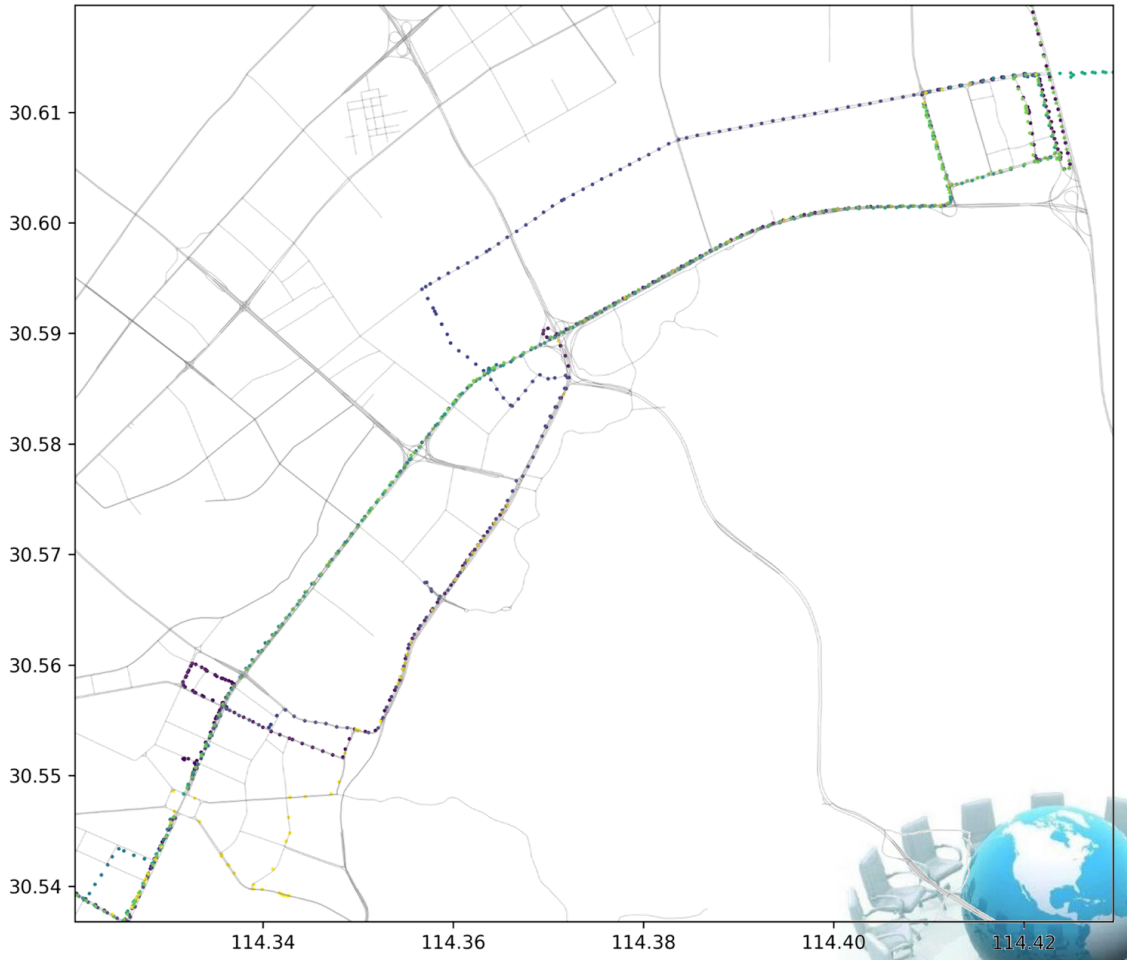
假设有 n 个待聚类的样本：

1.初始化,将每个样本都视为一个聚类；

2.计算各个聚类之间的相似度；

3.寻找最近的两个聚类，将他们归为一类；

4.重复2，3；直到所有样本归为一类。



层次聚类

相似度确定

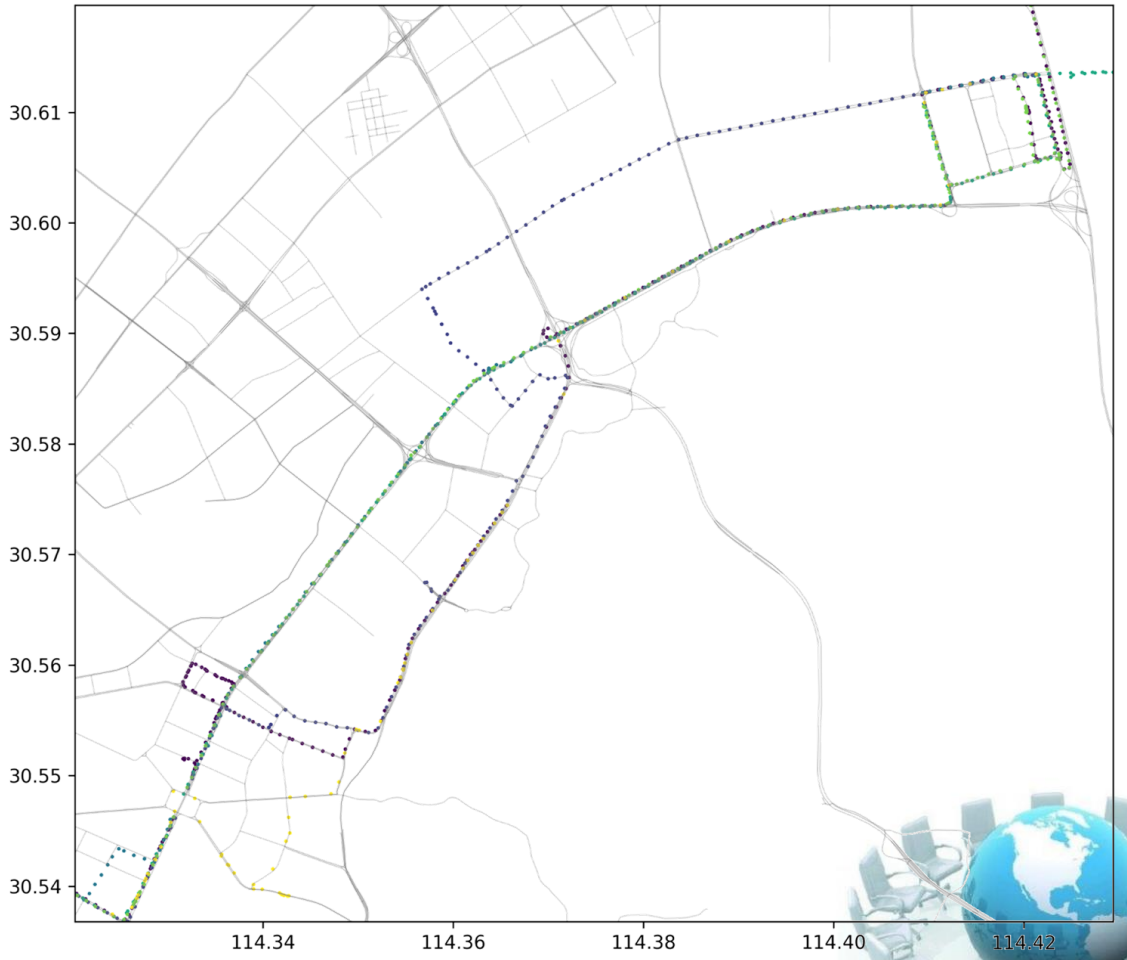
原则：最大化类间距离、最小化类内距离

$$SSW(M) = \max_{t}\left\{\max_{i,j}(1 - IED(T_i, T_j)_{T_i \neq T_j \in C_t})\right\} + \sum_{|C_t=1|} 1$$

$$SSB(M) = \sum_{t=1}^{M}\sum_{s>t}^{M} \min\left(1 - IED(T_i, T_j)_{T_i \in C_t, T_j \in C_s}\right)$$

