# 以 OPTICS 演算法識別階層性密度差異的時空群聚結構
# An OPTICS-based Algorithm for Identifying Spatio-Temporal Density Faults in Hierarchical Clustering Structures

游孟純 Meng-Chun You

統計碩士學程學位考試

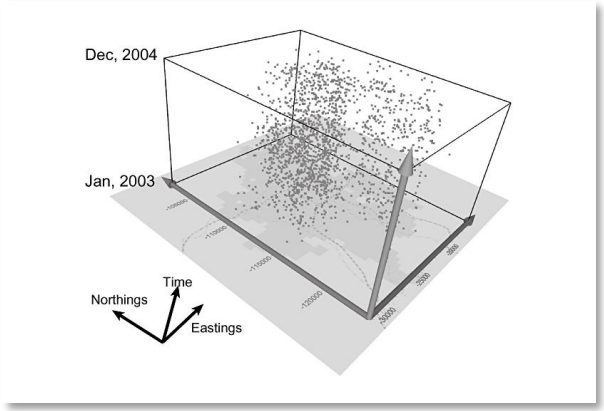
2024/7/16

# 目錄

# 緒論

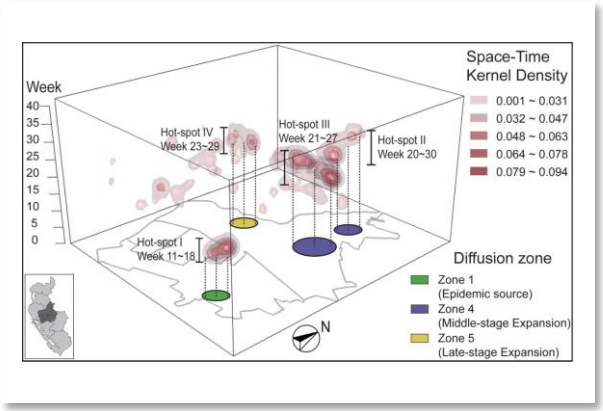
# 時空群聚所在位置為事件在時間和空間上密集發生的地區，亦即事件發生熱區

## 犯罪高風險區



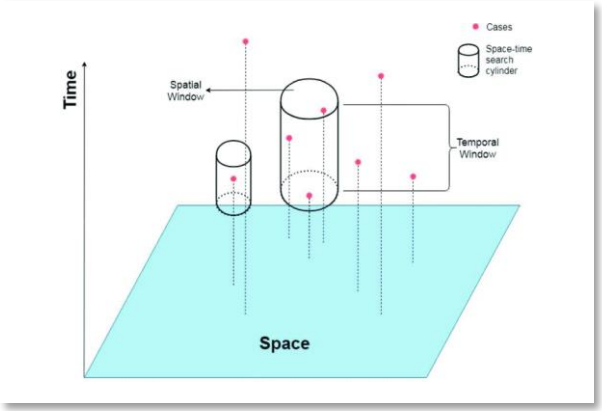

Song et al. (2018)

## 傳染病嚴重區




Kuo *et al.* (2018)

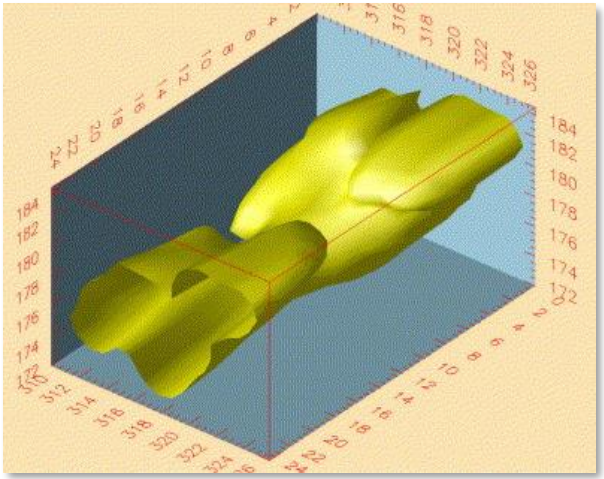
## 交通之易肇事區




Nakaya & Yano (2010)

# 現有時空群聚演算法在四個群聚結果特性上，已有成熟的方法研究

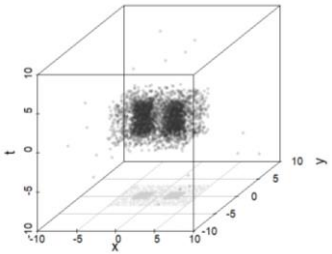
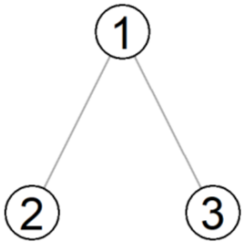
| 群聚結果特性 | 用途 | 可達成方法 |
|---|---|---|
| 總群聚數非經給定 | 可基於點事件分布而自動查找適合數量的群聚 | ST-KDE*<br>ST-DBSCAN**<br>ST-OPTICS*** |
| 排除雜訊點 | 排除不位於高密度範圍內的事件點進而更關注於熱區 | ST-Hierarchical Clustering****<br>ST-DBSCAN**<br>ST-OPTICS*** |
| 形狀任意 | 查找出非特定形狀的群聚進而更貼近更真實之群聚範圍 | ST-KDE*<br>ST-DBSCAN**<br>ST-OPTICS*** |
| 範圍明確 | 回應群聚查找目的後解決問題的資源之配置以及管理策略之制定 | SaTScan*****<br>ST-DBSCAN**<br>ST-OPTICS*** |



