



Aalto University  
School of Engineering

# Spatial Clustering

*Advanced Spatial Analytics*  
*17 January 2019*  
*Marko Kallio*

# Contents

**Multivariate data analysis – recap**

**Clustering**

**Clusterins algorithms**

**Regionalization (spatialization) of clustering algorithms**

# Multivariate data analysis

Any analysis that simultaneously handles of more than two variables measured on the same individual or object (vs. univariate or bivariate analysis)

## **Multivariable vs multivariate techniques:**

- Multivariable: One response variable
- Truly multivariate: "All variables are random and interrelated in a way that their different effects cannot be meaningfully interpreted separately", **multiple response** variables (Hair et al, 2006, Multivariate Data Analysis, p. 4)

# Multivariate methods

## Dependence techniques

- When one variable (or a set of variables) can be defined as **dependent**, to be predicted by **independent** variables
- Structural equation modelling (SEM), canonical correlation analysis, multivariate ANOVA, *multiple regression*, conjoint analysis, multiple discriminant analysis, linear probability models

## Interdependence techniques

- When **no** variables can be defined as **dependent/independent**
- PCA, factor analysis, **clustering**, multidimensional scaling, correspondende analysis

Hair et al, 2006

# Some important concepts:

## Variate

**(Linear) variate is a linear combination of variables with empirically determined weights.**

$$\textit{Variate value} = w_1X_1 + w_2X_2 + \cdots + w_nX_n$$

**Variables are specified by the user.**

**Weights are determined by the multivariate technique.**

# Some important concepts:

## Measurement scale

**Nonmetric – discrete characteristics or properties having a single value, with all other possible values excluded. E.g. Gender (male/female).**

- Nominal
- Ordinal

**Metric – amount or degree of an attribute. Constant unit of measurement.**

- Interval – Arbitrary zero point (e.g. Celsius, Fahrenheit)
- Ratio – Absolute zero point. The most accurate of all scales.

# Some important concepts: Similarity / dissimilarity measures

## Distance measures

- Based on magnitudes
- Measures multidimensional proximity between observations
- Euclidean, statistical, manhattan, mahalanobis distance etc...

## Correlational measures

- Based on patterns
- Correlation of variables (normal data matrix) , OR
- Correlation of observations (transposed data matrix)

# Spatial Clustering



Aalto University  
School of Engineering



# Clustering

**Clustering is the grouping of observations to groups in which the observations are as *similar* as possible, and which are as *dissimilar* as possible with observations in other groups.**

**Clustering is an unsupervised method.** It results in classes (groups), but which are not known in advance (in classification, the classes are known a priori).

- In classification training and testing data are needed; in clustering they are irrelevant

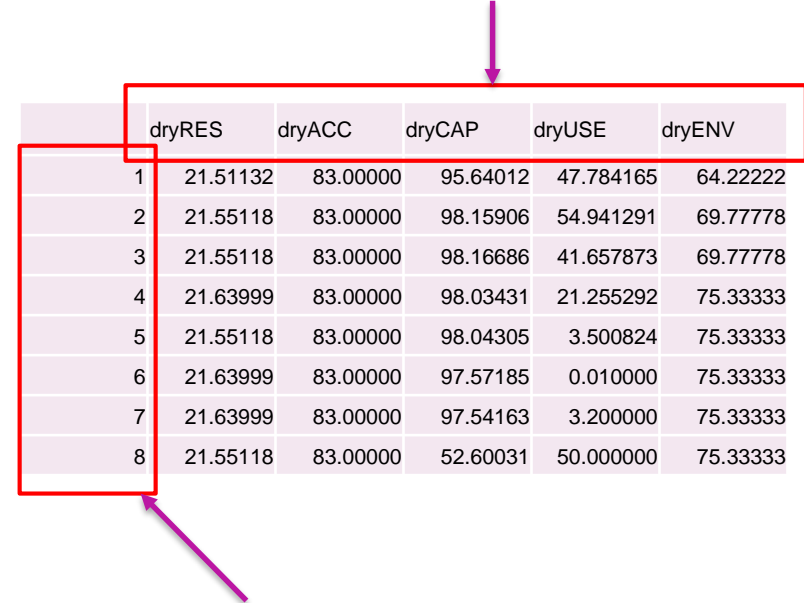
# Clustering vs PrinCompAnalysis

Both methods attempt to assess the structure of the dataset.

In clustering we group *observations* (objects) based on a dissimilarity measure.

PCA groups *variables* using patterns of variation (or correlations).