其中IED为两轨迹直接距离的描述,M为设置的聚类数。

SSW越小则类内距离越小，SSB越大则类间距离越大，聚类效果越好。



层次聚类

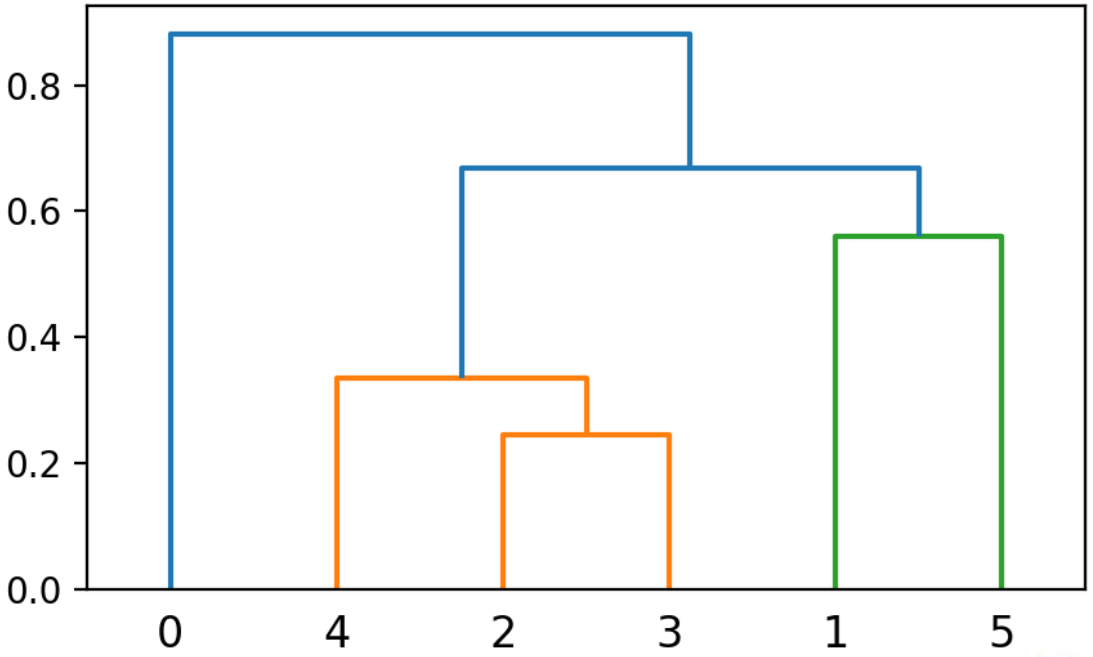
统计指标 — 聚类数确定

$$\text{WB-index} = M * SSW(M)/SSB(M)$$

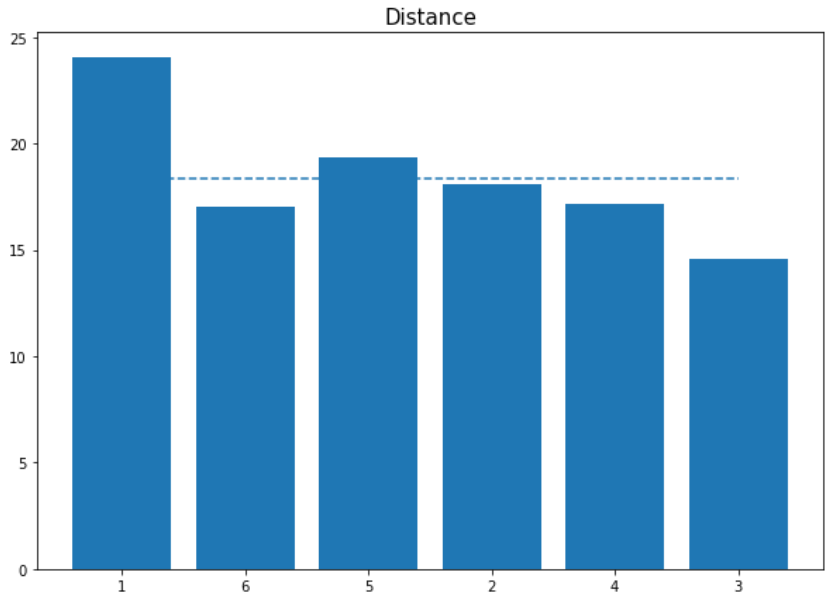
$$\text{CH-index} = \frac{SSB(M)/M - 1}{SSW(M)/N - M}$$

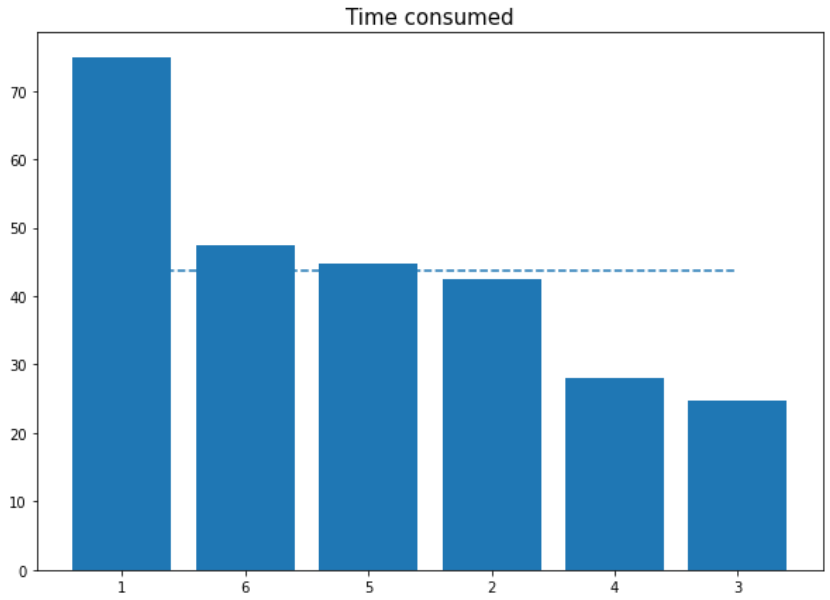
$$\text{Xu-index} = \log\sqrt{(SSW(M))/N^2} + logM$$