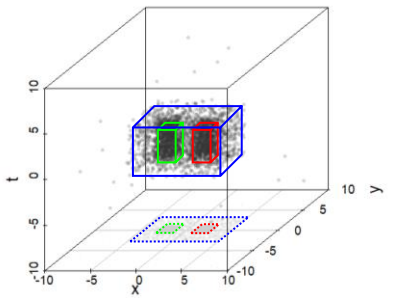
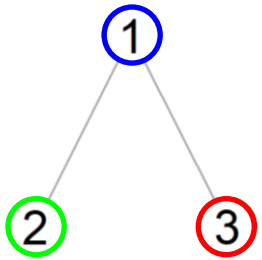
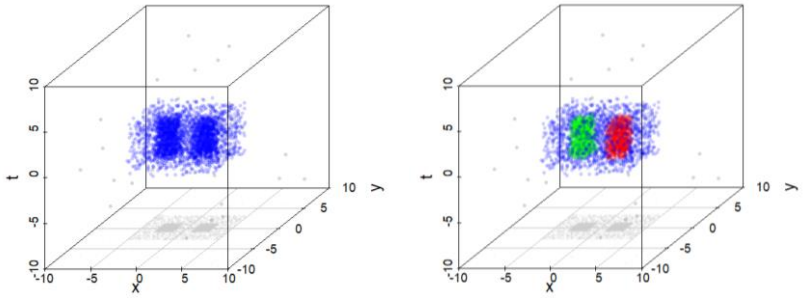
Brunsdon et al. (2007)

*(Brunsdon et al., 2007) **(Birant and Kut, 2007) ***(Agrawal et al., 2016) ****(Lamb et al., 2020) *****(Song et al., 2018)

# 現有時空群聚演算法在具密度差異與階層性兩個特性群聚結果上之方法，相對處於發展中之狀態

| 群聚結果特性 | 用途 | 空間群聚<br>可達成方法 | 時空群聚<br>嘗試方法 |
|---|---|---|---|
| 具有密度差異<br>  | 了解各群聚的密度<br>進而知悉不同群聚範圍內的事件嚴重程度 | KDE<br>(Silverman, 2018)<br>OPTICS<br>(Ankerst et al., 1999) | ST-KDE<br>(Brunsdon et al., 2007)<br>ST-OPTICS<br>(Agrawal et al., 2016) |
| 具有階層性<br>  | 同時了解同一母群聚之下群聚的相似性，又可以了解子群聚的特殊性 | OPTICS<br>(Ankerst et al., 1999) | ST-OPTICS<br>(Agrawal et al., 2016) |

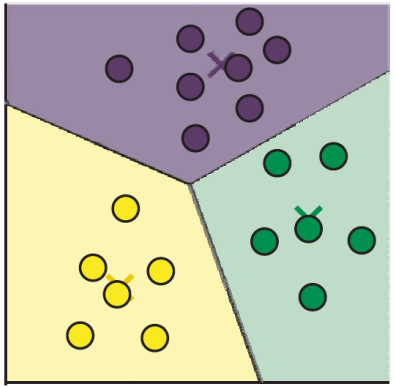
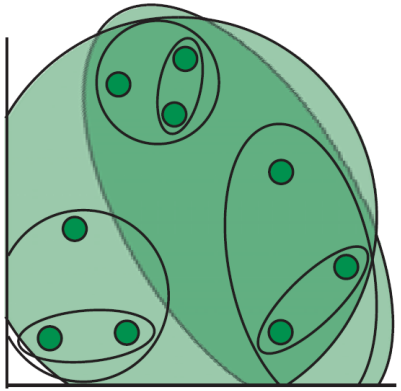
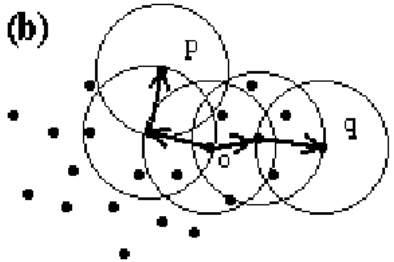
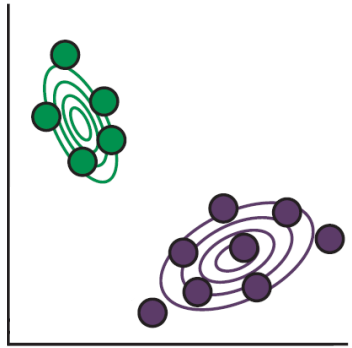
# 研究目的：發展基於 OPTICS* 之階層性時空群聚演算法，以達成三項目的

| | 研究目的（一） | 研究目的（二） | 研究目的（三） |
|---|---|---|---|
| 目的 | 可識別**時空密度斷層**位置，<br>以作為不同密度群聚之劃分依據 | 可識別出群聚間的<br>**階層性關係**，<br>以描述時空群聚結構 | 可**彈性**調整時空密度斷層定義之嚴格度，<br>以廣泛考量斷層存在的可能 |
| 示意圖 |  |  |  |

* (Ankerst et al., 1999)