PCA groups variables



The diagram illustrates the difference between PCA and Clustering. A purple arrow points from the text 'PCA groups variables' to a red box highlighting the variable headers (dryRES, dryACC, dryCAP, dryUSE, dryENV) in the table. Another purple arrow points from the text 'Clustering groups observations' to a red box highlighting the observation indices (1 through 8) in the same table.

	dryRES	dryACC	dryCAP	dryUSE	dryENV
1	21.51132	83.00000	95.64012	47.784165	64.22222
2	21.55118	83.00000	98.15906	54.941291	69.77778
3	21.55118	83.00000	98.16686	41.657873	69.77778
4	21.63999	83.00000	98.03431	21.255292	75.33333
5	21.55118	83.00000	98.04305	3.500824	75.33333
6	21.63999	83.00000	97.57185	0.010000	75.33333
7	21.63999	83.00000	97.54163	3.200000	75.33333
8	21.55118	83.00000	52.60031	50.000000	75.33333

Clustering groups observations

# Variate in Clustering and PCA:

**The variate (principal component) in PCA is straightforward:**

A PC is a linear combination of the original data values where weight is the coefficient load. Variate is determined empirically by PCA.

$$\text{Variate value}(PC \text{ score}) = w_1X_1 + w_2X_2 + \cdots + w_nX_n$$

  
**W determined by PCA**      **X given by user**

# Variate in Clustering and PCA:

## Clustering differs from other multivariate methods:

The variate is defined by the user, rather than the method!

- "The cluster variate is the set of variables representing the characteristics used to compare objects" (Hair et al, 2006)
- In clustering we compare objects based on the specified variate

# Dissimilarity

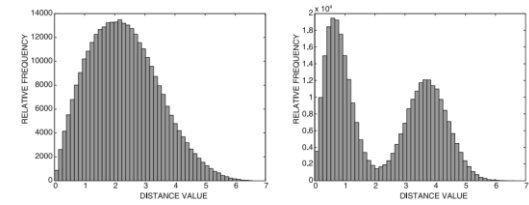
## Recall the dissimilarity measures

- Correlation measures
- Distance measures

## Either can be used, but notice the difference:

- Correlation based clustering results in clusters with similar *pattern/behaviour*
- Distance clustering results in clusters with similar magnitudes
- (Multivariate) clustering is commonly based on a distance measure

# Clustering workflow



1. **Preprocess data** (data types? Variables with different scales?)
2. **Variable selection**
  - a) Expert knowledge.
  - b) Filter models – based on some similarity measure.
  - c) Wrapper models – iterative clustering procedure with selection based on some internal validity criterion.
3. **Select and run clustering algorithm**
4. **Validate**
  - a) Internal criteria, e.g. intra-cluster to inter-cluster ratio of sum of squared error.
  - b) External criteria. Usually not available for real datasets.

# Selecting the number of clusters

## Based on prior knowledge

- Number of clusters in the data is known

## Based on expert knowledge

- The expert knows the data intimately and can determine, or guess, the number of clusters in the data

## Using computational indices

- Large number of indices can be used to guide the selection
- Warning: Do not rely on indices alone! Use your knowledge of the data!

# Clustering methods:

## Partitioning

- Breaks the data into  $k$  clusters, with  $k$  known a priori
- E.g. k-means

## Hierarchical

- Creates a nested tree of clusters – number of clusters can be determined in postprocessing

