Geographic Information System

Spatial Statistics II

Dr. Chan, Chun-Hsiang
Department of Geography
National Taiwan Normal University

Photo credit: ChatGPT

Outline

- Spatial Data :: Global vs Local Patterns
- Spatial Analysis :: A Local View
- Anselin Local Moran's I (LISA)
- Hot Spot Analysis (Getis-Ord Gi*)
- Density-based Clustering
- Spatial Outlier Detection
- False Discovery Rate Correction
- Multivariate Clustering
- Machine Learning :: Clustering
- Lab#01 Physical Meanings

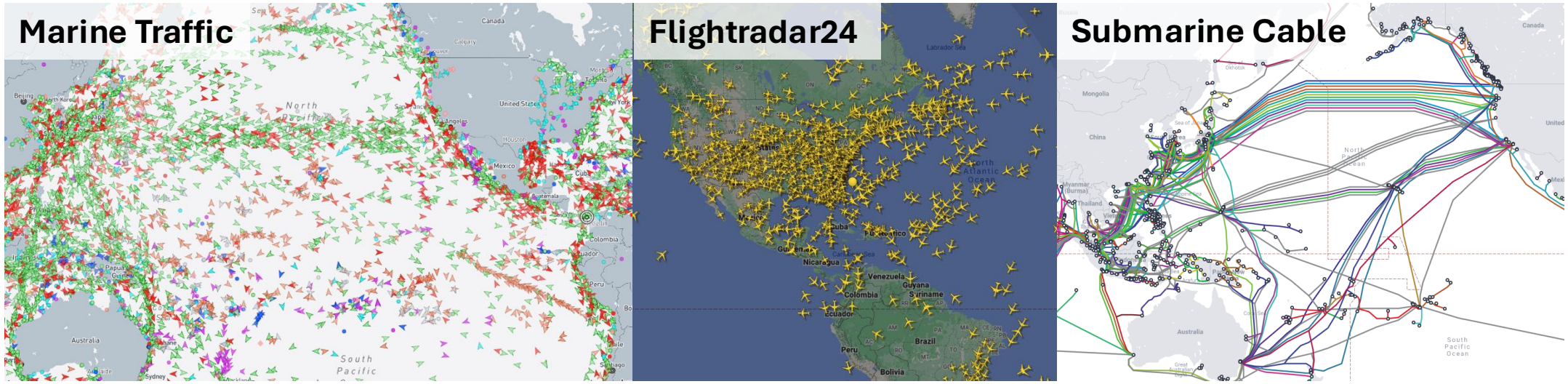
Spatial Data :: Global vs Local Patterns

- We already introduced several quantitative methods to depict the central tendency and spatial distribution of spatial data.
- However, these functions only describe a general/ global data tendency rather than a local view.
- Why do we need to investigate a spatial data distribution from a local view?
- Take a look at the right-hand-side picture.

source: https://www.bloomberg.com/news/storythreads/2020-06-12/covid-19-s-second-assault-on-america-is-underway

Spatial Analysis :: A Local View

- ArcGIS Pro provides a Mapping Clusters toolset containing tools that perform cluster analysis to identify the locations of statistically significant hot spots, cold spots, spatial outliers, and similar features or zones. These tools are useful when action is needed based on the location of one or more clusters.



Spatial Analysis :: A Local View

Functions	Definition
Anselin Local Moran’s I	Given a set of weighted features, identifies statistically significant hot spots, cold spots, and spatial outliers using the Anselin Local Moran's I statistic.
Hot Spot Analysis (Getis-Ord Gi*)	Given a set of weighted features, identifies statistically significant hot spots and cold spots using the Getis-Ord Gi* statistic.
Density-based Clustering	Finds clusters of point features within surrounding noise based on their spatial distribution. Time can also be incorporated to find space-time clusters.
Spatial Outlier Detection	Identifies global or local spatial outliers in point features.
Similarity Search	Identifies which candidate features are most similar or most dissimilar to one or more input features based on feature attributes.
Multivariate Clustering	Finds natural clusters of features based solely on feature attribute values.

Source: https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/an-overview-of-the-mapping-clusters-toolset.htm

Anselin Local Moran's I (LISA)

- Given a set of weighted features, identifies statistically significant hot spots, cold spots, and spatial outliers using the Anselin Local Moran's I statistic.

- A high positive z-score for a feature indicates that the surrounding features have similar values (either high or low values). The COType field will be HH for a statistically significant cluster of high values and LL for a statistically significant cluster of low values.
- A low negative z-score for a feature indicates a statistically significant spatial data outlier. The COType field will indicate if the feature has a high value and is surrounded by features with low values (HL) or if the feature has a low value and is surrounded by features with high values (LH).

Source: https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/cluster-and-outlier-analysis-anselin-local-moran-s.htm

Anselin Local Moran's I (LISA)

- When chordal distances are used in the analysis, the **Distance Band** or **Threshold Distance** parameter, if specified, should be given in meters.
- **Fixed distance band:** Uses a Distance Band or Threshold Distance and it will ensure each feature has at least one neighbor.
- **Inverse distance or Inverse distance squared**
- When zero is entered for the Distance Band or Threshold Distance parameter, all features are considered neighbors of all other features; when this parameter is left blank, the default distance will be applied.

Source: https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/cluster-and-outlier-analysis-anselin-local-moran-s.htm

Anselin Local Moran's I (LISA)

- The Local Moran's I statistic of spatial association is given as:

$$I_i = \frac{x_i - \bar{X}}{S_i^2} \sum_{j=1, j \neq i}^{n} w_{i,j}\left(x_j - \bar{X}\right)$$

where x_i is an attribute for feature i, $\bar{X}$ is the mean of the corresponding attribute, $w_{i,j}$ is the spatial weight between feature i and j, and:

$$S_i^2 = \frac{\sum_{j=1, j \neq i}^{n}\left(x_j - \bar{X}\right)^2}{n - 1}$$

with n equating to the total number of features.

Source: https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/cluster-and-outlier-analysis-anselin-local-moran-s.htm

Anselin Local Moran's I (LISA)

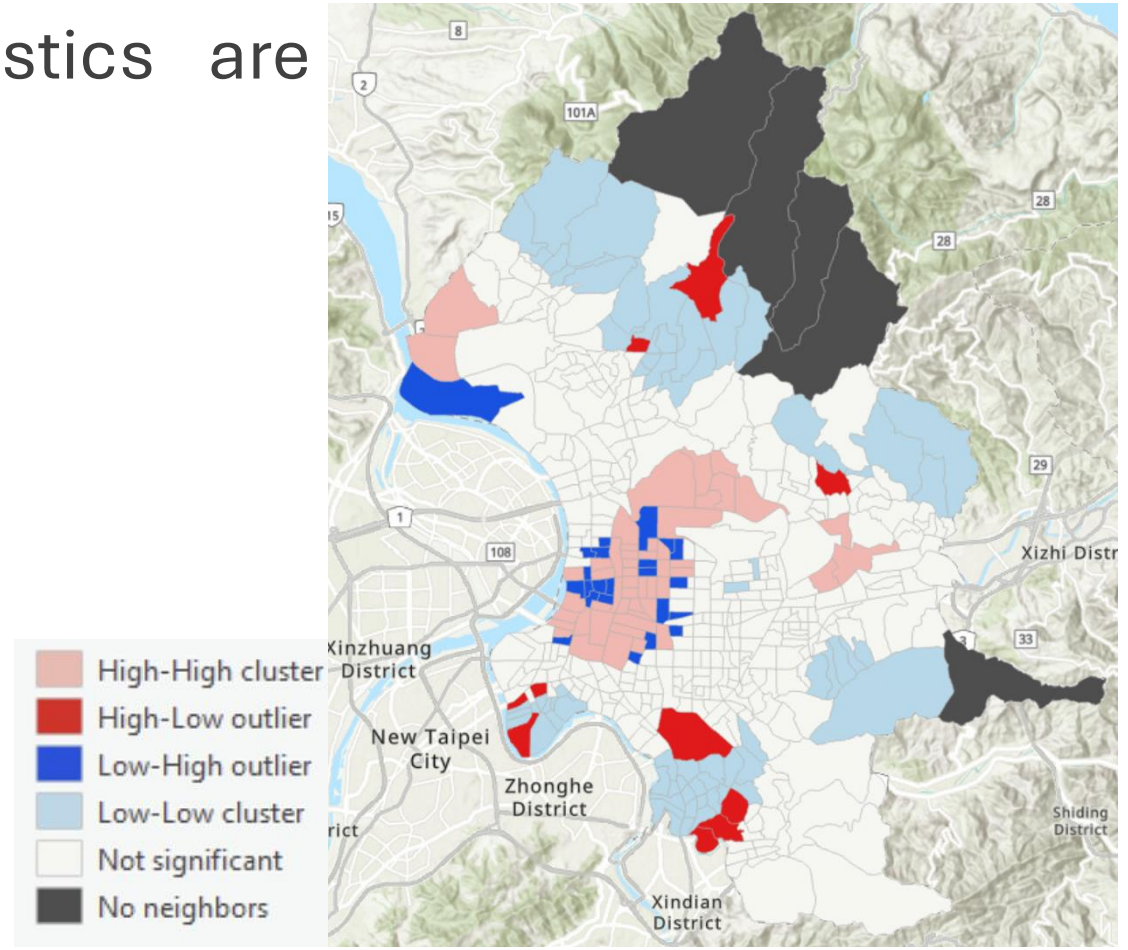
- The $z_{I_i} - score$ for the statistics are computed as:

$$Z_{I_i} = \frac{I_i - E[I_i]}{\sqrt{V[I_i]}}$$

where:

$$E[I_i] = -\frac{\sum_{j=1, j\neq i}^{n} w_{i,j}}{n-1}$$

$$V[I_i] = E\left[I_i^2\right] - E[I_i]^2$$



Source: https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/cluster-and-outlier-analysis-anselin-local-moran-s.htm

Hot Spot Analysis (Getis-Ord Gi*)

- Given a set of weighted features, identifies statistically significant hot spots and cold spots using the Getis-Ord Gi* statistic.
- This tool identifies statistically significant spatial clusters of high values (hot spots) and low values (cold spots). It creates an output feature class with a z-score, *p*-value, and confidence level bin field (Gi_Bin) for each feature in the input features.
- In effect, they indicate whether the observed spatial clustering of high or low values is more pronounced than would be expected in a random distribution of those same values.