# 文獻回顧

# 群聚演算法：在群聚識別概念大致可分為四種類型*，基於不同概念所識別的群聚結果，有不同的特性與適用情境

| 查找中心點 | 鄰近整合 | 基於密度 | 基於混合模型 |
|---|---|---|---|
|  |  |  |  |
| **k-means**<br>(MacQueen, 1967) | **hierarchical clustering**<br>(Ward Jr, 1963) | **KDE****<br>**DBSCAN*****<br>**OPTICS****** | **Gaussian mixture modeling**<br>(Stauffer and Grimson, 1999) |

*(Rhys, 2020) **(Silverman, 2018) ***(Ester et al., 1996) ****(Ankerst et al., 1999)

## 時空群聚演算法：對於時空資料進行特定的考量，將群聚演算法延伸發展而來，其識別概念同樣可分為四種類型，其所具有的時空群聚結果特性不同，而在使用上仍有其缺失

| | 查找中心點 | 鄰近整合 | 基於密度 | 基於混合模型 |
|---|---|---|---|---|
| 方法 | **fuzzy C-means** (Izakian et al., 2013) | **ST-Hierarchical Clustering** (Lamb et al., 2020) | **ST-KDE*** **ST-DBSCAN**** **ST-OPTICS***** | **STGMM**** **(ST-gaussian mixture model)** |
| 時空群聚結果特性 | 群聚中心明確的特色 | 層次結構明確<br>可以顯示不同層級群聚 | 形狀任意且可排除雜訊點 | 能發現重疊階層性群聚 |
| 缺失 | 對於非球形的群聚查找效果不佳 | 層級之選定缺乏依據 | 對於參數的設定較為敏感 | 計算複雜度較高 |

*(Brunsdon et al., 2007) **(Birant and Kut, 2007) ***(Agrawal et al., 2016) ****(Kim et al., 2012)

## 時空鄰近定義：代表事件點在時空上相似程度與鄰近程度的衡量，也影響群聚存在與否之判定。參考時空相依檢定中時空鄰近之定義，分為三個類型

| | 閾值切分 | 距離遞減 | 前 k 鄰近 |
|---|---|---|---|
| 時空鄰近定義 | 只有兩事件點時間與空間上皆小於給定之閾值時，彼此才具有鄰近關係 | 所有點間皆具有鄰近關係，鄰近程度則隨距離上升而遞減 | 各事件點對於鄰近與否的衡量會因為該點附近事件點密度的差異而有不同 |
| 時空相依檢定方法 | **Knox's Test*** <br> **Baker's Max Test**** <br> **Diggle's Test***** | **Mantel's Test** <br> (Mantel, 1967) | **Jacquez's k-NN Test** <br> (Jacquez, 1996) |

*(Knox and Bartlett, 1964) **(Baker, 1996) ***(Diggle et al., 1995)

## 基於密度之時空群聚演算法：在特定分析流程下，對具密度差異與階層性的群聚結構有識別潛力，但會因其發展時所參考的一般群聚演算法特性，而有時空群聚結構的識別限制

| | ST-KDE* | ST-DBSCAN** | ST-OPTICS*** |
|---|---|---|---|
| 識別潛力 | 在選取適當的多個**切分閾值**後，可以查找出**不同密度**之群聚結果，並可同時了解整體之**時空群聚結構**與**階層性**關係 | 在不同**參數設定**下，可查找出**不同密度**之群聚結果<br>透過**統整**多次群聚結果，可嘗試獲取整體之時空群聚結構以及群聚間的**階層性**關係 | 可透過**可及圖**與 **steepness 參數**設定，查找有密度差異的群聚，並統整出整體之時空群聚結構以及群聚間的階層性關係 |
| 識別限制 | 需**自行指定**切分之閾值，閾值之選定**缺乏依據**，是否需切分出新的子群聚缺乏**密度變異程度**的考量 | 需設定適當的參數值，於統整時需**自行歸納**群聚間的階層性關係 | 群聚查找結果受參數設定影響很大， steepness 參數切分過於嚴格，導致子群聚的切分與否**缺乏彈性**<br>研究***中最後一步驟階層性時空群聚結構中的小範圍群聚依相似性合併，**失去階層性的特性** |

*(Brunsdon et al., 2007) **(Birant and Kut, 2007) ***(Agrawal et al., 2016)

## 綜合評析：綜合過去研究成果，目前在時空群聚結構識別的領域中，尚面臨三大問題

| | 問題（一） | 問題（二） | 問題（三） |
|---|---|---|---|
| 問題敘述 | **密度各異群聚**識別 | **階層性時空群聚結構**識別 | **密度斷層**識別彈性問題 |
| 識別目標的意義 | 代表事件在各個時空範圍內的嚴重程度 | 代表子群聚有與母群聚類似的特性，但又同時進一步具備相對於母群聚更細的特徵 | 代表了密度的急遽變化處，其存在與否可作為是否應切分出新的子群聚之依據 |

# 研究方法

# 研究流程：三個階段，發展 HST-OPTICS 演算法、驗證演算法成效、提供演算法參數設定建議

## 第一階段：發展 HST-OPTICS 演算法

時空可及圖：

1. 定義時空資料點鄰近關係
2. 定義時空距離
3. 獲取時空可及圖

時空密度斷層與群聚識別：

1. 基於 OPTICS 陡度概念，定義搜尋窗大小 window size 以及可及分數差異 diff 兩個參數
2. 識別出密度斷層範圍
3. 劃分階層性時空群聚結構