## Density



# Clustering methods:

## Hierarchical Clustering

**Creates a nested hierarchy of objects (a dendrogram).**

**Agglomerative clustering - all observations start in their own cluster**

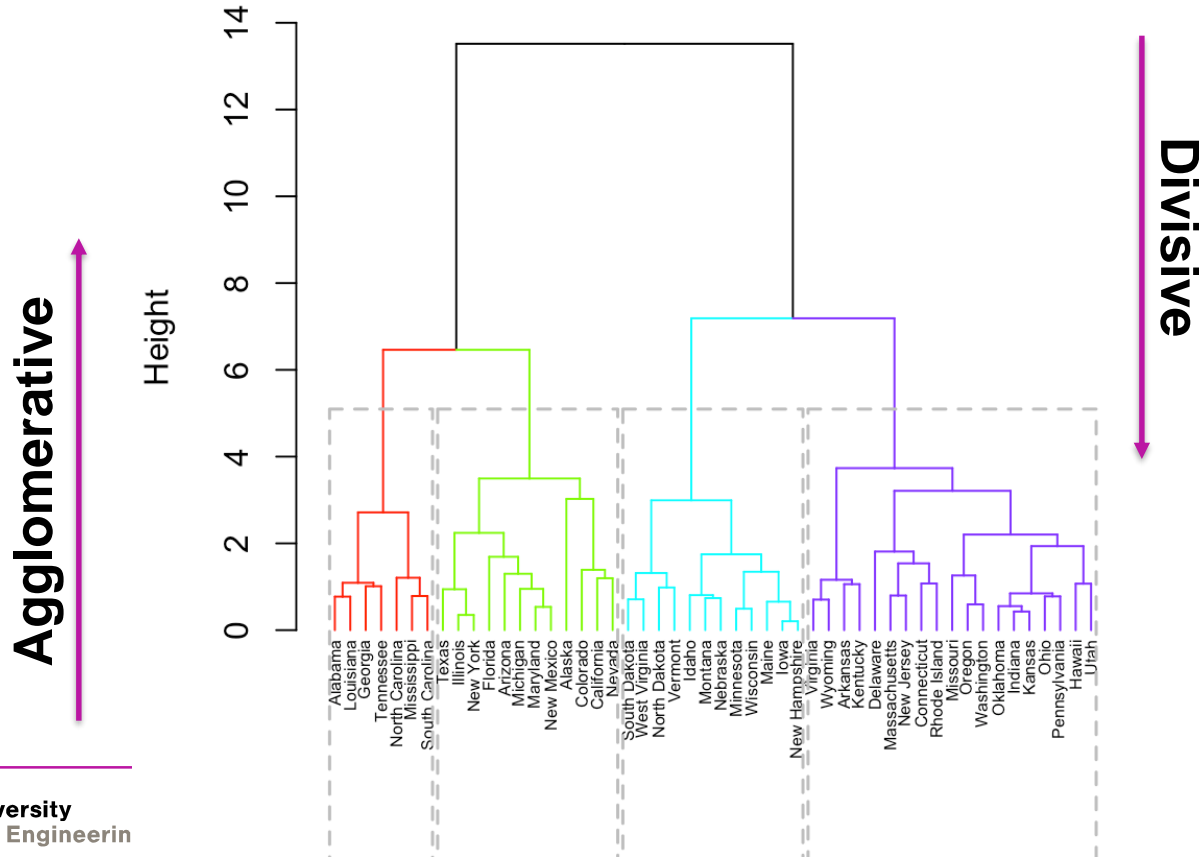
- Bottom-up algorithm
- Merges most similar clusters based on some distance measure

**Divisive clustering - all observations in one cluster**

- Top-down algorithm
- Splits clusters into smaller ones based on some distance measure

# Hierarchical clustering

Cluster Dendrogram



# Hierarchical clustering

## Several methods to determine cluster merging or splitting:

- **Single-linkage:** combining two clusters that contain the closest pair of elements not yet belonging to the same cluster as each other
- **Complete-linkage:** combines clusters that contain shortest distance between the farthest pair of elements
- **Average-linkage:** combines clusters with shortest average distance between all pairs of elements
- **Ward's method:** combine the clusters that lead to minimum increase in total within-cluster variance after merging.
  - *Many variations of Ward's method exist. The above is only one of many possibilities (the minimum variance criterion).*

