5. 事件點的空間群聚(2)

Spatial Point Clustering

https://ceiba.ntu.edu.tw/1062_Geog5016

授課教師：溫在弘
E-mail: wenthung@ntu.edu.tw

Density-based Clustering: DBSCAN

dbscan: Fast Density-based Clustering with R

Michael Hahsler
Southern Methodist University

Matthew Piekenbrock
Wright State University

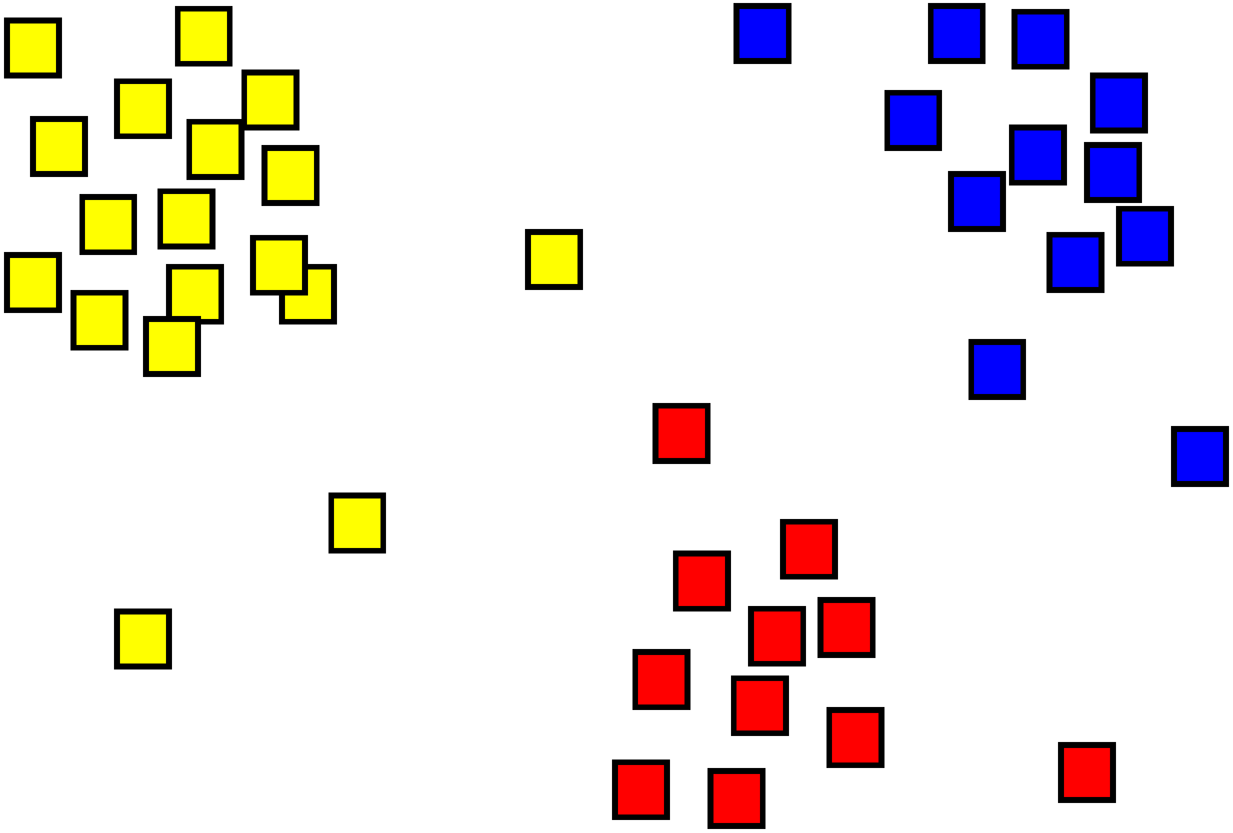
Derek Doran
Wright State University

Abstract

This article describes the implementation and use of the R package **dbscan**, which provides complete and fast implementations of the popular density-based clustering algorithm DBSCAN and the augmented ordering algorithm OPTICS. Compared to other implementations, **dbscan** offers open-source implementations using C++ and advanced data structures like k-d trees to speed up computation. An important advantage of this implementation is that it is up-to-date with several primary advancements that have been added since their original publications, including artifact corrections and dendrogram extraction methods for OPTICS. Experiments with **dbscan**'s implementation of DBSCAN and OPTICS compared and other libraries such as FPC, ELKI, WEKA, PyClustering, SciKit-Learn and SPMF suggest that **dbscan** provides a very efficient implementation.

Keywords: DBSCAN, OPTICS, Density-based Clustering, Hierarchical Clustering.

DBSCAN: Density Based Spatial Clustering of Applications with Noise



DBSCAN: Concepts