Source: https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/hot-spot-analysis.htm

Hot Spot Analysis (Getis-Ord Gi*)

- The Getis-Ord local statistic is given as:

$$G_i^* = \frac{\sum_{j=1}^{n} w_{i,j}x_j - \bar{X}\sum_{j=1}^{n} w_{i,j}}{S\sqrt{\frac{\left[n\sum_{j=1}^{n} w_{i,j}^2 - \left(\sum_{j=1}^{n} w_{i,j}\right)^2\right]}{n-1}}}$$

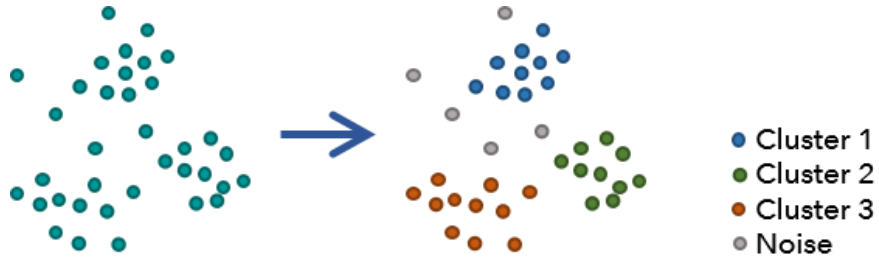
- Where x_j is the attribute value for feature j, $w_{i,j}$ is the spatial weight between feature i and j, n is equal to the total number of features and:

$$\bar{X} = \frac{1}{n}\sum_{j=1}^{n} x_j \, , S = \sqrt{\frac{1}{n}\sum_{j=1}^{n} x_j^2 - (\bar{X})^2}$$

- The G_i^* statistic is a $z - score$ so no further calculations are required.

Source: https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/hot-spot-analysis.htm

Density-based Clustering



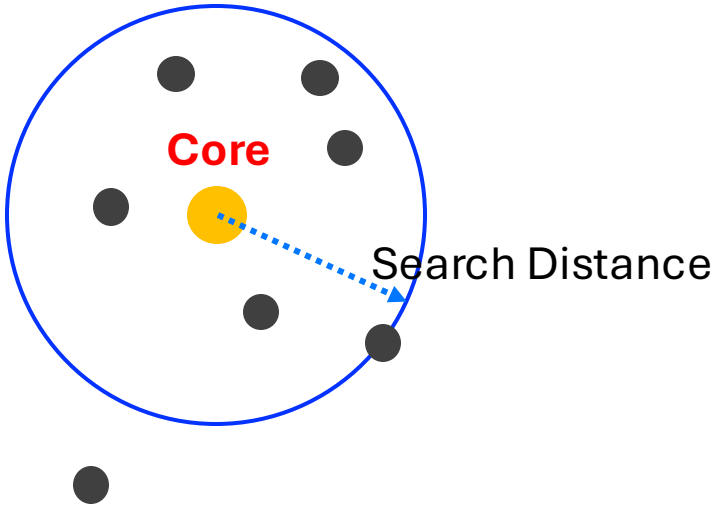
- Finds clusters of point features within surrounding noise based on their spatial distribution. Time can also be incorporated to find space-time clusters.
- This tool extracts clusters from the **Input Point Features** parameter value and identifies any surrounding noise.

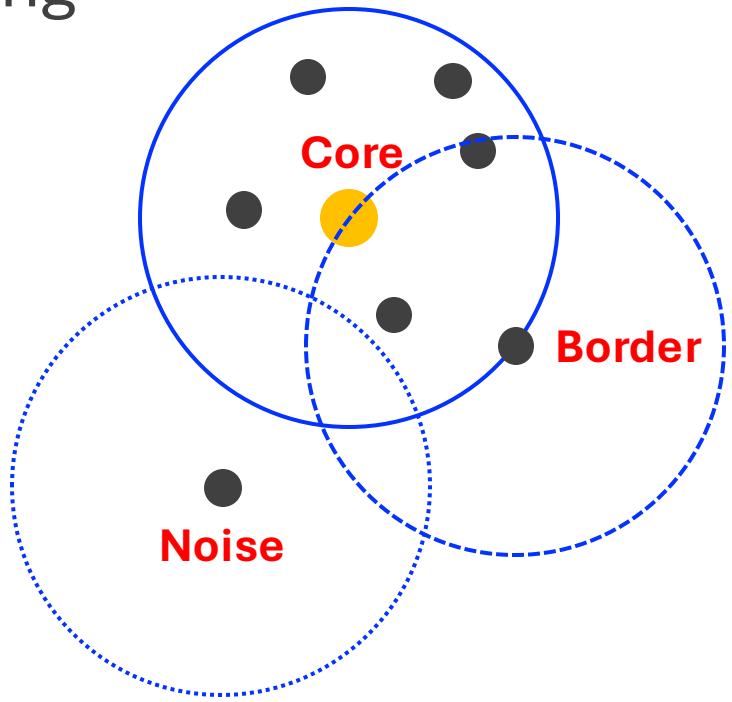
Algorithm	Characteristics
DBSCAN	finds clusters of points that are in close proximity based on a specified search distance.
HDBSCAN	finds clusters of points similar to DBSCAN but uses varying distances, allowing for clusters with varying densities based on cluster probability.
OPTICS	orders the input points based on the smallest distance to the next point.

Source: https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/densitybasedclustering.htm

Density-based Clustering

The core-distance is related to the Search Distance parameter, which is used by both the Defined distance (DBSCAN) and Multi-scale (OPTICS) clustering methods.





1 Core Point

The number of points within the search distance of a **core point** is higher than the minimum points.

2 Border Point

A **border point** is a point that is within the search distance of a core point but does not itself have the minimum number of features within the search distance.

3 Noise Point

If a point does not have the minimum number of features within the search distance and is not within the search distance of another core point

Source: https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/how-density-based-clustering-works.htm

Density-based Clustering :: OPTICS

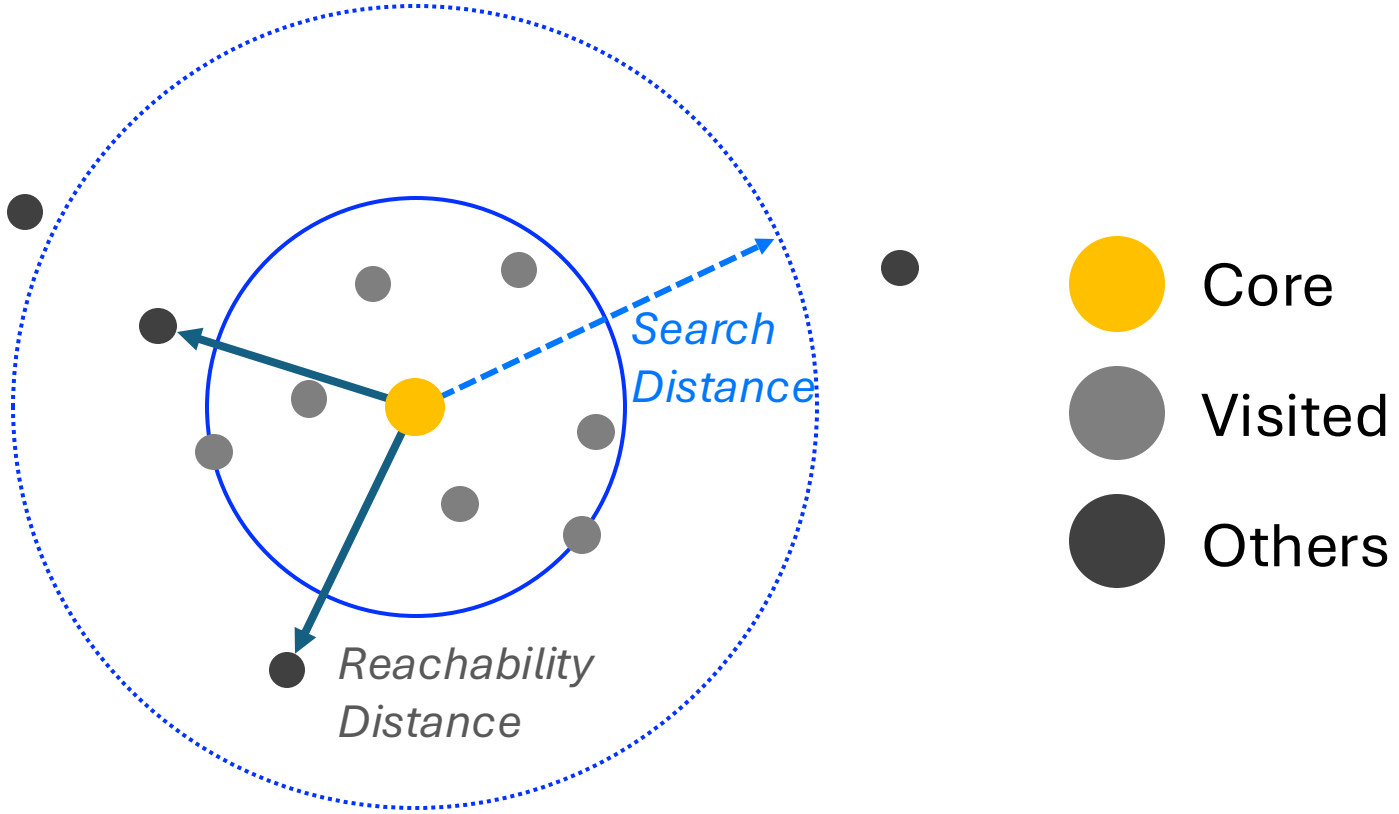
- For the **Multi-scale (OPTICS)** clustering method, the search distance value is treated as the maximum distance that will be compared to the core distance.
- OPTICS uses a concept of a minimum reachability distance, which is the distance from a point to its nearest neighbor that the search has not yet visited.
- OPTICS will search all neighbor distances within the specified search distance, comparing each of them to the core distance.

Situations	Core Distance Update	Core Distance
$\forall NeighborDist < CoreDist$	False	Original core distance
$\forall NeighborDist > CoreDist$	True	Replace by Min(Neighbor distance)

Source: https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/how-density-based-clustering-works.htm

Density-based Clustering :: OPTICS

- Revisit the situation again with a graph demonstration.