# Hierarchical clustering

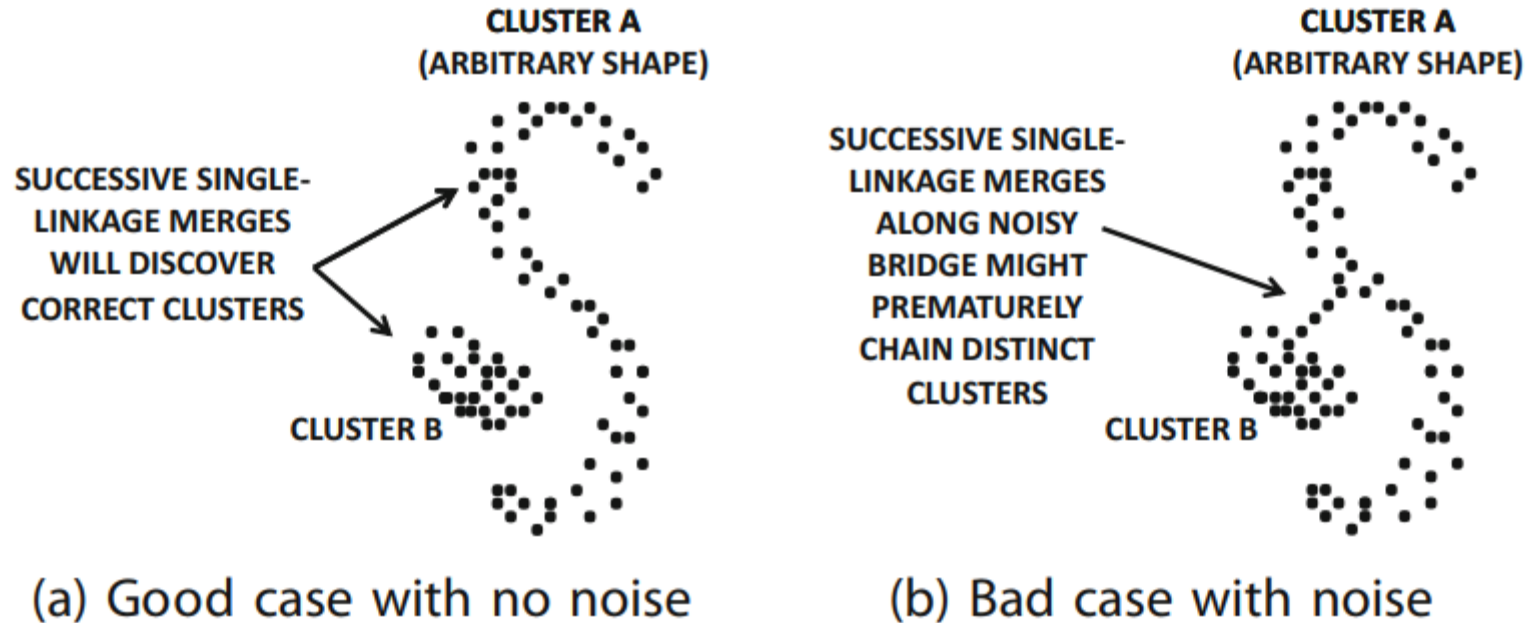


Figure 6.9: Good and bad cases for single-linkage clustering

# Clustering methods: *k*-representatives

Partitioning methods where the number  $k$  of clusters given a priori.

1. **Assign all points to some cluster**
2. **Compute a representative for each cluster**
3. **One by one, do for each point**
  1. Determine which cluster's representative is closest to the point
  2. Reassign point to that cluster
  3. Recalculate the representative for the affected clusters
  4. Repeat until no more reassignments are required

# Clustering methods: *k*-representatives

## ***k*-means**

Representative point is the mean of all members of the cluster.

*Euclidean k-means creates a multidimensional Voronoi Diagram.*

## ***k*-median**

Representative point is the median member of the cluster

## ***k*-medoid**

Representative is the medoid – the point with the shortest distance to all other cluster members.

# Clustering methods: *k*-representatives

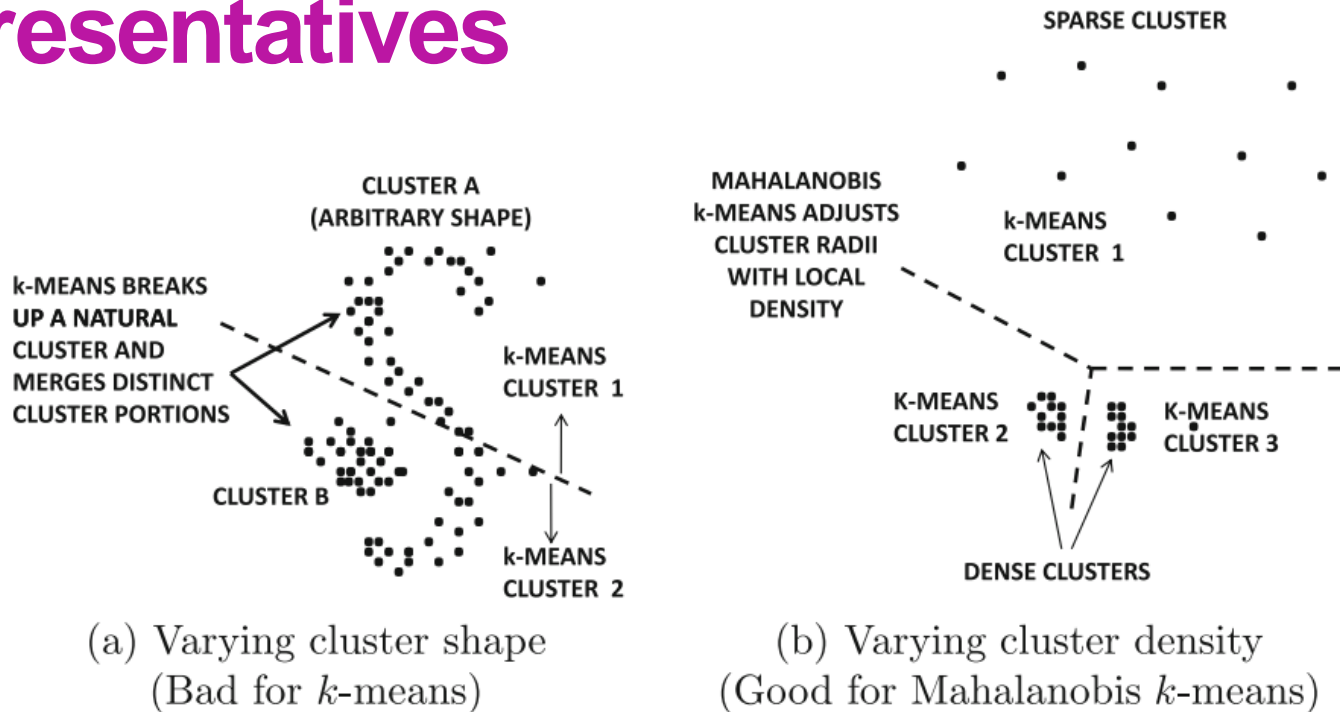


Figure 6.4: Strengths and weaknesses of *k*-means

# Clustering methods: Grid based methods

Similar to density based clustering methods.