## 第二階段：驗證演算法成效

1. 模擬七組群聚數、階層關係各異之時空群聚結構
2. 衡量研提演算法與既有演算法之群聚查找成效

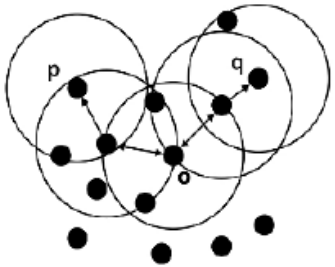
## 第三階段：提供演算法參數設定建議

1. 進行搜尋窗大小與可及分數差異兩個參數之敏感度分析
2. 提供演算法參數設定的建議

# 研提演算法參考 DBSCAN 中之參數定義、鄰近關係定義以及核心點的概念

## DBSCAN

density-based spatial clustering of application with noise (Ester et al., 1996)




- ✓ Eps: 空間搜尋範圍
- ✓ MinPts: 群聚最少點數

透過兩參數將點分為核心點、邊緣點和雜訊點
其中當點的 Eps 內的點數 > MinPts，則為**核心點**

可及、可連結 → 群聚

## ST-DBSCAN

spatial–temporal - density-based spatial clustering of application with noise (Birant and Kut, 2007)

引入了非空間屬性的距離參數，在衡量時空鄰近關係時考量了事件點間**空間**與**非空間**的相似性

## MST-DBSCAN

modified space–time - density-based spatial clustering of application with noise (Kuo et al., 2018)

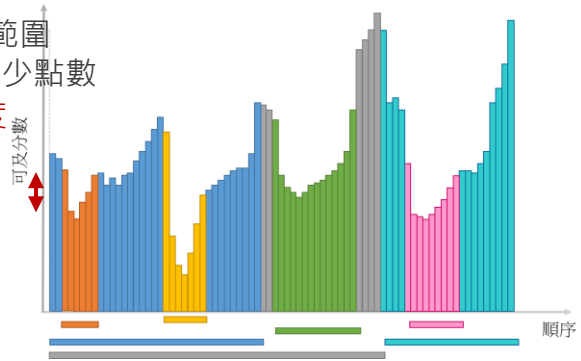
改進時空鄰近上的定義，設定新的參數，使得在判定事件點鄰近關係時，僅將**具備潛在傳染關係**的點視為潛在鄰近點

# 研提演算法參考 OPTICS 中將點排序繪製可及圖與透過陡度切分出群聚以識別階層性群聚之概念

## OPTICS

ordering points to identify the clustering structure
(Ankerst et al., 1999)

- ✓ Eps: 空間搜尋範圍
- ✓ MinPts: 群聚最少點數
- ✓ steepness: 陡度




定義核心距離、可及距離概念，逐一造訪事件點，
將點排序並記錄下可及分數，以獲取可及圖

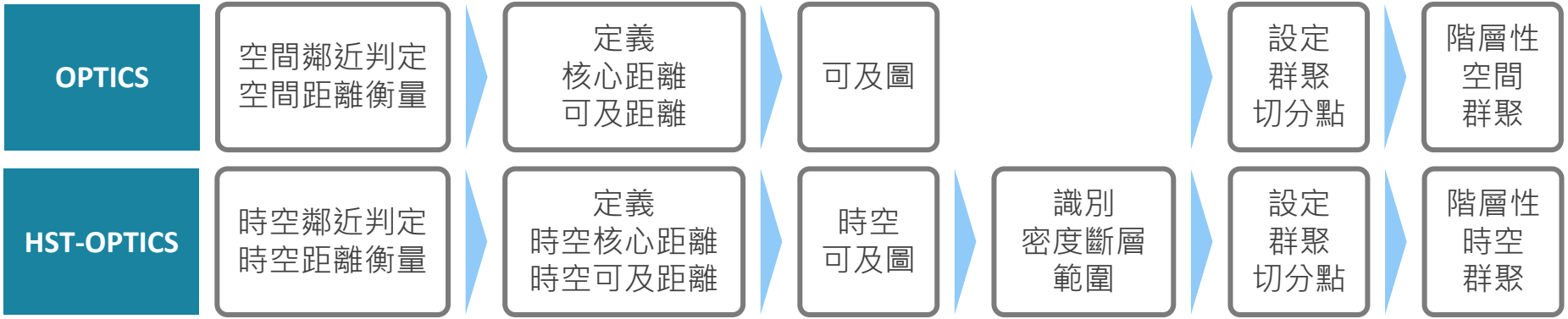
可及圖 → 階層性群聚

## ST-OPTICS

spatio-temporal - ordering points to identify clustering structure
(Agrawal et al., 2016)