Source: https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/how-density-based-clustering-works.htm

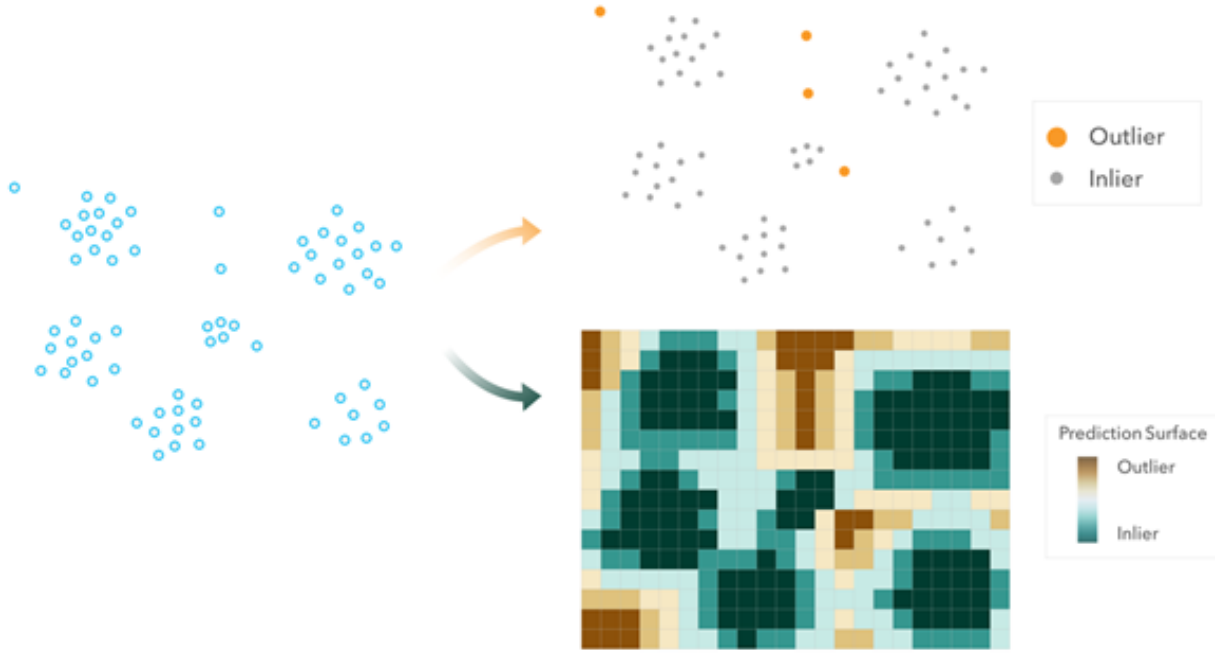
Density-based Clustering :: OPTICS

- For both Defined distance (DBSCAN) and Multi-scale (OPTICS), if no distance is specified, the default search distance is the highest core distance found in the dataset, excluding those core distances in the top 1 percent (in other words, excluding the most extreme core-distances).
- OPTICS is an ordered algorithm that starts with the feature with the smallest ID and goes from that point to the next to create a plot. The order of the points is fundamental to the results.

Source: https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/how-density-based-clustering-works.htm

Spatial Outlier Detection

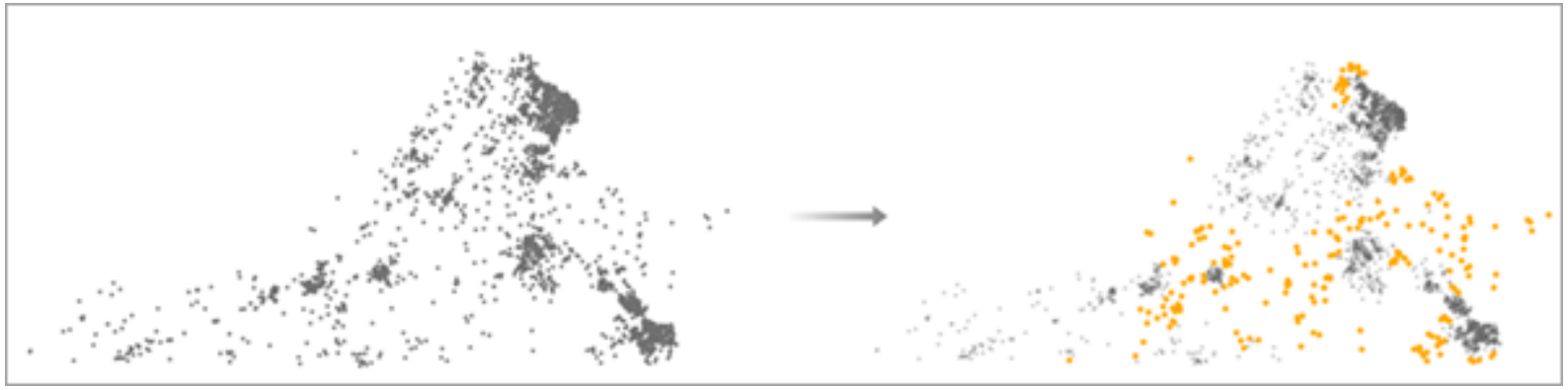
- Identifies global or local spatial outliers in point features.
- A **global outlier** is a point that is far away from all other points in a feature class.
- Global outliers are detected by examining distances between each point and one of its closest neighbors (by default, the closest neighbor) and detecting points where the distance is large.

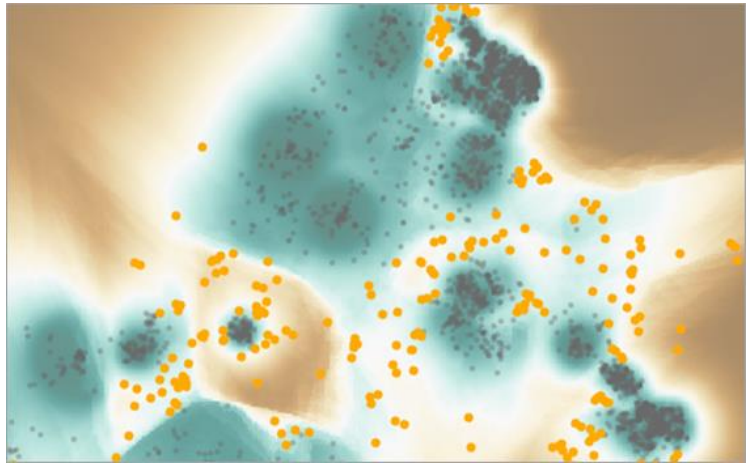


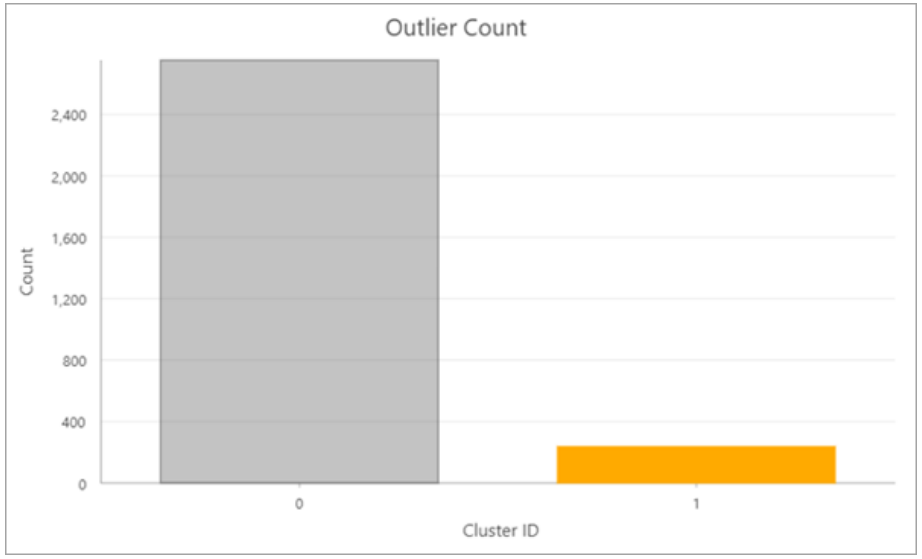
Spatial Outlier Detection

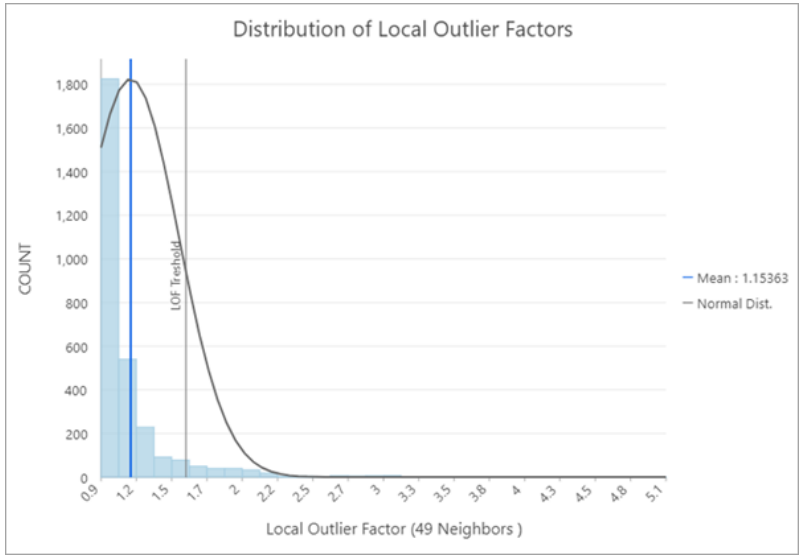
- A **local outlier** is a point that is farther away from its neighbors than would be expected by the density of points in the surrounding area. Local outliers are detected by calculating the local outlier factor (LOF) of each feature.
- The LOF is a measure that describes how isolated a location is compared to its local neighbors. A higher LOF value indicates greater isolation.
- The tool can also be used to produce a raster prediction surface that can be used to estimate whether new features will be classified as outliers based on the spatial distribution of the data.