1. Choose a resolution
2. Choose density threshold
3. Choose merging neighbourhood

Figure 6.12: Generic grid-based algorithm

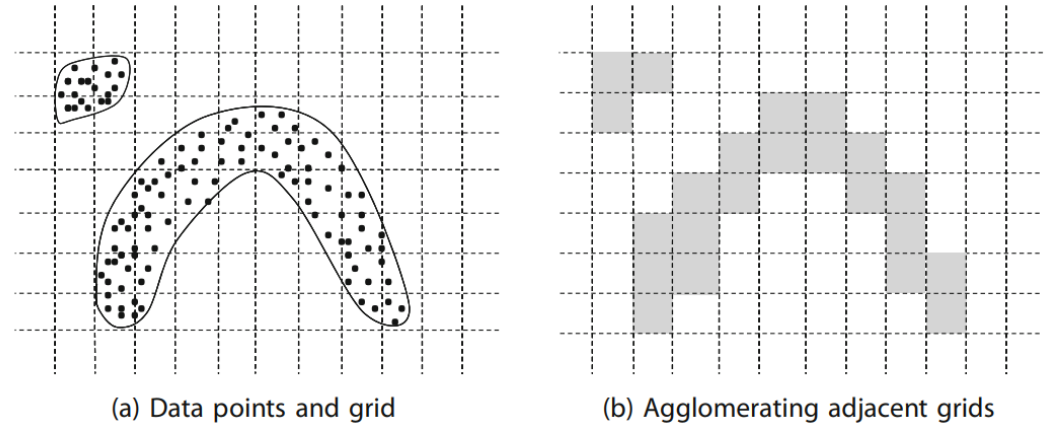


Figure 6.13: Agglomerating adjacent grids



# Clustering methods:

## Density based methods

Density based clustering results in clusters of dense regions (many observations), separated by regions of low density (few observations).

Several possible algorithms, e.g.:

- DBSCAN – based on discrete data points
- DENCLUE – based on kernel density
- OPTICS

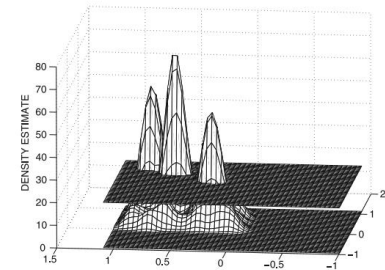


Figure 6.18: Density-based profile with lower density threshold

# Clustering methods:

## DBSCAN

1. **DBSCAN searches for clusters by inspecting the  $\epsilon$ -neighbourhood for each point in the data**
2. **If the  $\epsilon$ -neighbourhood of an object consist of more than minimum number of objects, a new cluster is created around the core object**
3. **DBSCAN iterately collects directly density reachable objects around the core objects and may merge density reachable clusters**
4. **Algorithm stops when there is no new points added to any cluster**