Definition 1. ϵ-Neighborhood. *The ϵ-neighborhood, $N_\epsilon(p)$, of a data point p is the set of points within a specified radius ϵ around p.*

$$N_\epsilon(p) = \{q \mid d(p, q) < \epsilon\}$$

where d is some distance measure and $\epsilon \in \mathbb{R}^+$. Note that the point p is always in its own ϵ-neighborhood, i.e., $p \in N_\epsilon(p)$ always holds.

Following this definition, the size of the neighborhood $|N_\epsilon(p)|$ can be seen as a simple un-normalized kernel density estimate around p using a uniform kernel and a bandwidth of ϵ. DBSCAN uses $N_\epsilon(p)$ and a threshold called *minPts* to detect dense regions and to classify the points in a data set into **core**, **border**, or **noise** points.

DBSCAN: 1. Defining the Neighborhood

Definition 1. ϵ-Neighborhood. *The ϵ-neighborhood, $N_\epsilon(p)$, of a data point p is the set of points within a specified radius ϵ around p.*

$$N_\epsilon(p) = \{q \mid d(p, q) < \epsilon\}$$

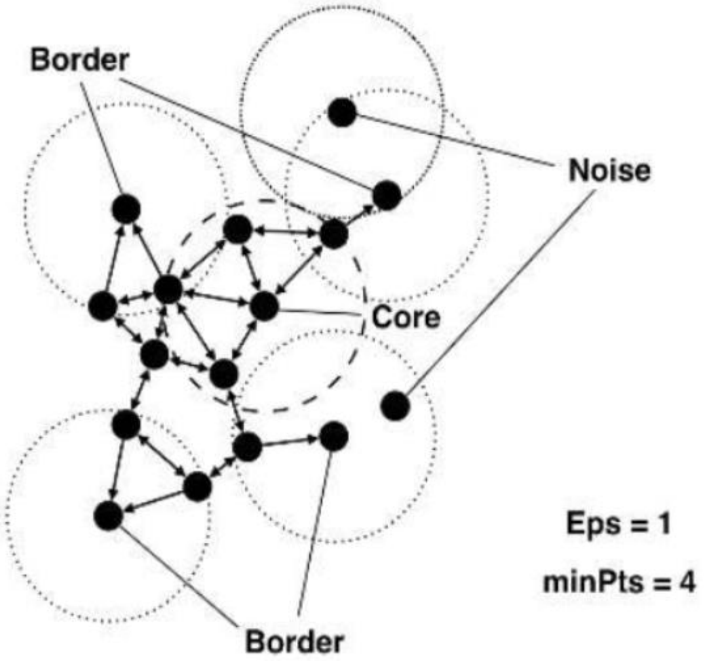
where d is some distance measure and $\epsilon \in \mathbb{R}^+$. Note that the point p is always in its own ϵ-neighborhood, i.e., $p \in N_\epsilon(p)$ always holds.

Following this definition, the size of the neighborhood $|N_\epsilon(p)|$ can be seen as a simple un-normalized kernel density estimate around p using a uniform kernel and a bandwidth of ϵ. DBSCAN uses $N_\epsilon(p)$ and a threshold called *minPts* to detect dense regions and to classify the points in a data set into **core**, **border**, or **noise** points.