Spatial Outlier Detection





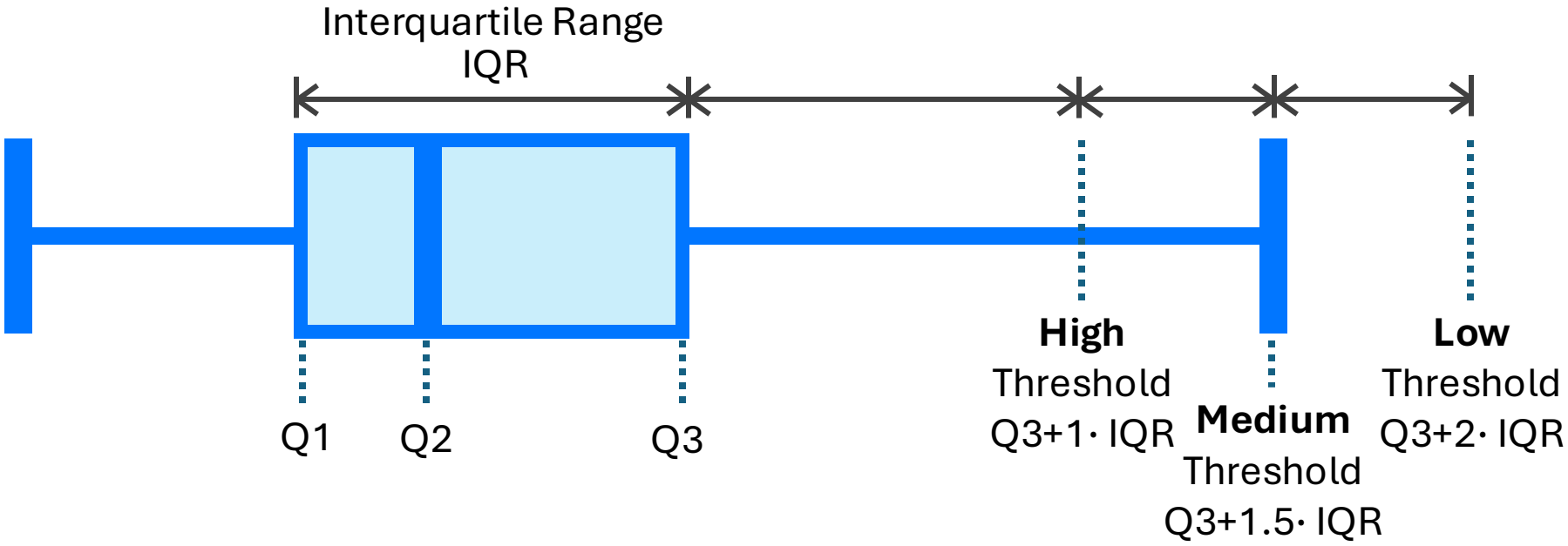




Source: https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/how-spatial-outlier-detection-works.htm

Spatial Outlier Detection

- **Global outliers** are simpler than local outliers. Global outliers are determined by calculating the distance to one of its closest neighbors, called the neighbor distance. By default, the closest neighbor is used, but you can change the number using the Number of Neighbors parameter.



Source: https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/how-spatial-outlier-detection-works.htm

Spatial Outlier Detection

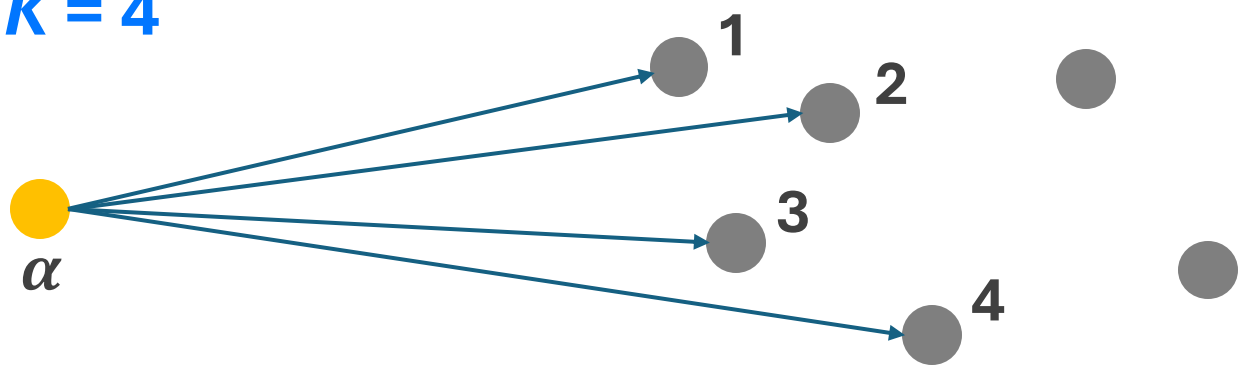
- For the tool to measure and identify **spatial outliers**, it requires a value for the **Number of Neighbors parameter** evaluated for each feature and a value for the **Percent of Locations Considered Outliers parameter** in the study area; these criteria are important when determining the size of the neighborhood in the LOF calculation and the threshold for designating outliers and inliers.

Source: https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/how-spatial-outlier-detection-works.htm

Spatial Outlier Detection

- A local neighborhood is established for each location using a specified minimum number of features. This approach is commonly referred to as *K*-nearest neighbors, where *K* corresponds to the specified minimum number of features in the vicinity of the currently analyzed feature.

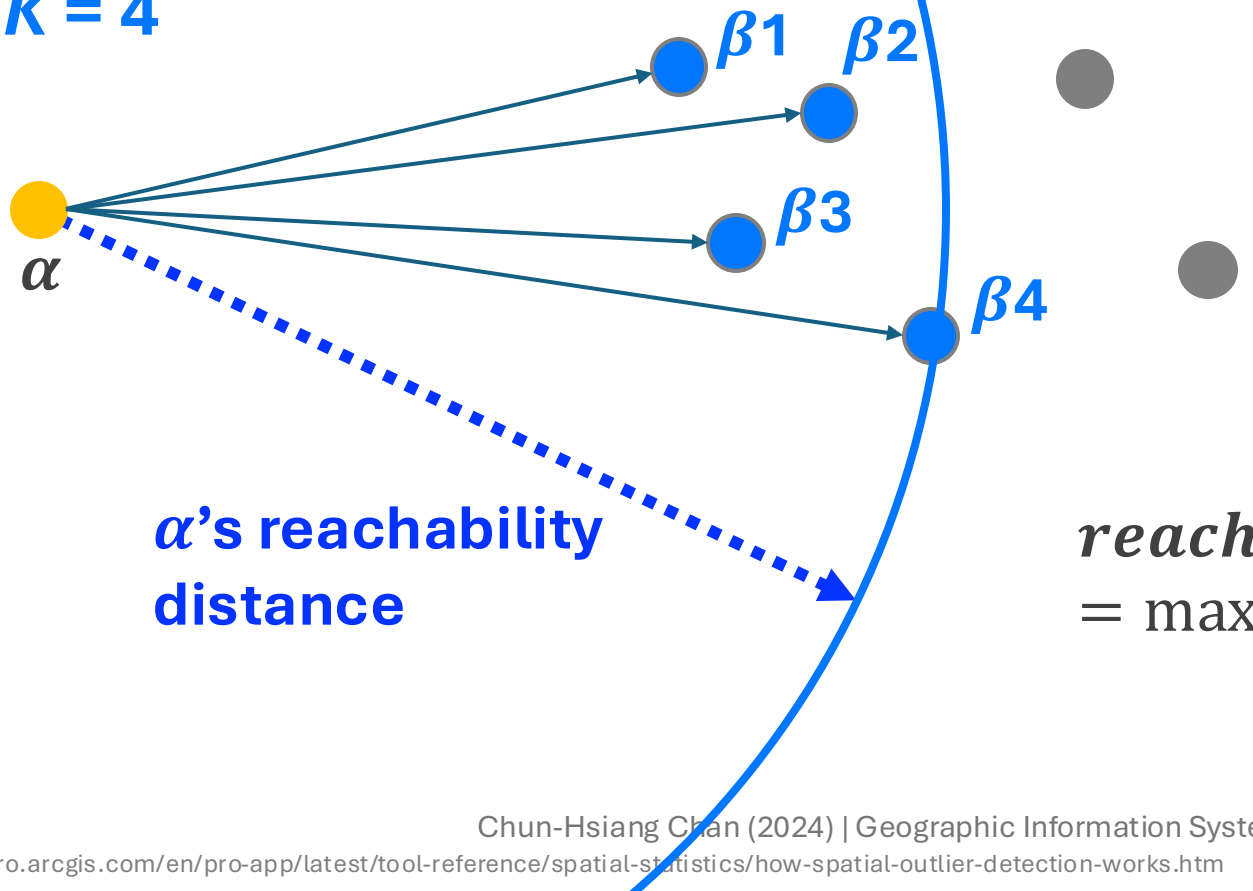
When ***K* = 4**



Source: https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/how-spatial-outlier-detection-works.htm

Spatial Outlier Detection

When $K = 4$

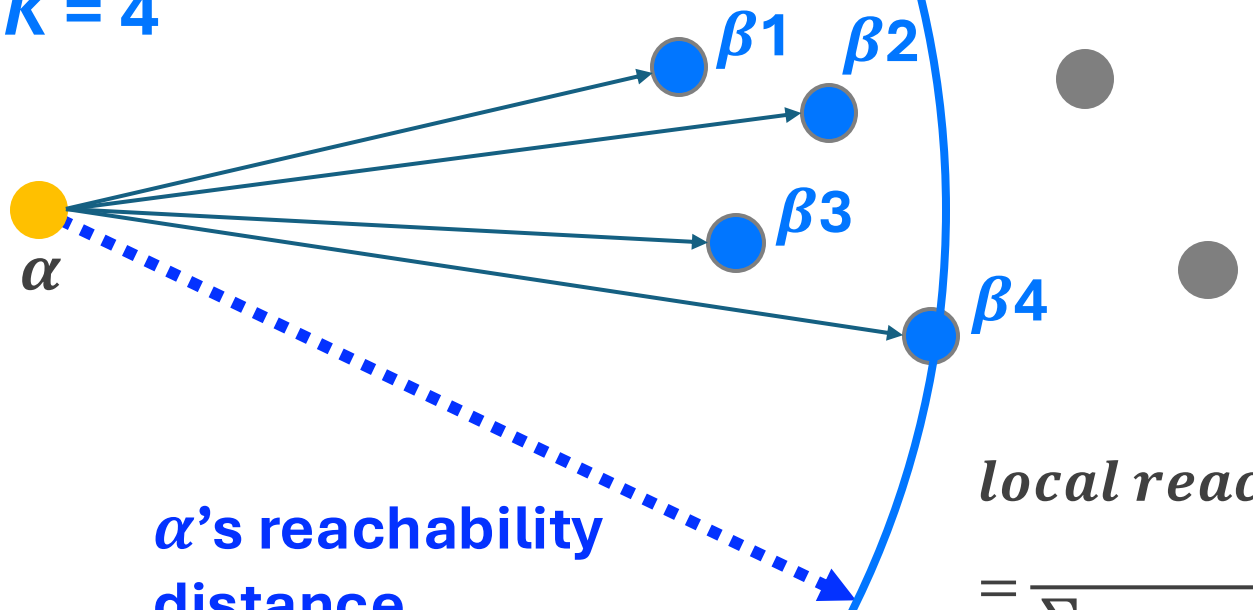


$$\boldsymbol{reachability - distance_k(\alpha, \beta)} = \max\{k - distance(\beta), d(\boldsymbol{\alpha}, \boldsymbol{\beta})\}$$