# Clustering methods:

## DBSCAN

### Relevant concepts:

**$\epsilon$ -neighbourhood:** radius  $\epsilon$  around an object  $A$

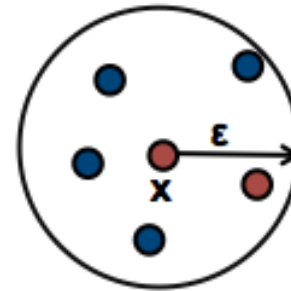
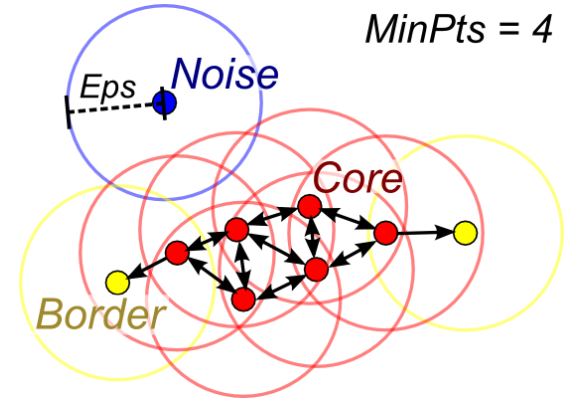
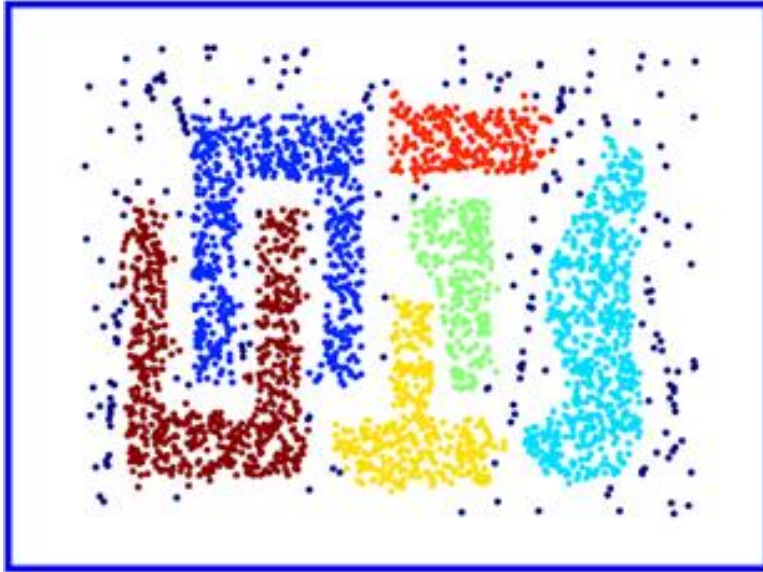
**Core object:** An object  $C$  with at least minimum number of objects within the specified  $\epsilon$ -neighbourhood

**Direct density reachable:** Point  $A$  is directly density reachable from point  $C$  if  $A$  is in the  $\epsilon$ -neighborhood of  $C$  and  $C$  is a core point

**Density reachable:** Point  $A$  is density reachable from  $B$  if there are a set of core points  $C$  leading from  $B$  to  $A$

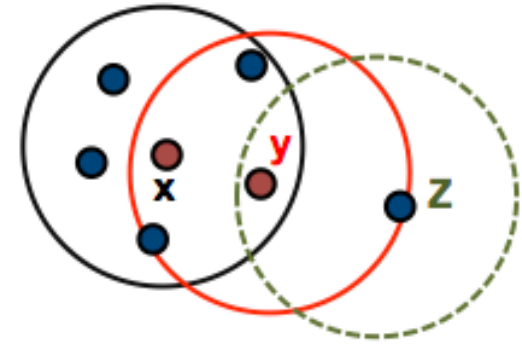
**Density connected:**  $A$  and  $B$  are density connected if there is a core point  $C$ , and both  $A$  and  $B$  are density reachable from  $C$

# Clustering methods: DBSCAN



MinPts = 6

(a)



(b)

# Spatial clustering AKA Regionalization

**As with other statistical measures, clustering can be done without spatial information (aspatially), and if location is known, clusters can be mapped.**

→ Clusters are spatially fragmented.

**In spatial data and patterns, location is a key element.**

→ Spatial autocorrelation or other measure of proximity or closeness should be incorporated. This includes possible physical obstacles in the distance!

# Regionalization methods

- 1. Optimization through trial and error:** starts with a random regionalization and iteratively improves the solution by switching the boundary objects between neighbouring regions.
- 2. Nonspatial multivariate clustering followed by spatial processing:** uses a general clustering method to derive clusters and then divides or merges the clusters to form regions.

# Regionalization methods

## **3. Clustering with spatially weighted dissimilarity measure:**

incorporates spatial information explicitly in the similarity measure for a general clustering method.

**4. Contiguity constrained clustering and partitioning:** Spatial information is included in a hierarchical clustering process by measures of contiguity.

# 3. Clustering with spatially weighted dissimilarity measure

**Simplest method is simply adding properly scaled coordinates as variables in clustering**

- Problem! There are two coordinates where as the spatial information should ideally be a single variable. How to transform 2D -> 1D?

**Spatial relationship can be either in the data model or in the algorithm.**

- Ward<sub>p</sub> method: dissimilarity measure is weighted by spatial distance or neighborhood information before clustering



# Exercise

**Form groups of 4, and answer the following questions**

1. Trial and error approach – how could geography be included in such an algorithm?
2. How would you postprocess non-spatial cluster solution to form spatial clusters?
3. How could geography be included in the dissimilarity computation used in clustering?

