Clustering

D G Rossiter

Cornell University, Soil & Crop Sciences Section

Nanjing Normal University, Geographic Sciences Department
南京师范大学地理学学院

March 18, 2022

Objective

- Given a set of objects with some **attributes** (measured properties) . . .
- . . . **group** the objects into groups = "clusters" . . .
- . . . such that members of each cluster are "similar" and the clusters are "dissimilar"
  - i.e., internally **homogeneous**, externally **heterogeneous**
- Two types:
  - **Centroid-based**: one set of k classes
  - **Hierarchical**: increasingly-general groupings, can form any number of classes from one hierarchy

What determines clustering?

- "nature" determines the degree to which clusters can/should be formed
  - is there a natural hierarchy or not?
  - how many clusters?
  - how "confused" are they?
- the analyst tries to find clusters that match this "natural" clustering
- clustering can also be for a pre-determined purpose (e.g., fixed number of sampling strata)

Examples

- Soil profiles: measurements of many properties at several depths
- People, households, census tracts . . . with attributes
- Space-time profiles of micropollutants in stream water [1]
- Metro stations: Time profiles of ridership; points of interest near stations

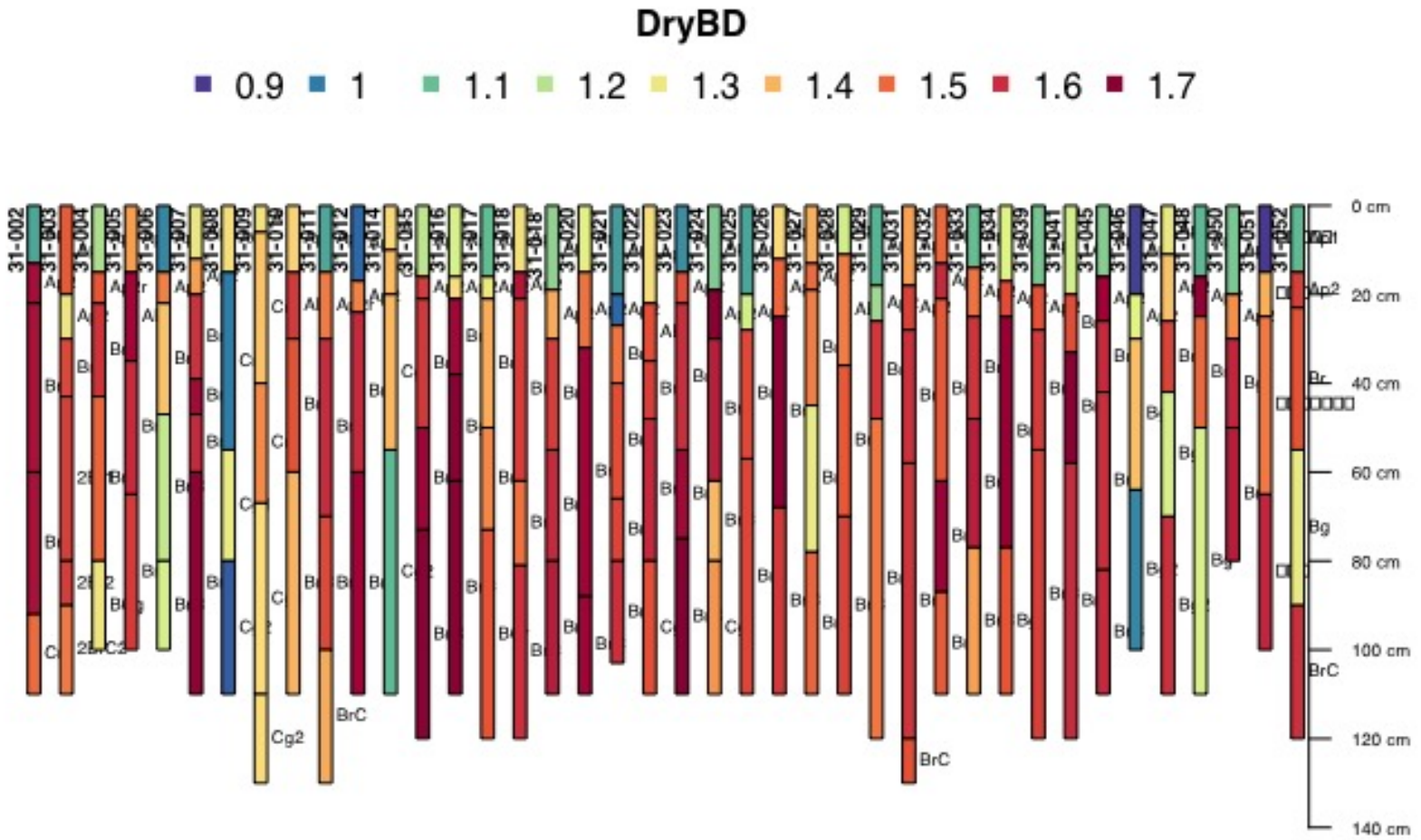
Questions

- How do we measure "similarity"?
- How do we build groups?
- How do we decide how many clusters k to make from n individuals?
  - hierarhical: where to cut the hierarchy
  - centroid-based: number to form