新增非空間距離之參數
額外計算最小可及距離 minRD 與最大核心距離 maxCD 來取代了原先的 Eps 參數

## 研提演算法：HST-OPTICS 修正時空鄰近關係與時空距離衡量方式，基於時空可及圖，彈性地識別密度斷層，以獲取最終之時空群聚結構。所需參數共有七個




### 建立時空可及圖

HST-OPTICS 所使用的參數：

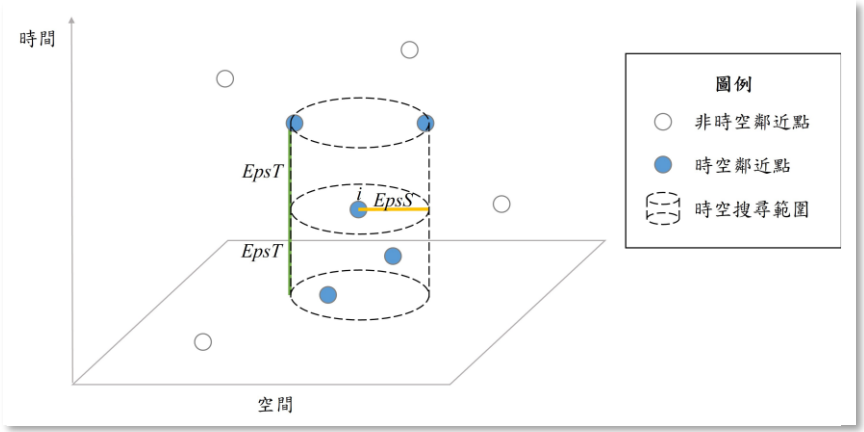
- ✓ EpsS: 空間搜尋範圍
- ✓ EpsT: 時間搜尋範圍
- ✓ MinPts: 群聚最少點數
- ✓ $W_S$ :空間距離權重
- ✓ $W_T$ :時間距離權重