**Use your knowledge of clustering algorithms. Use any sources you need.**

# 1. Trial and error

1. **Select random seeds**
2. **For each observation which is not a seed**
  1. Find all seeds within a specified distance
  2. Assign observation to the seed's cluster with smallest multivariate distance
  3. If no seeds/clusters within specified distance, become seed

**For a real algorithm, see Automated Zoning Procedure(AZP) algorithm.**

## 2. How to spatially process clusters

**One possible solution:**

- 1. Cluster data, e.g. of countries of the world**
- 2. Split clusters using interesting areas, e.g. continents**
- 3. Compare clusters within a continent**
  1. Go through all clusters; split them if cluster quality measure is below threshold
  2. Merge clusters if their similarity is below threshold

See e.g. Fovell and Fovell, 1993, Climate Zones of the Conterminous United States Defined Using Cluster Analysis. Journal of American Meteorological Society

# 3. Geography incorporated to dissimilarity measure

Inspired by Spatial AutoRegressive (SAR) process

$$y = \rho W y + \beta X + \varepsilon$$

Spatial correlation

Weight matrix

1. Compute Moran's I for each variable
2. Compute spatial weights
3. Compute dissimilarity with

$$\rho_{\text{vector of Moran's I}} W X$$

OR

Distance measure:  $\rho_{\text{multivariate}} W \text{distance}$

Wartenberg D. 1985.

Multivariate spatial correlation: A method for exploratory geographical analysis, Geographical Analysis, 17: 263–283

# 4. Contiguity constrained clustering and partitioning

**Restrict potential cluster merging by some proximity measure**

**Spatial 'K'luster Analysis (SKATER):** hierarchical clustering where only neighboring clusters can be merged

→ Implemented in R package "spdep"; Will be applied in the assignment

→ Also implemented in ArcGIS

**REDCAP:** A family of six methods to merge neighboring clusters.

→ Implemented in a stand-alone software package, availability not clear from website. <http://www.spatialdatamining.org/software/redcap>

# Clustering methods: SKATER

**SKATER = Spatial 'k'luster  
Analysis by Tree Edge  
Removal**

**Based on a neighbourhood  
measure i.e. a graph!**

**→ All neighboring objects  
connected, OR e.g. a  
Minimum Spanning Tree.**

**Joins only neighboring  
objects.**

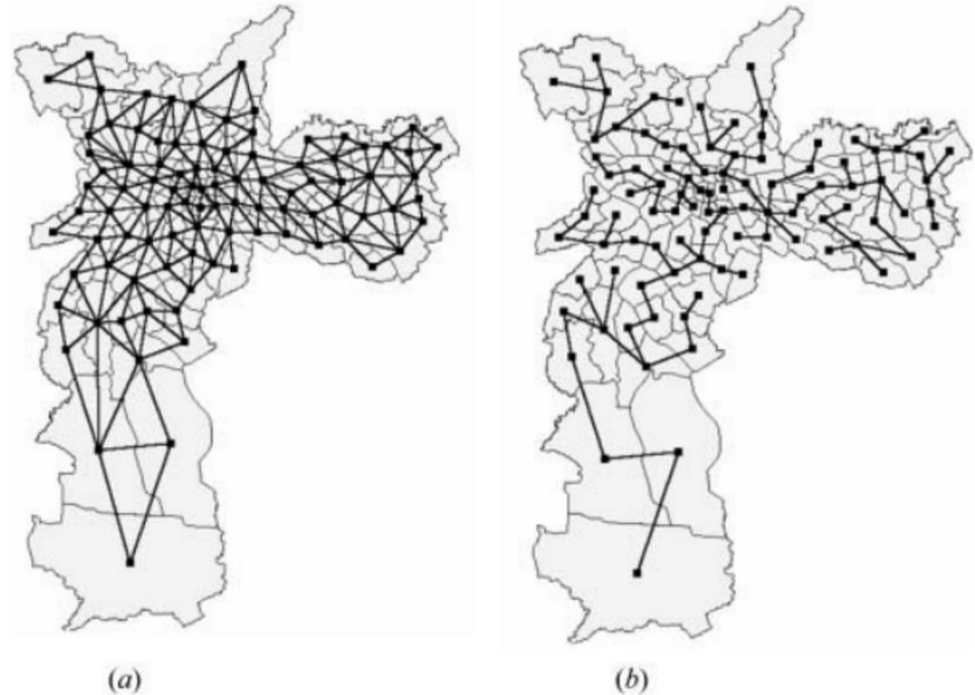


Figure 1. (a) Connectivity graph. (b) Minimum spanning tree.