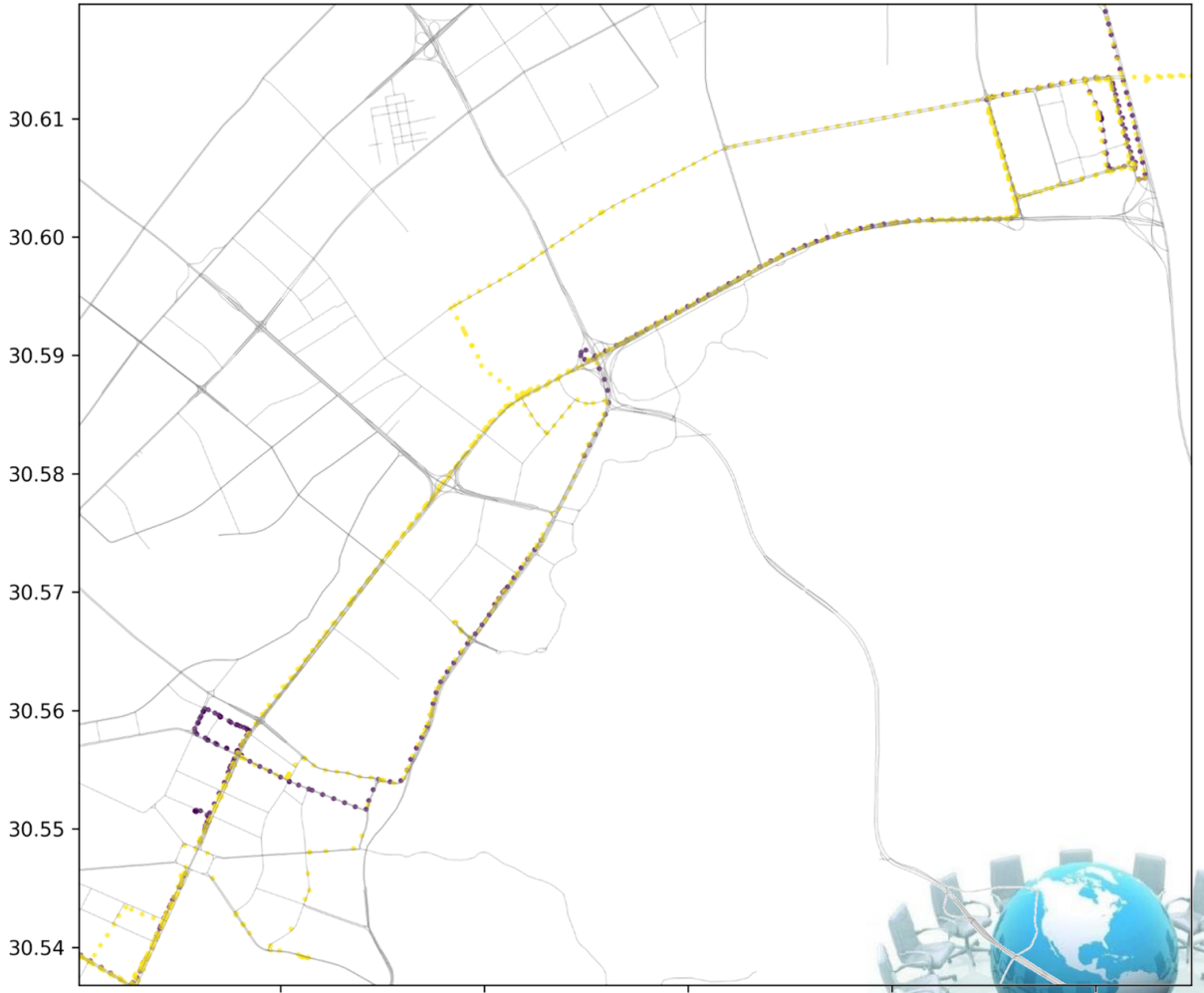
	WB-index	CH-index	Xu-index
2	3.83517	0.41871	-1.17908
3	4.31683	0.28088	-0.77362
4	5.75577	0.22085	-0.48593
5	7.19472	0.18148	-0.26279



异常轨迹探测结果


Distance
25
20
15
10
5
0
1
6
5
2
4
3


Time consumed
70
60
50
40
30
20
10
0
1
6
5
2
4
3


30.61
30.60
30.59
30.58
30.57
30.56
30.55
30.54
114.34
114.36
114.38
114.40
114.42