2. Define Point Classes

Definition 2. Point classes. *A point $p \in D$ is classified as*

- *a* **core point** *if $N_\epsilon(p)$ has high density, i.e., $|N_\epsilon(p)| \geq minPts$ where $minPts \in \mathbb{Z}^+$ is a user-specified density threshold,*
- *a* **border point** *if p is not a core point, but it is in the neighborhood of a core point $q \in D$, i.e., $p \in N_\epsilon(q)$, or*
- *a* **noise point**, *otherwise.*



3. Density-reachable and connected

Definition 3. Directly density-reachable. *A point $q \in D$ is directly density-reachable from a point $p \in D$ with respect to ϵ and minPts if, and only if,*

1. $|N_\epsilon(p)| \geq minPts$, *and*
2. $q \in N_\epsilon(p)$.

That is, p is a core point and q is in its ϵ-neighborhood.

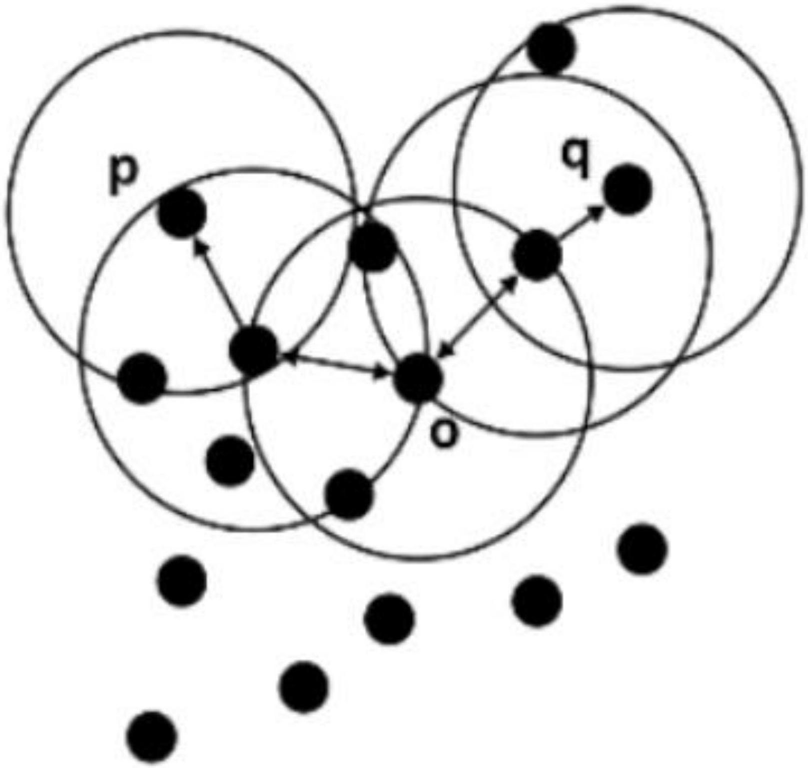
Definition 4. Density-reachable. *A point p is density-reachable from q if there exists in D an ordered sequence of points $(p_1, p_2, ..., p_n)$ with $q = p_1$ and $p = p_n$ such that $p_i + 1$ directly density-reachable from p_i $\forall\ i \in \{1, 2, ..., n-1\}$.*

Definition 5. Density-connected. *A point $p \in D$ is density-connected to a point $q \in D$ if there is a point $o \in D$ such that both p and q are density-reachable from o.*

A Cluster:

1. **Maximality**: *If $p \in C$ and q is density-reachable from p, then $q \in C$; and*
2. **Connectivity**: $\forall\ p, q \in C$, *p is density-connected to q.*