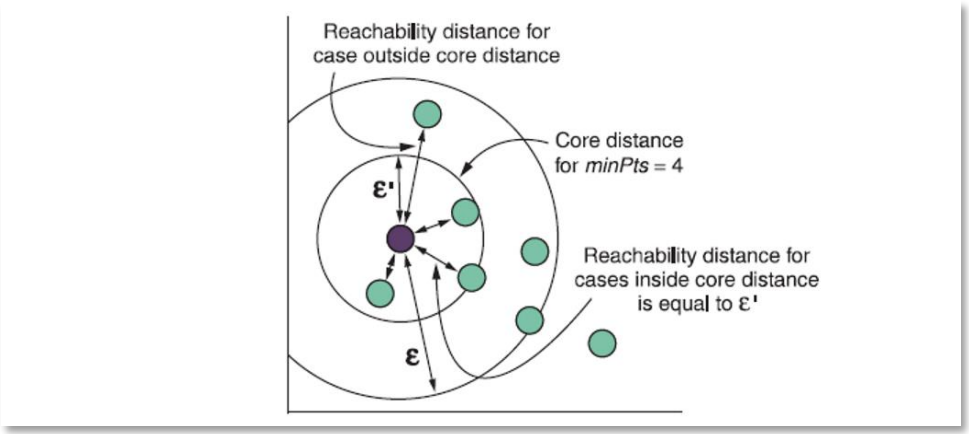
### 彈性地識別密度斷層與時空群聚結構

HST-OPTICS 所使用的參數：

- ✓ window size:搜尋窗大小
- ✓ diff:可及分數差異

# 時空可及圖：基於 OPTICS 中之概念進行調整，重新定義時空鄰近關係、時空距離、時空核心距離與時空可及距離




## 時空資料點鄰近關係

位於時空搜尋範圍內

## 核心點

鄰近點數 ≥ MinPts

## 時空核心距離

點至其第 MinPts 個鄰近點的時空距離

## 時空距離*

$$\sqrt{(W_S \times \text{空間距離})^2 + (W_T \times \text{時間差異})^2}$$

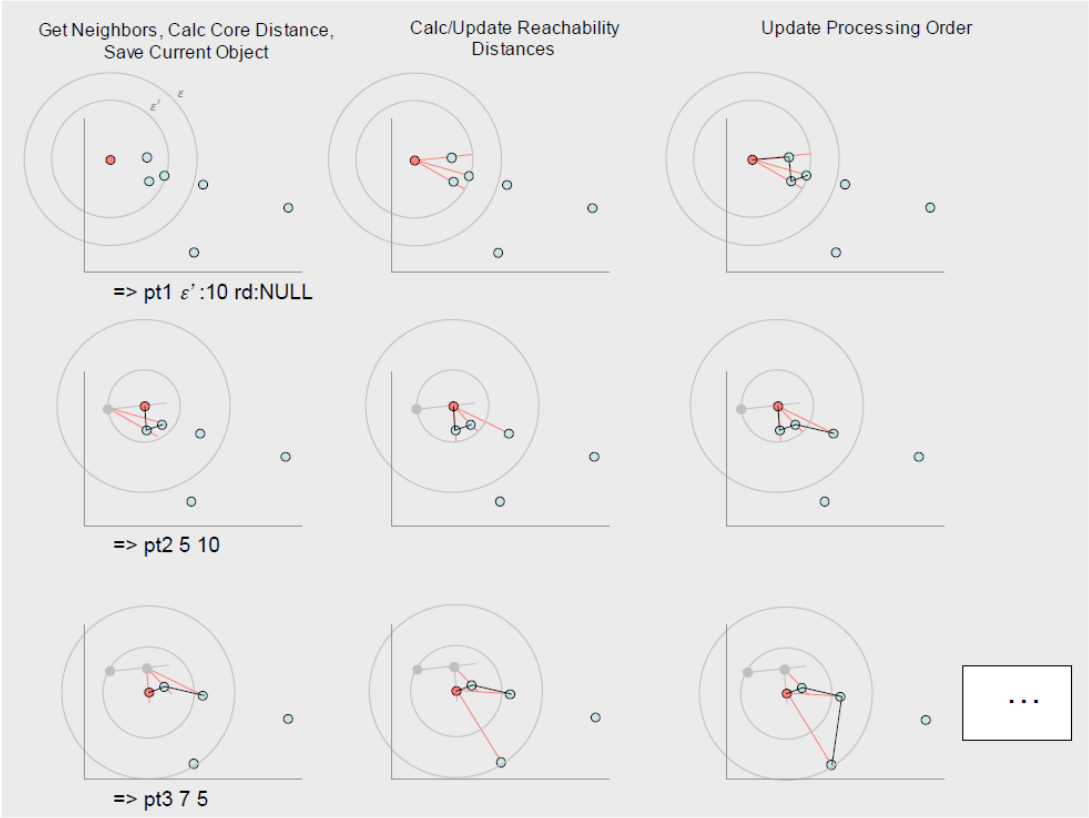
## 時空可及距離

點至其未訪鄰近點之時空距離，至小為時空核心距離

* 本研究中設定為兩權重皆為 1

註：以二維圖示意方便說明

# 時空可及圖：HST-OPTICS 在逐點造訪運作後，會記錄下各點之排序與對應之時空可及分數，繪製成時空可及圖




## 由編號 1 的點逐點造訪運作流程：

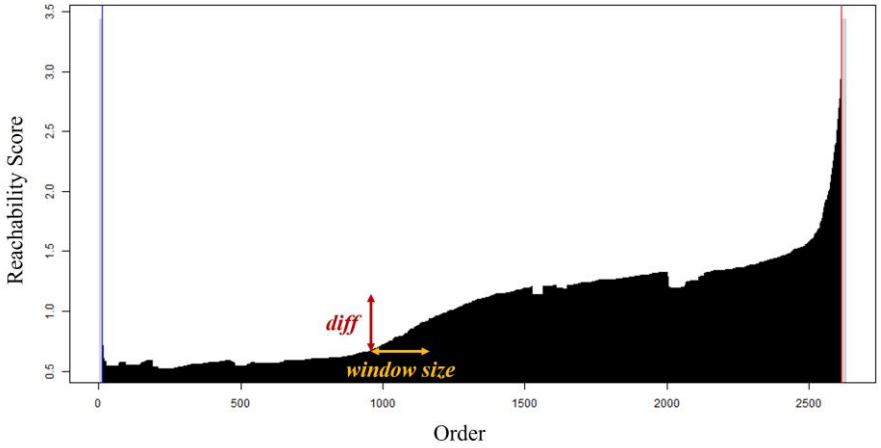
1. 紀錄該點既有**時空可及距離**中最小的數值當作**時空可及分數**（其中編號為 1 之事件點的時空可及分數不存在）

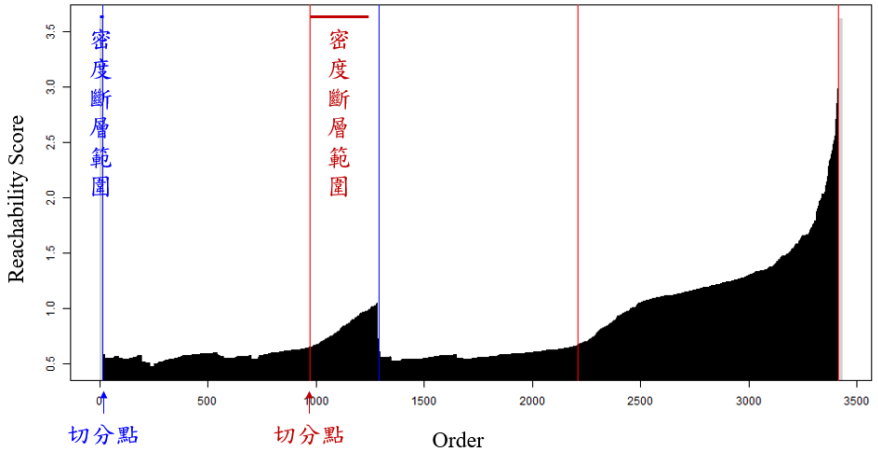
2. 計算**時空核心距離**，若該點為**核心點**，則計算它對時空搜尋範圍內所有點的時空可及距離

3. 按現有的時空可及距離由小到大排列，更新點的造訪順序

註：以二維圖示意方便說明

## 時空密度斷層與群聚識別：定義搜尋窗大小 window size 以及可及分數差異 diff 兩個參數，在時空可及圖中，更彈性地識別出密度斷層發生範圍，並設定出切分點以獲取時空群聚結構

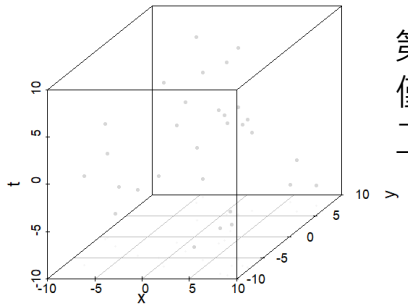


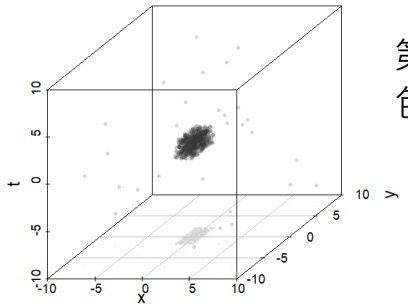

搜索各點排序前後**搜尋窗範圍 window_size** 內的點，是否有可及**分數差異 diff** 夠大的點，若有達成條件，則該點被視為位於**密度斷層發生範圍**內