Source: https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/how-spatial-outlier-detection-works.htm

Spatial Outlier Detection

When $K = 4$

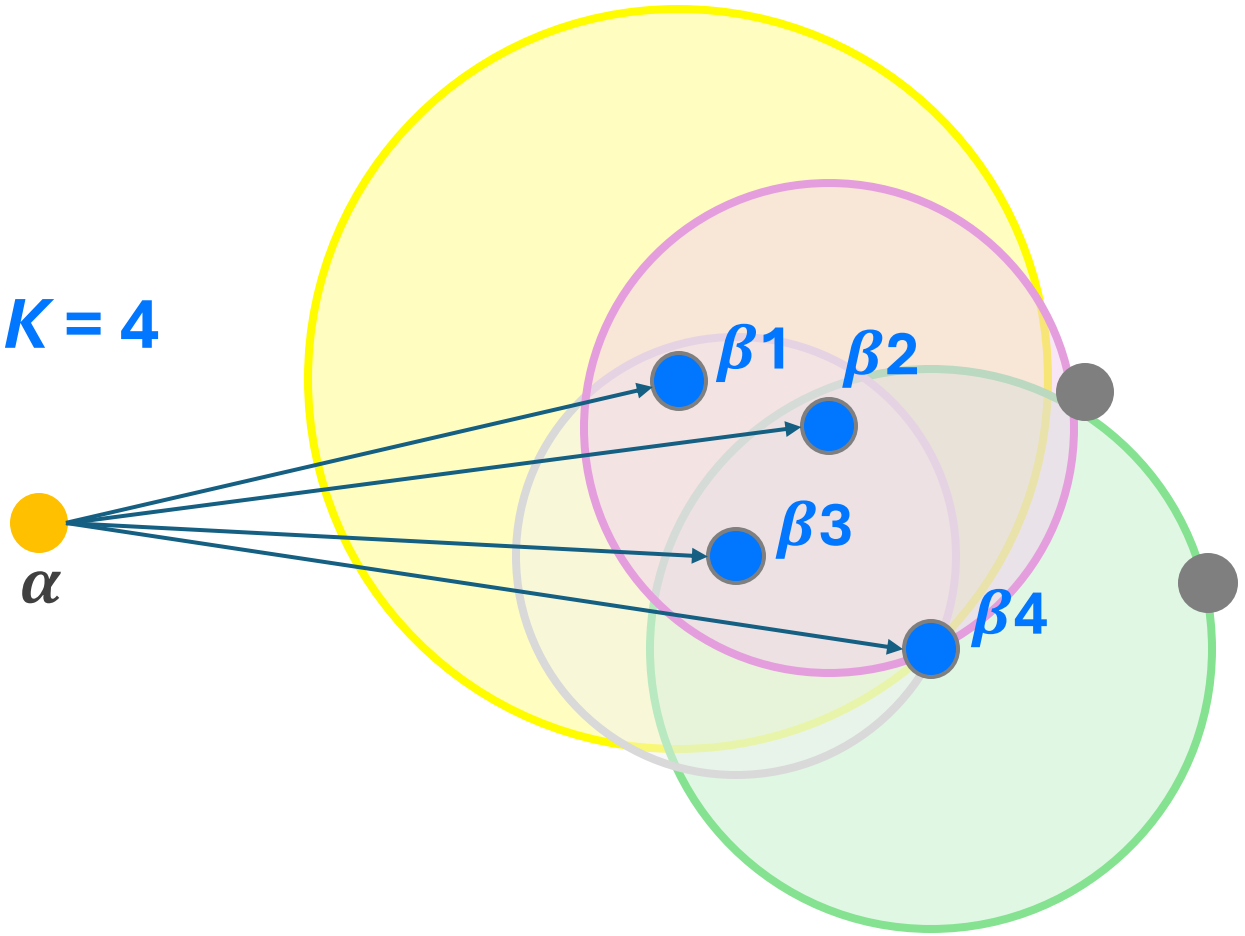


$$\boldsymbol{local\ reachability\ density}_k(\alpha) = \frac{1}{\frac{\sum_{\beta \in N_{a(\alpha)}} readchability - distance_k(\boldsymbol{\alpha}, \boldsymbol{\beta})}{|N_a(\alpha)|}}$$

Source: https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/how-spatial-outlier-detection-works.htm

Spatial Outlier Detection

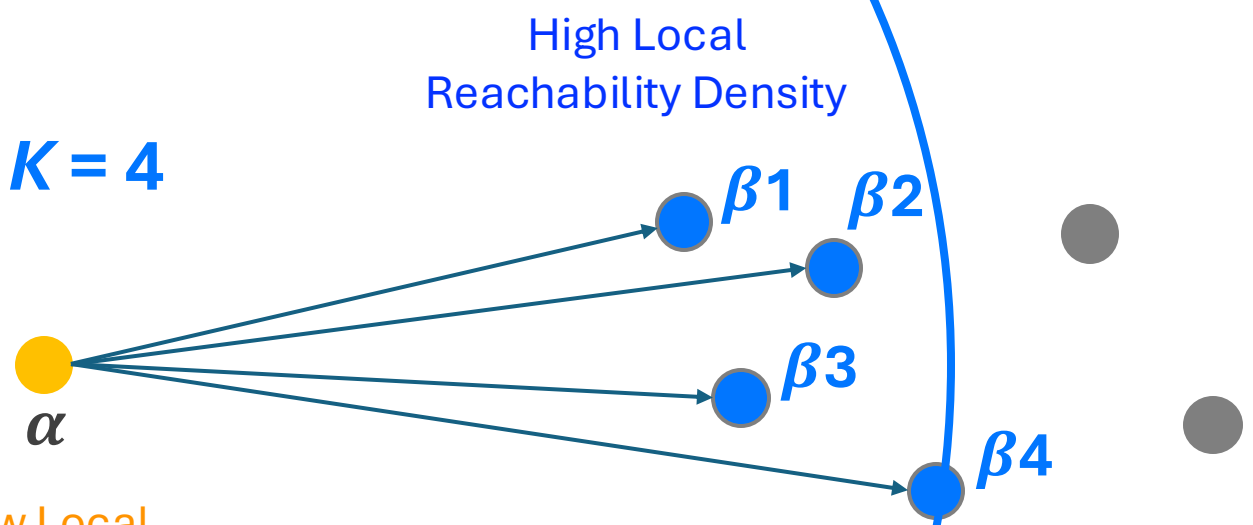
When $K = 4$



Source: https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/how-spatial-outlier-detection-works.htm

Spatial Outlier Detection

When $K = 4$



$$\boldsymbol{local\ outlier\ factor_k(\alpha)} = \frac{\frac{\sum_{\beta \in N_a(\alpha)} local\ density_k(\beta)}{local\ \ density_k(\alpha)}}{|N_a(\alpha)|}$$

Source: https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/how-spatial-outlier-detection-works.htm

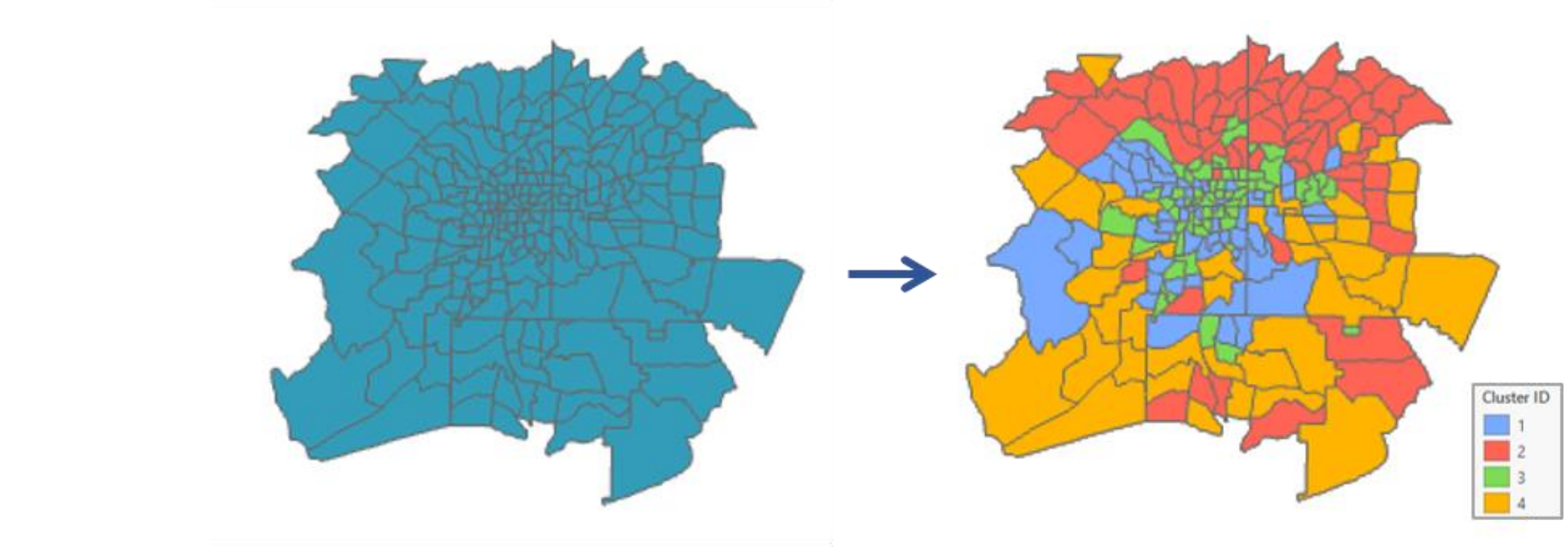
False Discovery Rate Correction

- The local spatial pattern analysis tools including Hot Spot Analysis and Cluster and Outlier Analysis Anselin Local Moran's I provide an optional Boolean parameter Apply False Discovery Rate (FDR) Correction.
- **Multiple Testing**—With a confidence level of 95 percent, probability theory tells us that there are 5 out of 100 chances that a spatial pattern could appear structured (clustered or dispersed) and could be associated with a statistically significant *p*-value, when in fact the underlying spatial processes promoting the pattern are truly random.
- **Spatial Dependency**—Features near to each other tend to be similar; more often than not spatial data exhibits this type of dependency. Nonetheless, many statistical tests require features to be independent. For local pattern analysis tools this is because spatial dependency can artificially inflate statistical significance.