Density-reachable and connected



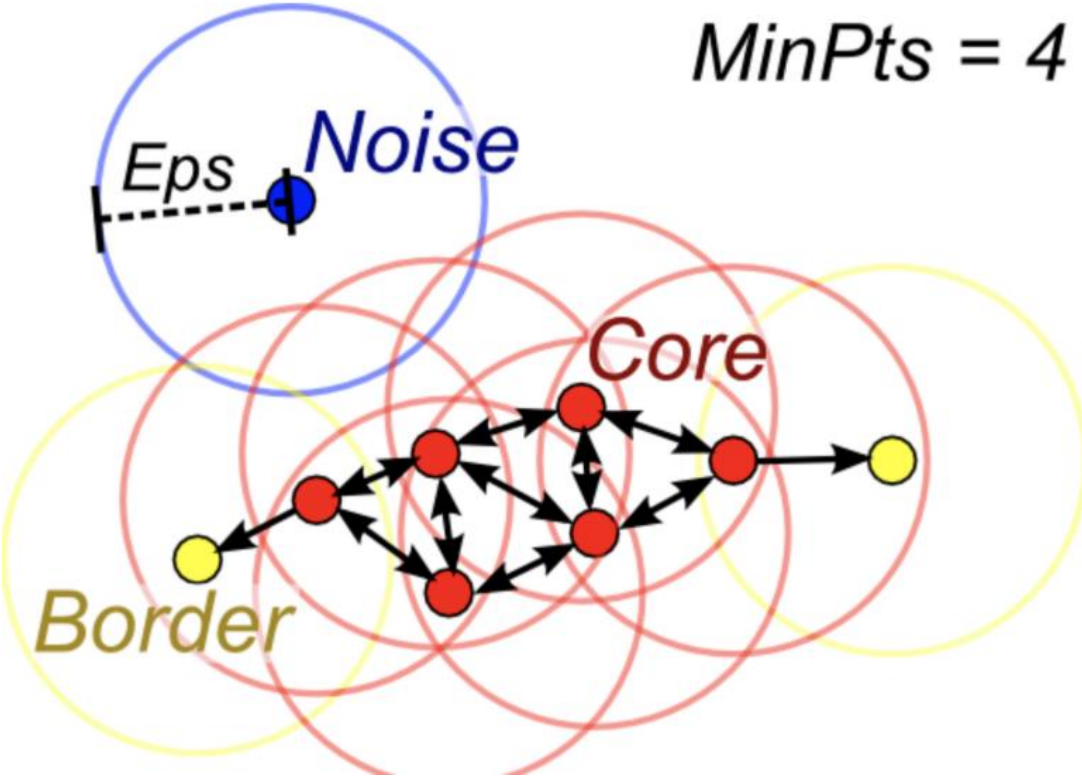
p and q are *density-reachable* from *o*