# Clustering algorithms: SKATER

**SKATER = Spatial 'k'luster  
Analysis by Tree Edge  
Removal**

**Based on a neighbourhood  
measure i.e. a graph!**

**→ All neighboring objects  
connected, OR e.g. a  
Minimum Spanning Tree.**

**Joins only neighboring  
objects.**

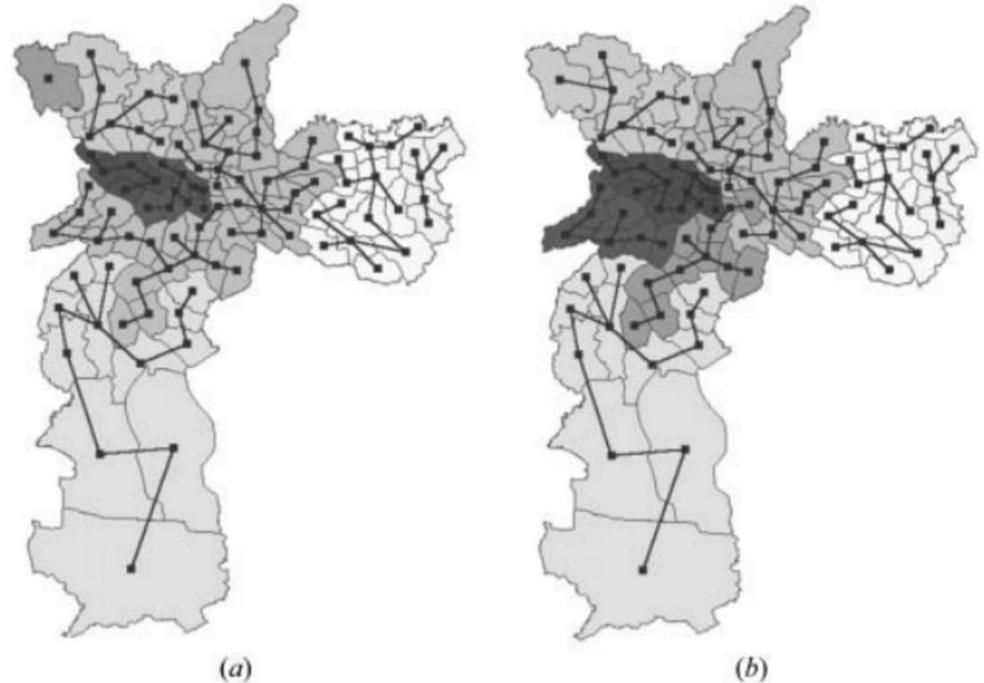


Figure 6. (a) Unrestricted regionalization. (b) Population-restricted regionalization.

# Other spatial clustering methods

**Spatial data can also be converted to timeseries data using Centroid Sweep method. This allows using timeseries clustering methods on spatial objects (shapes).**

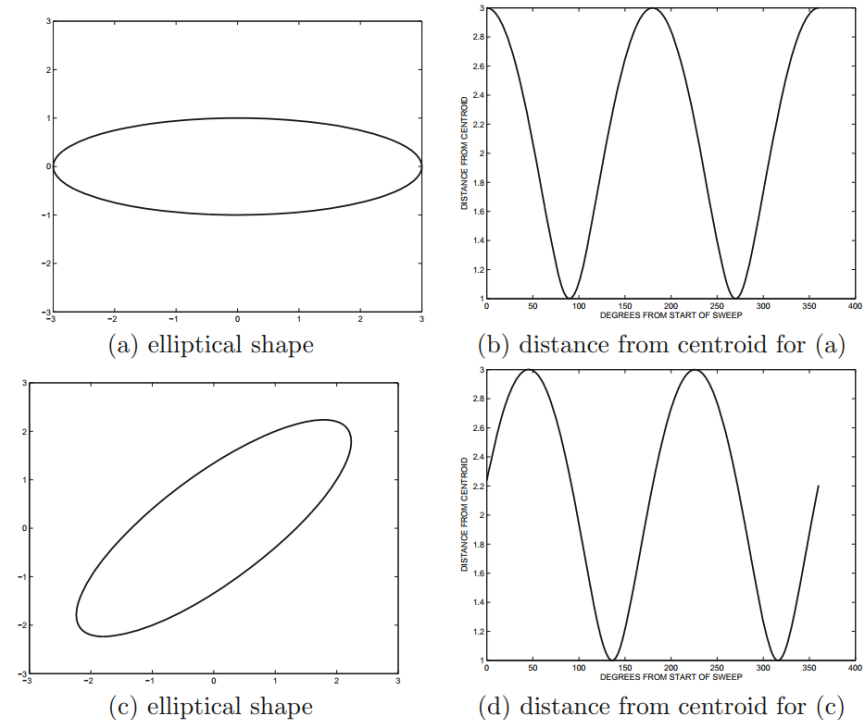


Figure 16.4: Conversion from shapes to time series



# Reading

## **Data Mining: The Textbook**

- Chapter 3 – Similarity and distances
- Chapter 6 – Cluster Analysis
- Optionally Chapter 7 – Cluster Analysis, Advanced Concepts

## **Geographic data mining and knowledge discovery**

- Chapter 12

**Other material provided in myCourses.**