Source: https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/how-multivariate-clustering-works.htm

Multivariate Clustering

- Finds natural clusters of features based solely on feature attribute values.



Source: https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/how-multivariate-clustering-works.htm

Multivariate Clustering

- ArcGIS Pro provides k-means clustering for grouping all features with similar attributes (Analysis Fields ← manual settings).
- **Input**
  1) Input Features
  2) Analysis Fields
  3) Number of Clusters
- **Output**
  1) Cluster Map
  2) Multivariate Clustering Box-Plots
  3) Features Per Cluster

Source: https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/how-multivariate-clustering-works.htm

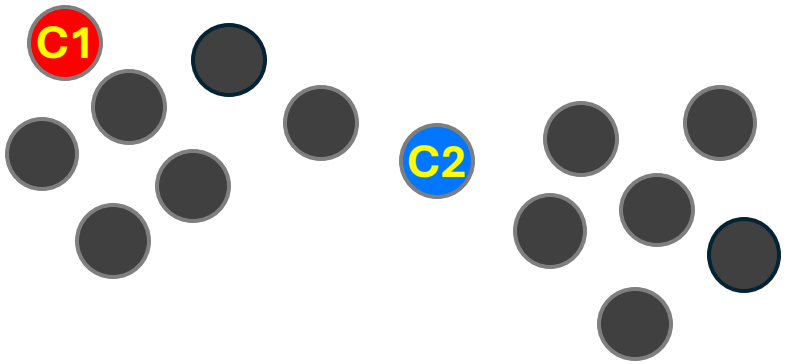
Multivariate Clustering

1 Define the number of clusters (*k*)

Given *k* value of 2

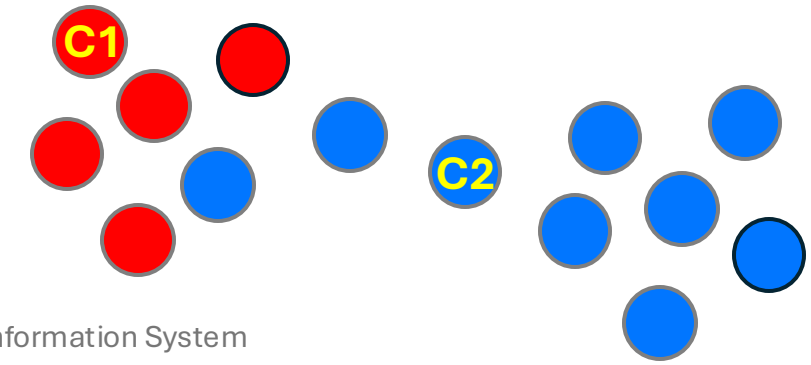
2 Select *k* random points from the dataset as centroids

The number of clusters is 2; therefore, we randomly select 2 data points as centroids



3 Assign all data points to the nearest centroid

For each data point has been assigned to a centroid by the closet distance

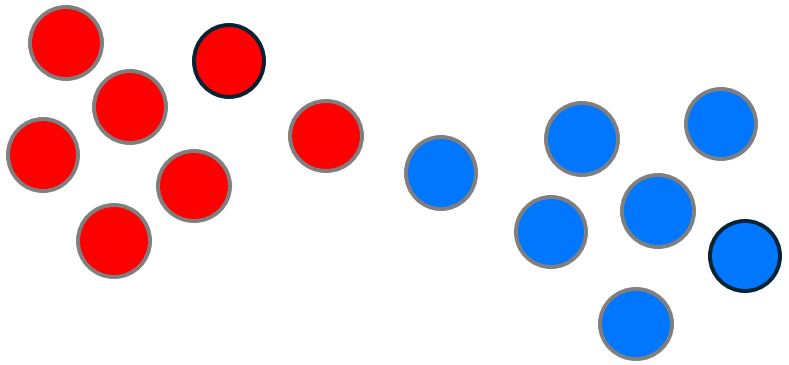


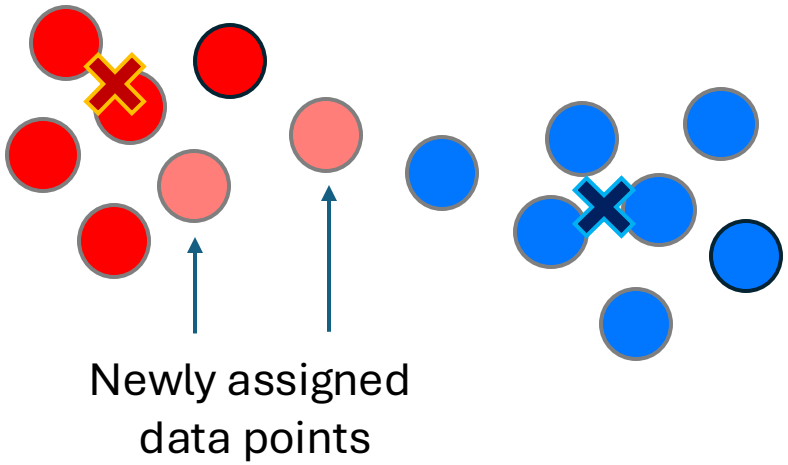
Source: https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/

Multivariate Clustering

4 Update centroid by mean centers

Find the mean center of each group, defined by the previous step.





5 Repeat step 3 and step 4

- Centroids of newly formed clusters do not change
- Points remain in the same cluster
- Maximum number of iterations is reached

Source: https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/

Multivariate Clustering

- Since *k* is a self-defined variable, how can the optimized *k* value be defined?
- In ArcGIS Pro, Grouping effectiveness is measured using the **Calinski-Harabasz pseudo F-statistic**, which is a ratio reflecting within-group similarity and between-group difference:

Between-cluster differences

$$F = \frac{\frac{R^2}{n_c - 1}}{\frac{1 - R^2}{n - n_c}}, where\ R^2 = \frac{SST - SSE}{SST},$$

$$\begin{cases} SST = \sum_{i=1}^{n_c} \sum_{j=1}^{n_i} \sum_{k=1}^{n_v} \left(V_{ij}^k - \overline{V^k}\right)^2 \\ SSE = \sum_{i=1}^{n_c} \sum_{j=1}^{n_i} \sum_{k=1}^{n_v} \left(V_{ij}^k - \overline{V_i^k}\right)^2 \end{cases}$$

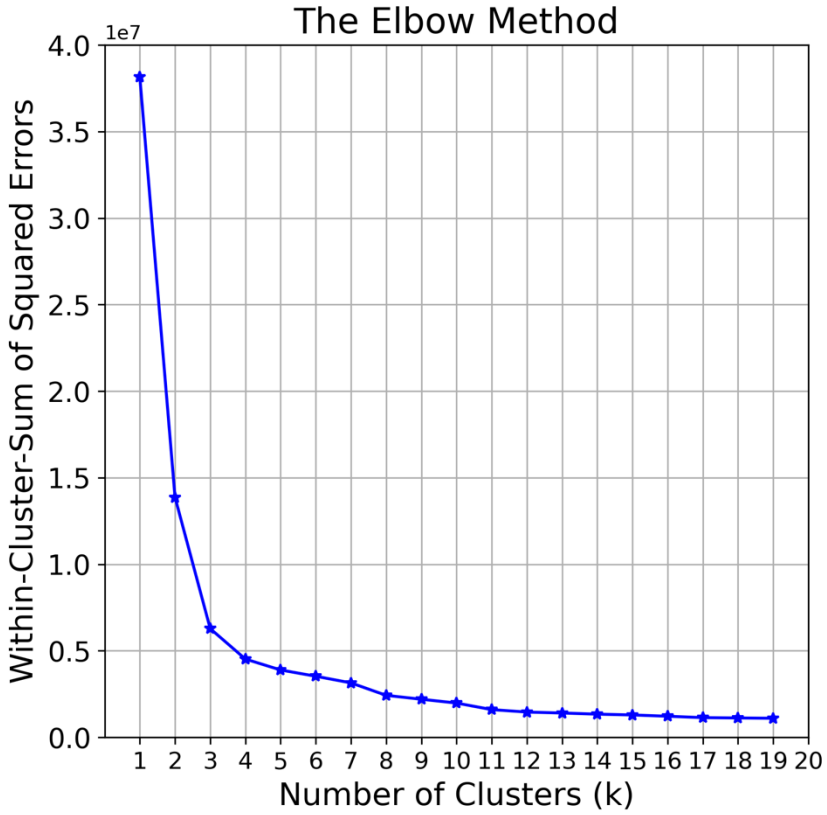
Within-cluster differences

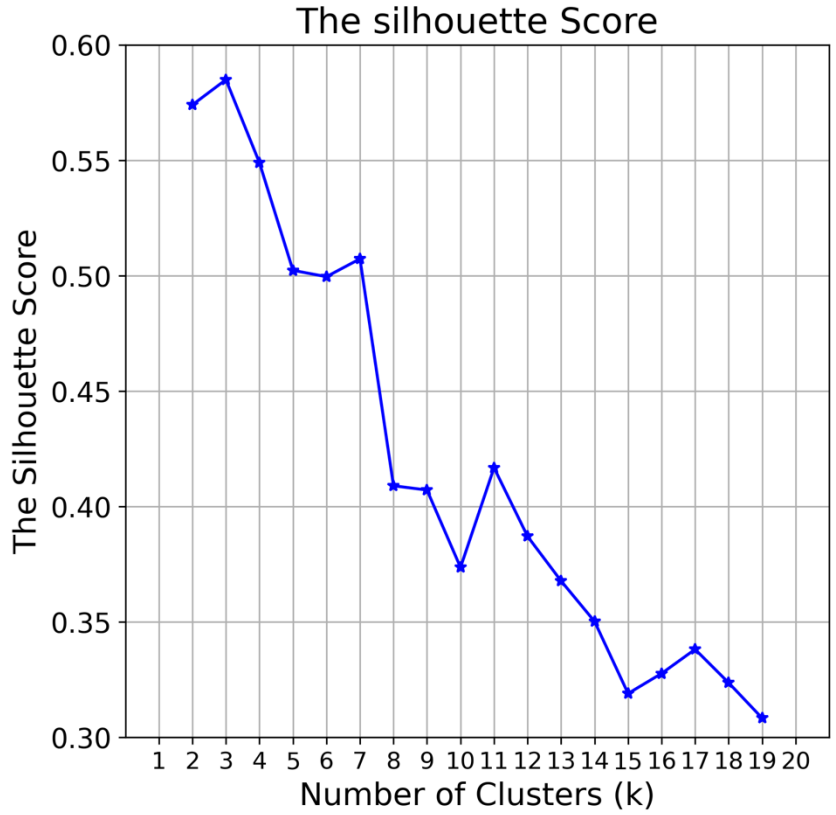
n: the number of features; n_i: the number of features in cluster i; n_c: the number of classes (clusters); n_v: the number of variables used to cluster features; V_{ij}^k: the value of the k^{th} variable of the j^{th} feature in the i^{th} cluster; $\overline{V^k}$: the mean value of the k^{th} variable; $\overline{V_i^k}$: the mean value of the k^{th} variable in cluster i.