Therefore p and q are *density-connected*

Eps = 1

minPts = 4

DBSCAN: Identifying Clusters

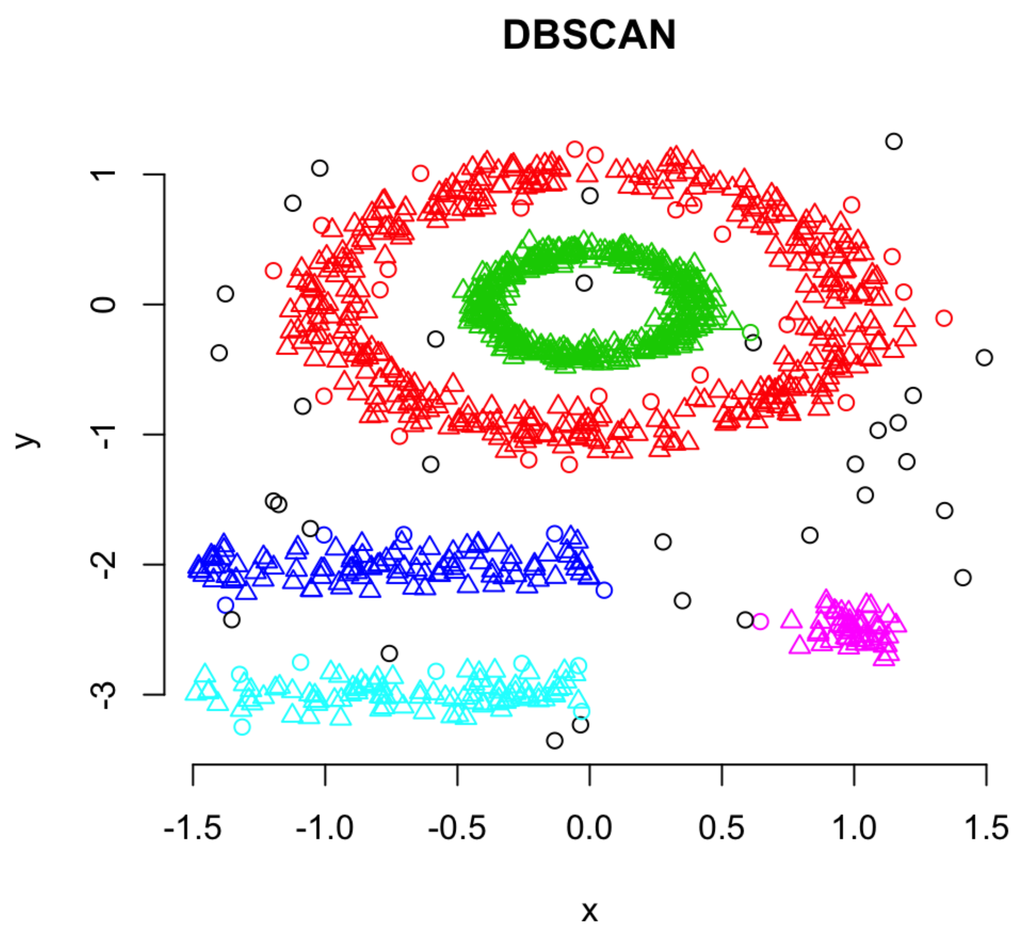


Red: Core Points

Yellow: Border points. Still part of the cluster because it's within epsilon of a core point, but not does not meet the min_points criteria

Blue: Noise point. Not assigned to a cluster

DBSCAN: Advantages



DBSCAN: Disadvantages

- Does not work well when dealing with clusters of varying densities. While DBSCAN is great at separating *high* density clusters from *low* density clusters, DBSCAN struggles with clusters of similar density.

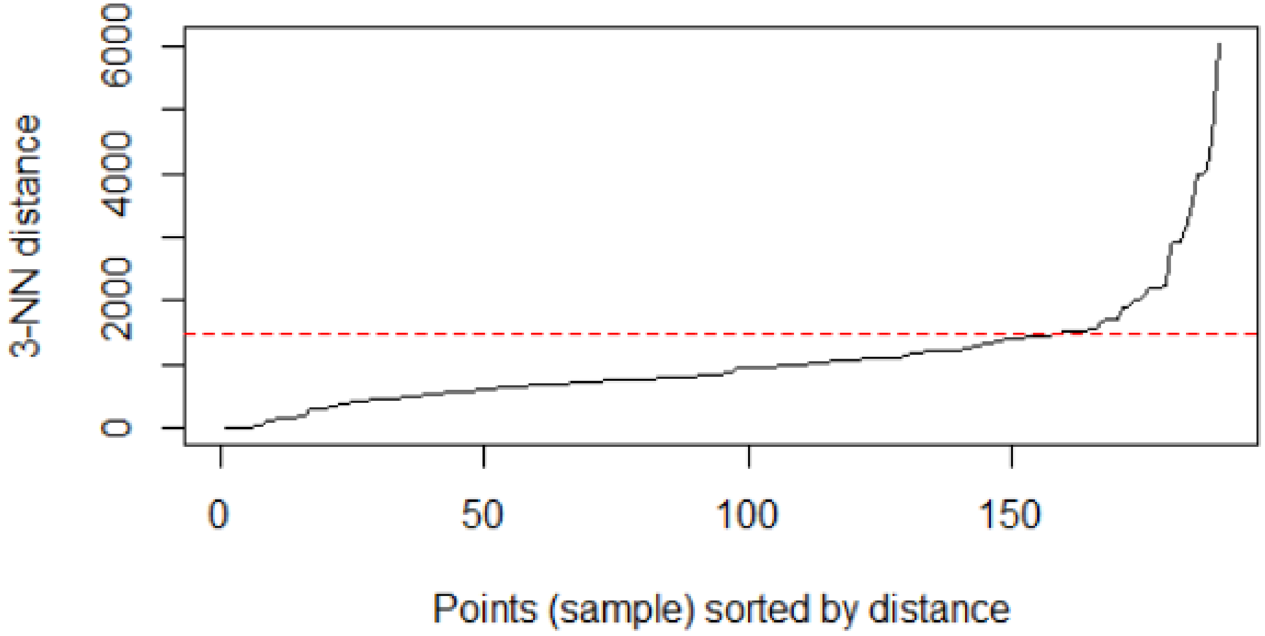
DBSCAN in R

```
install.packages("dbscan")
library("dbscan")

Pts0 <- cbind(data[,2], data[,3])



res <- dbscan(Pts0, eps = 1500, minPts = 3)
```