密度斷層範圍中**最後**或**最前**排序的點被設定為群聚切分的**起始**與**終止**位置，以劃分出各群聚的範圍，並獲得整體之階層性時空群聚結構

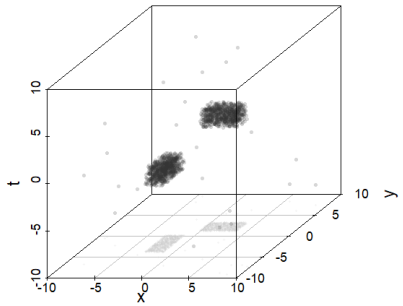
# 演算法成效驗證：模擬七組群聚數、階層關係不同之時空群聚結構，衡量研提演算法與既有演算法之群聚查找成效




第一組
僅包含雜訊點
二至七組亦包含此雜訊點

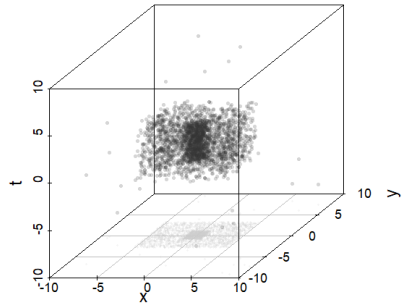



第二組
包含一個群聚

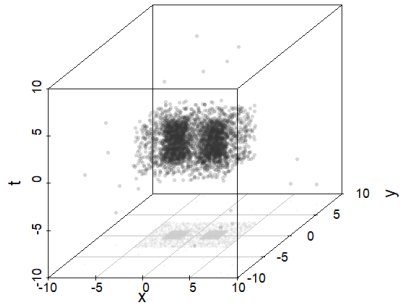



第三組
包含兩個群聚

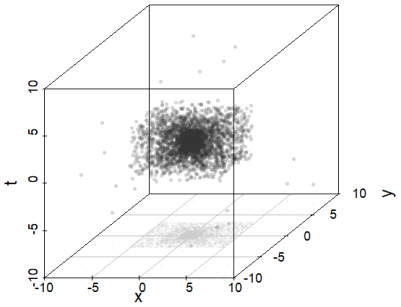
# 演算法成效驗證：模擬七組群聚數、階層關係各異之時空群聚結構，衡量研提演算法與既有演算法之群聚查找成效



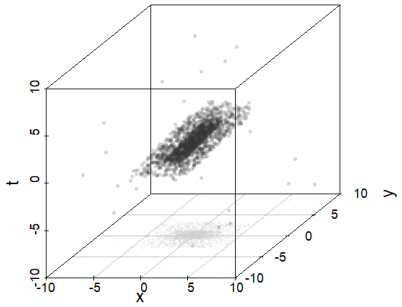

第四組
包含兩個群聚




第五組
包含三個群聚




第六組
包含三個群聚




第七組
包含兩個群聚
空間群聚範圍隨時間改變

## 密度斷層敏感度分析：由於 window size 和 diff 對識別密度斷層與群聚結構的結果有決定性影響，需進行敏感度分析以提供設定建議，共形成 $3 \times 3 = 9$ 組參數設定方式

| 數值大小 | 參數：搜尋窗大小 window size | 參數：可及分數差異 diff |
|---|---|---|
| 最小值 | 1 | 任兩非雜訊點時空可及分數差異的最小值 |
| **敏感度分析低值** | **20% × 非雜訊點數** | **0.5 × 非雜訊點可及分數標準差*** |
| **敏感度分析中值** | **30% × 非雜訊點數** | **非雜訊點可及分數標準差*** |
| **敏感度分析高值** | **40% × 非雜訊點數** | **2 × 非雜訊點可及分數標準差*** |
| 最大值 | 非雜訊點的總數量 | 任兩非雜訊點時空可及分數差異的最大值 |