Source: https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/how-grouping-analysis-works.htm#ESRI_SECTION1_0DDB93A2A36148A29E3CA2D19B3C6C64

Multivariate Clustering

- In general, we usually adopt "**The Elbow Method**" or "**The Silhouette Method**" to determine the best number of clusters.



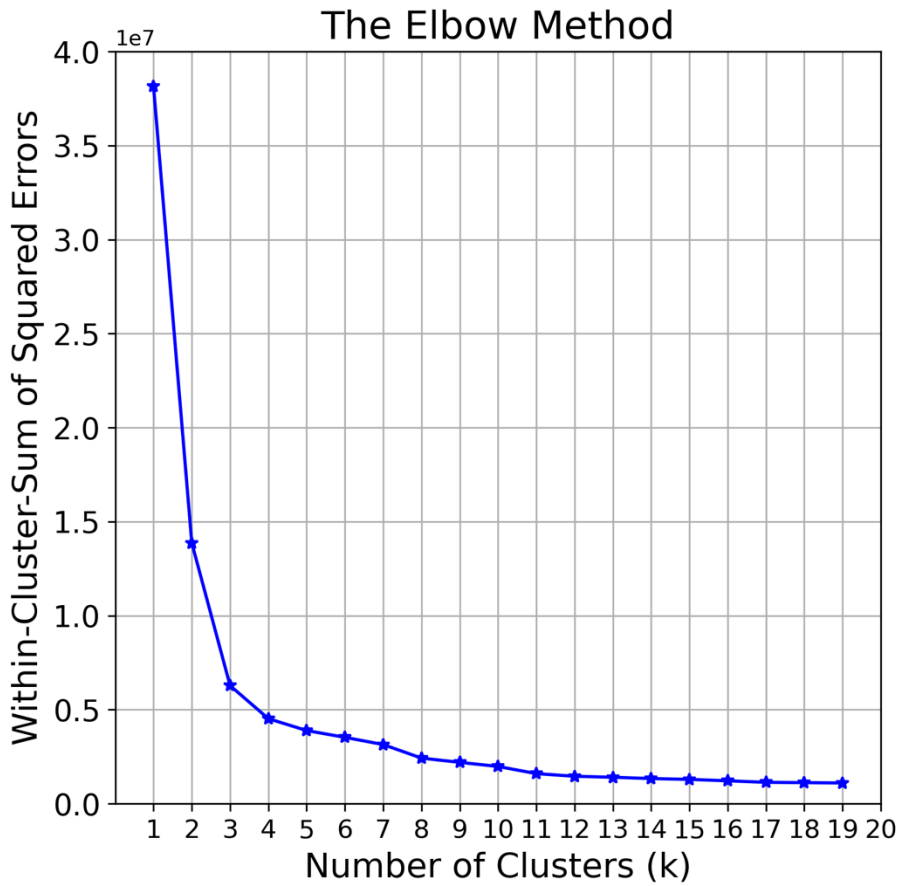


Multivariate Clustering :: The Elbow Method

- WCSS, also known as inertia, measures the sum of squared distances between each data point and the centroid of its assigned cluster.
- The centroid is the mean position of all the points in the cluster.

$$WCSS = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2$$

k is the number of cluster; C_i represents the ith cluster; x denotes the data points in cluster C_i; μ_i is the centroid of cluster C_i; $\|x - \mu_i\|^2$ is the squared Euclidean distance between point x and centroid μ_i.



Source: https://medium.com/@sreeku.ralla/wcss-how-many-clusters-are-good-enough-74f91c06dc75

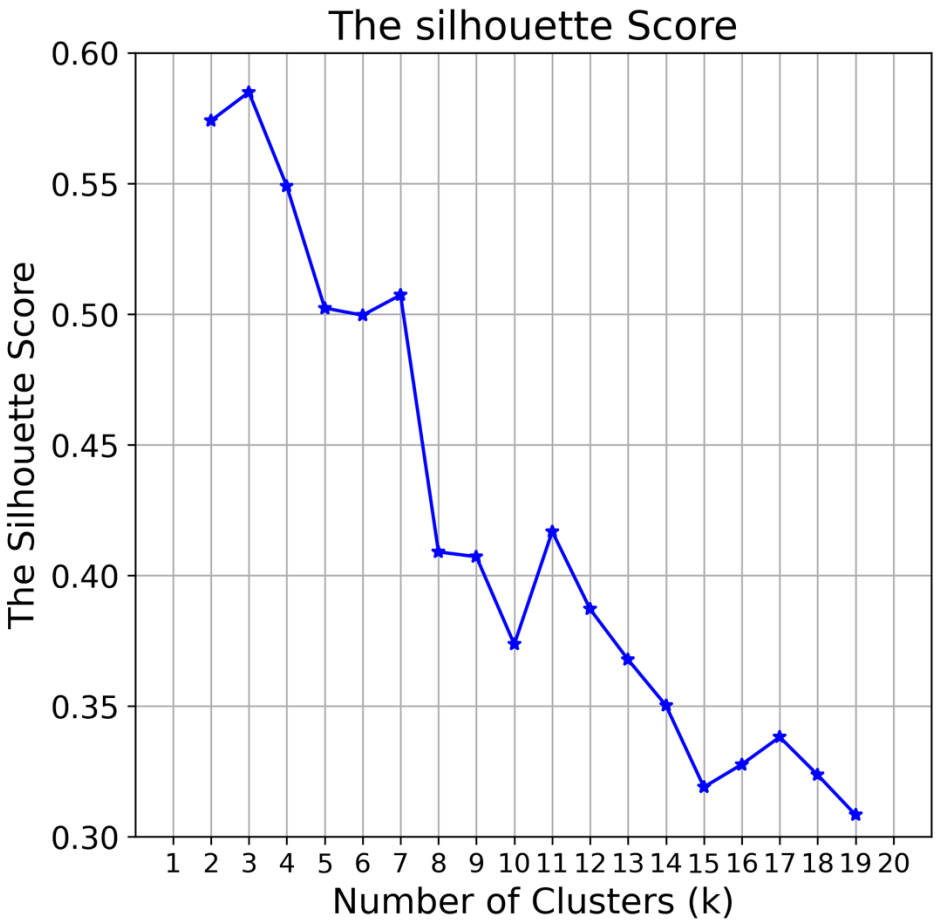
Multivariate Clustering :: The Elbow Method

- Calculate the **Within-Cluster-Sum of Squared Errors (WCSS)** for different values of *k*, and choose the k for which WSS first starts to diminish. In the plot of WCSS-versus-*k*, this is visible as an elbow.
- The Squared Error for each point is the square of the distance of the point from its representation, i.e., its predicted cluster center.
- The WCSS score is the sum of these Squared Errors for all the points.
- Any distance metric like Euclidean or Manhattan Distance can be used.

Source: https://medium.com/@sreeku.ralla/wcss-how-many-clusters-are-good-enough-74f91c06dc75

Multivariate Clustering :: The Silhouette Value

- The silhouette value measures how similar a point is to its own cluster (cohesion) compared to other clusters (separation).
- The range of the Silhouette value is between +1 and −1.
- A high value is desirable and indicates that the point is placed in the correct cluster.
- If many points have a negative Silhouette value, it may indicate that we have created too many or too few clusters.



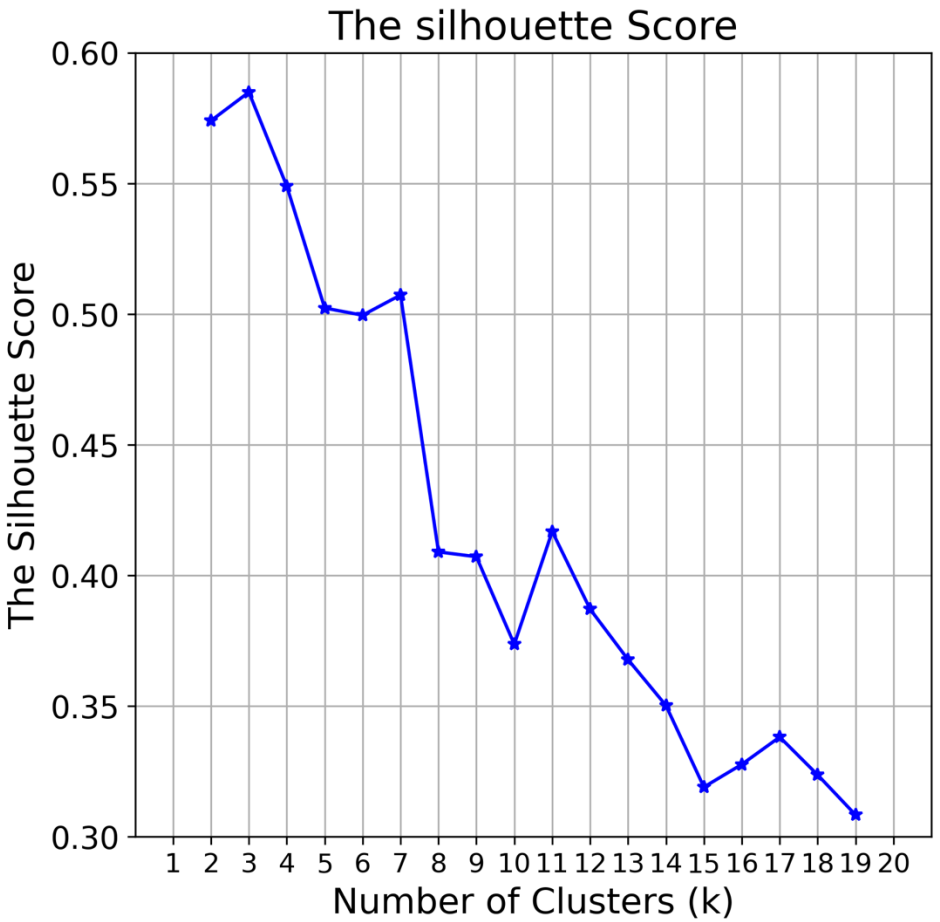
Source: https://medium.com/analytics-vidhya/how-to-determine-the-optimal-k-for-k-means-708505d204eb