How to determine searching radius

K-nearest neighbor (k-NN) distance

```
kNNdistplot(Pts0, k = 3)
```

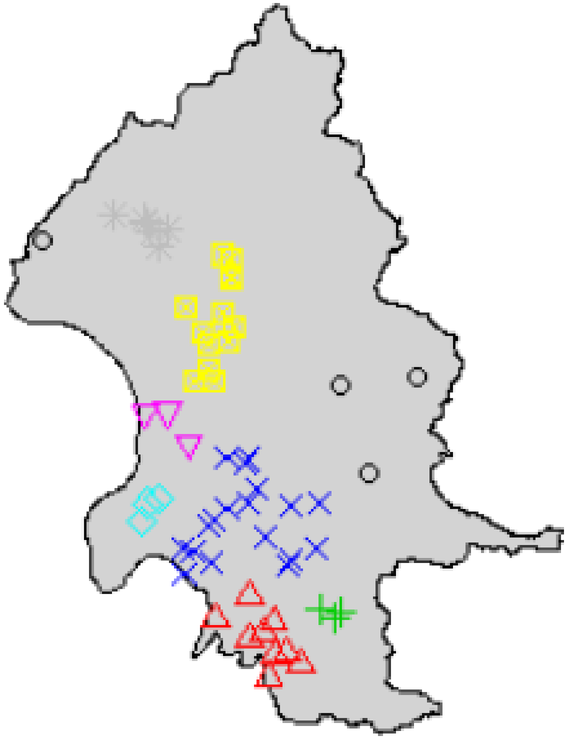


DBSCAN results

```
DBSCAN clustering for 63 objects.
Parameters: eps = 1500, minPts = 3
The clustering contains 7 cluster(s) and 4 noise points.

 0  1  2  3  4  5  6  7
 4  9  4 19  3  3 14  7

Available fields: cluster, eps, minPts
```

Plotting DBSCAN results

```
polymap(Pts_bnd, col="lightgray")
pointmap(Pts0, col = res$cluster + 1, pch = res$cluster + 1, add=T)
```



本週作業

圖資：台北市速食店 Tpe_Fastfood.shp

- 1. 參考 Reading_Dual.KDE.pdf 這篇論文關於 market dominance的定義，用dual KDE分析台北市 MIC 或 KFC 市場主導的空間分布。

- 2. 利用DBSCAN找出 MIC 與 KFC 的空間群聚。並討論不同參數設定 (eps, minPts)，對於群聚結果的敏感性。