* 非雜訊點可及分數之標準差，代表任意非雜訊點與鄰近其中一個核心點的距離相對於平均值的變異量

# 研究結果

# 時空群聚結果比較：HST-OPTICS 相對於 ST-DBSCAN，除了可以排除雜訊點、查找出不同範圍的群聚外，亦可查找出具有階層關係之時空群聚結構

(a) ST-DBSCAN (b) HST-OPTICS

Figure 4.1: 第一組模擬資料時空群聚結果

(a) ST-DBSCAN (b) HST-OPTICS

Figure 4.2: 第二組模擬資料時空群聚結果

(a) ST-DBSCAN (b) HST-OPTICS

Figure 4.3: 第三組模擬資料時空群聚結果

(a) ST-DBSCAN (b) HST-OPTICS

Figure 4.4: 第四組模擬資料時空群聚結果

(a) ST-DBSCAN (b) HST-OPTICS

Figure 4.5: 第五組模擬資料時空群聚結果

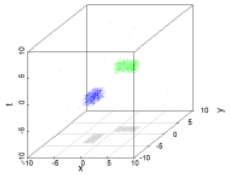
(a) ST-DBSCAN (b) HST-OPTICS

Figure 4.6: 第六組模擬資料時空群聚結果

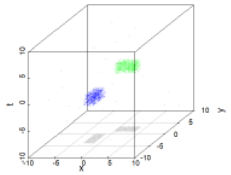
(a) ST-DBSCAN (b) HST-OPTICS

Figure 4.7: 第七組模擬資料時空群聚結果

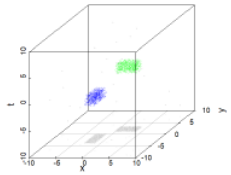
# 密度斷層敏感度分析結果：七組模擬資料皆包含九組參數設定下的群聚結果



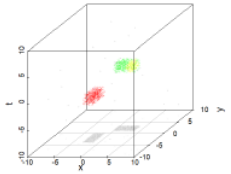
(a) *window size* 低, *diff* 低



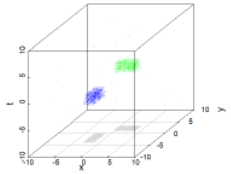
(b) *window size* 中, *diff* 低



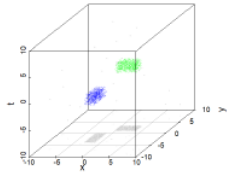
(c) *window size* 高, *diff* 低



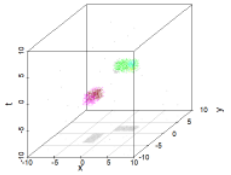
(d) *window size* 低, *diff* 中



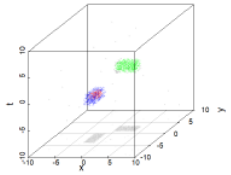
(e) *window size* 中, *diff* 中



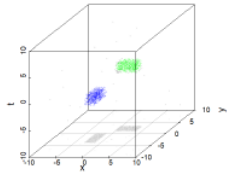
(f) *window size* 高, *diff* 中



(g) *window size* 低, *diff* 高



(h) *window size* 中, *diff* 高

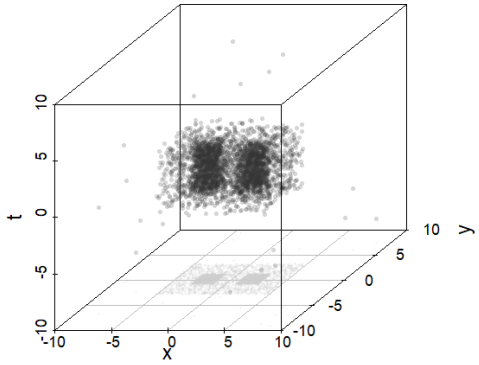


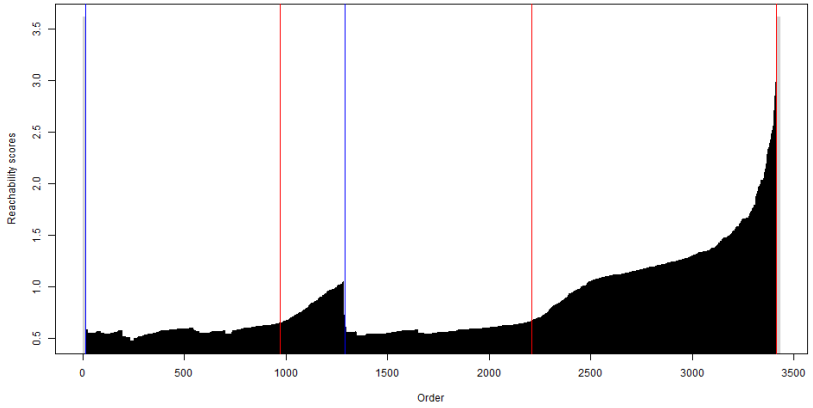
(i) *window size* 高, *diff* 高

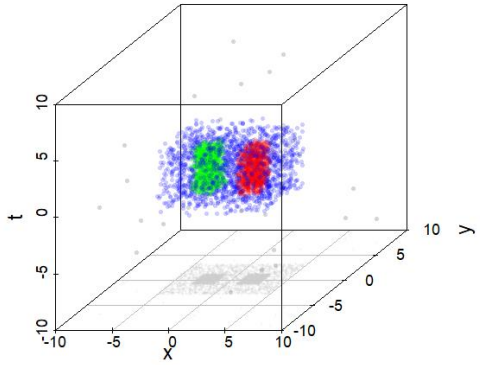
Figure B.10: 第三組時空模擬資料在各參數設定下之 HST-OPTICS 群聚圖

# 討論

# 時空群聚結構與密度斷層：本研究所研提之 HST-OPTICS 演算法因能夠識別時空密度斷層，有助於點事件的階層性時空群聚結構之建立




事件點分布 → 時空可及圖 → 時空密度斷層範圍 → 時空群聚切分點 → 時空群聚結構

# 參數設定

## 參數設定意義

兩個參數的設定，主要依據使用者對群聚寬限範圍和密度差異的期望

| 對密度差異的要求 | 搜尋窗大小 window size 參數設定值 | 可及分數差異 diff 參數設定值 | 群聚切分點意義 |
|---|---|---|---|
| **嚴格** | 小 | 大 | 代表較為區域化的密度差異發生處 |
| **寬鬆** | 大 | 小 | 反映大範圍中密度斷層存在的情況 |

## 未來應用參數設定

根據資料本身的可能群聚結構來設定參數

| 事件點分布傾向 | 需識別之時空群聚細緻度 |
|---|---|
| 出現小範圍內的密度變化 | **高** |
| 在大範圍內才具有密度變化 | **低** |

# 結論

# 結論

## 本篇研究

- 新的基於密度群聚演算法 HST-OPTICS
- 放寬 OPTICS 中陡度之識別嚴格度
- 識別密度斷層以劃分出群聚範圍
- 能識別包含不同密度之階層性時空群聚結構

## 未來研究

- 實務應用
- 演算法效能提高
- 參數指引
  - 識別密度斷層：
    搜尋窗大小 window_size、可及分數差異 diff
  - 計算時空距離：
    空間權重$W_S$、時間權重$W_T$

# Thank you for listening.