Multivariate Clustering :: The Silhouette Value

- The Silhouette value $s(i)$ for each data point i is defined as follows:

$$\begin{cases} s(i) = \dfrac{b(i) - a(i)}{max\{a(i), b(i)\}}, if\ |C_i| > 1 \\ s(i) = 0, if |C_i| = 1 \end{cases}$$

- $s(i)$ is defined to be zero if i is the only point in the cluster. This is to prevent the number of clusters from increasing significantly with many single-point clusters.



Source: https://medium.com/analytics-vidhya/how-to-determine-the-optimal-k-for-k-means-708505d204eb

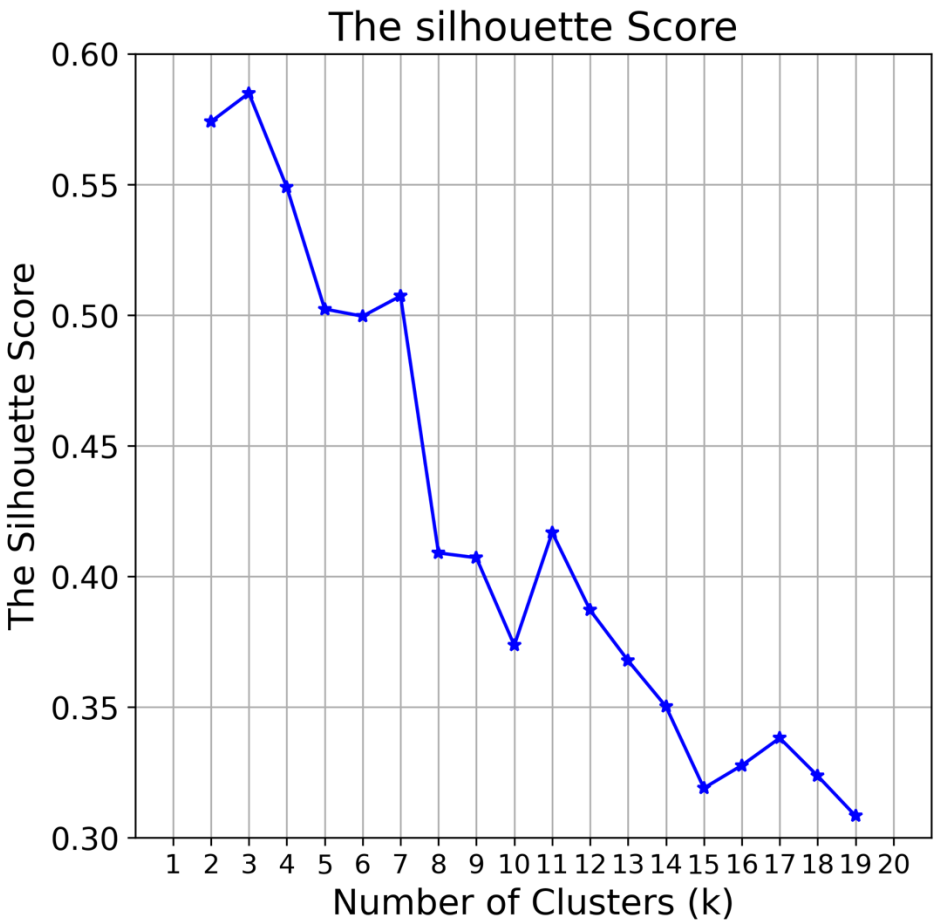
Multivariate Clustering :: The Silhouette Value

- Here, $a(i)$ is the measure of similarity of the point i to its own cluster. It is measured as the average distance of i from other points in the cluster.

$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} d(i,j)$$

- Similarly, $b(i)$ is the measure of dissimilarity of i from points in their clusters.

$$b(i) = \min_{i \neq j} \frac{1}{|C_j|} \sum_{j \neq C_j} d(i,j)$$

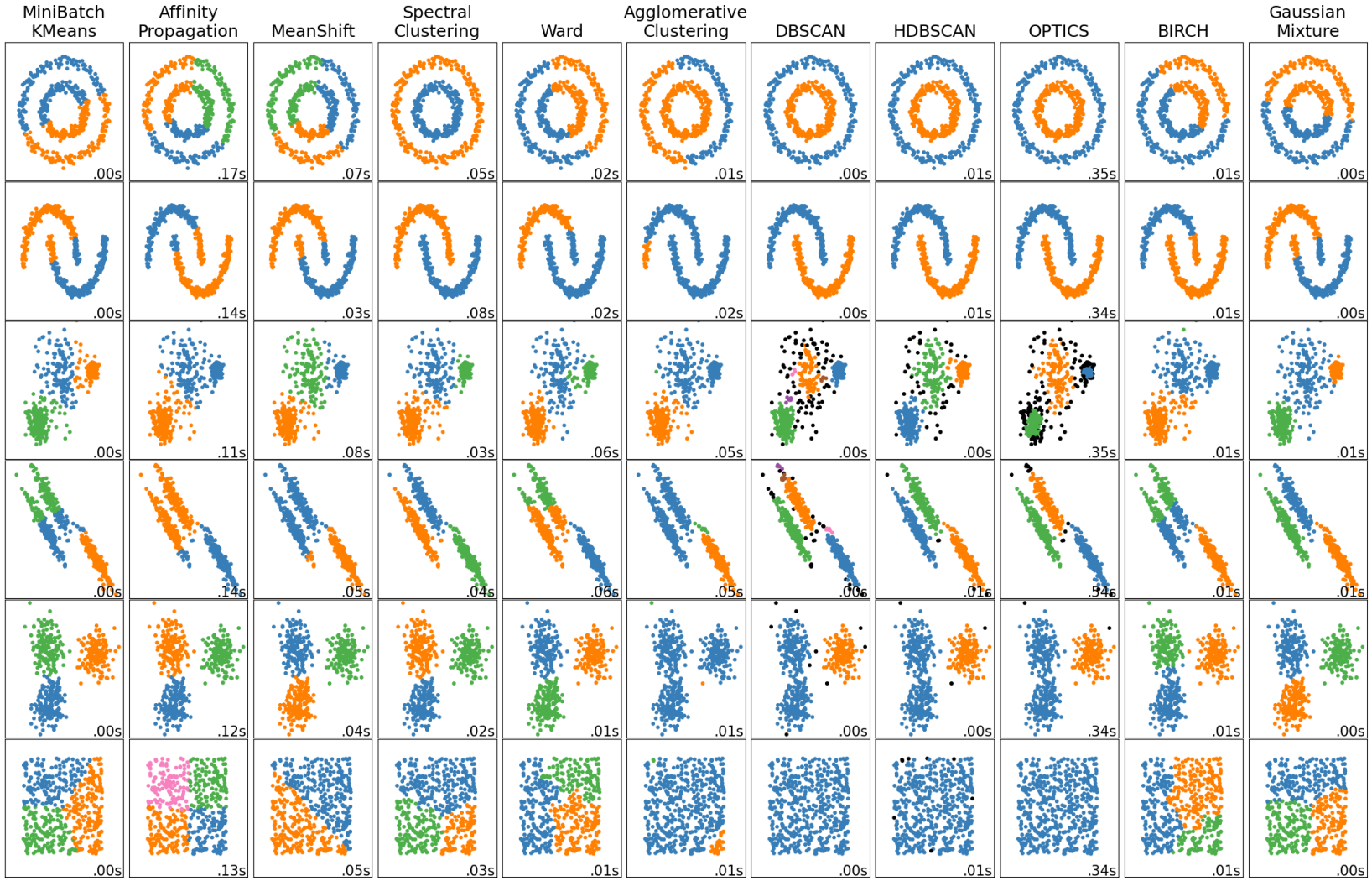


Source: https://medium.com/analytics-vidhya/how-to-determine-the-optimal-k-for-k-means-708505d204eb

Machine Learning :: Clustering

- Clustering analysis groups samples or data units based on the similarity of their features or attributes.
- Hence, clustering analyses are widely conducted to characterize or rapidly understand the data patterns and distribution.
- Clustering analysis is also an unsupervised learning method that provides insightful information for feature extraction, data engineering, and dimension reduction.

Machine Learning :: Clustering



Source: https://scikit-learn.org/1.5/auto_examples/cluster/plot_cluster_comparison.html

Lab#01 Physical Meanings

- How to define HH, LL, HL, and LH in statistics and explain each cluster?
- How do we define hot and cold spots in statistics and give corresponding explanations for them?
- What are the differences between DBSCAN, OPTICS, and *k*-means clustering from an algorithm and output (result) perspective?
- How do you solve the cluster sensitivity problem in the DBSCAN and OPTICS?
- What are the definitions of core, border, and noise points in DBSCAN?
- When you conduct OPTICS for density-based clustering, what kinds of issues that you have to aware?

Lab#01 Physical Meanings

- Since the number of clusters in k-means is self-defined, how do you determine the best k?
- What's WCSS? What's the foundation of WCSS in determining the number of clusters for k-means clustering?
- What's Silhouette value? What's the foundation of the Silhouette value in determining the number of clusters for k-means clustering?

The End

Thank you for your attention!

Email: chchan@ntnu.edu.tw
Web: toodou.github.io