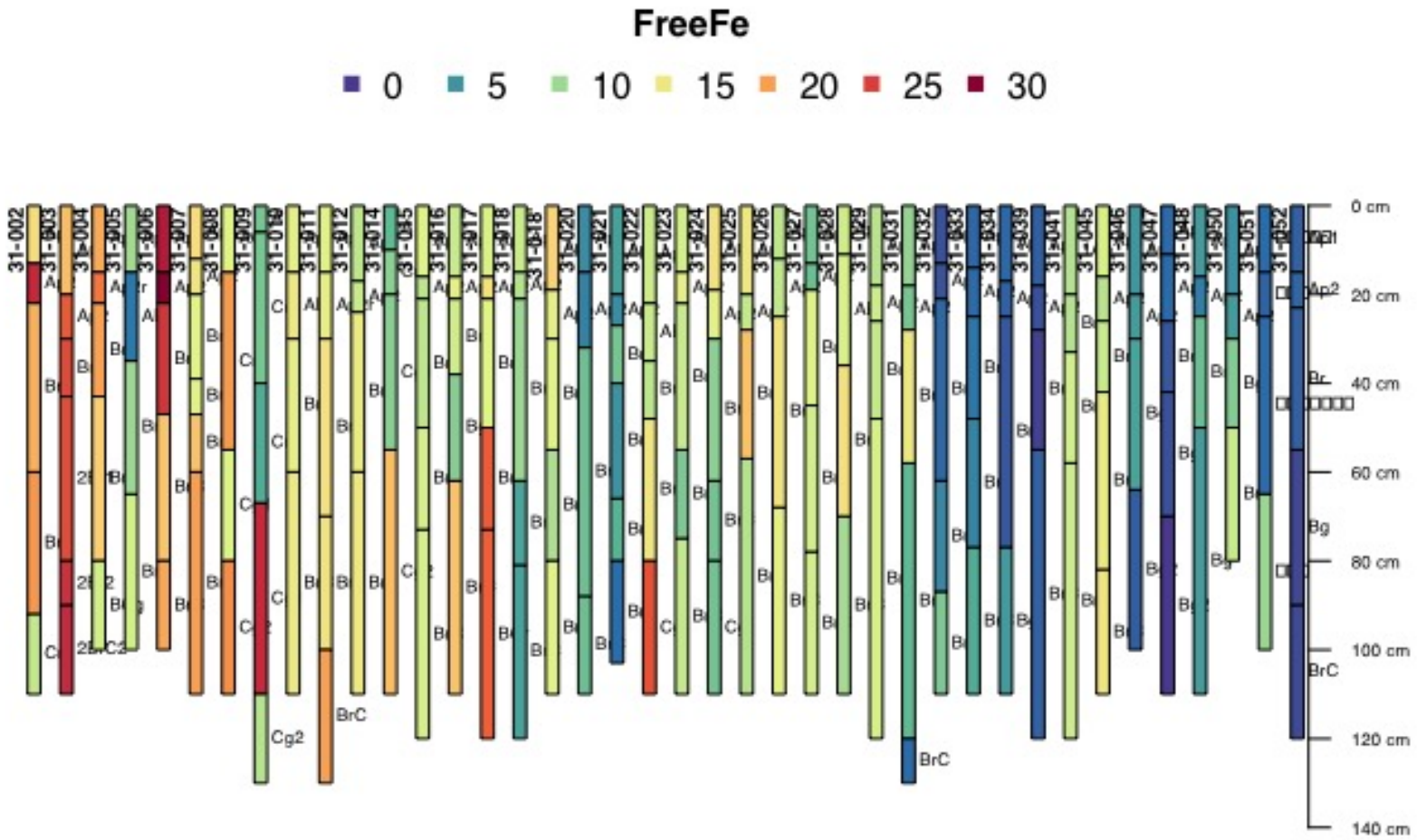
Example 1: soil profiles

- 40 soil profiles from Shanghai City
- 24 properties measured as averages or single values within surveyor-determined "genetic horizons" (layers)
- Genetic horizons *not* at fixed depths
  - need some way to harmonize these: by horizon type? by depth slice?
- Aim: cluster these "soil series" into functional groups, e.g., for management recommendations

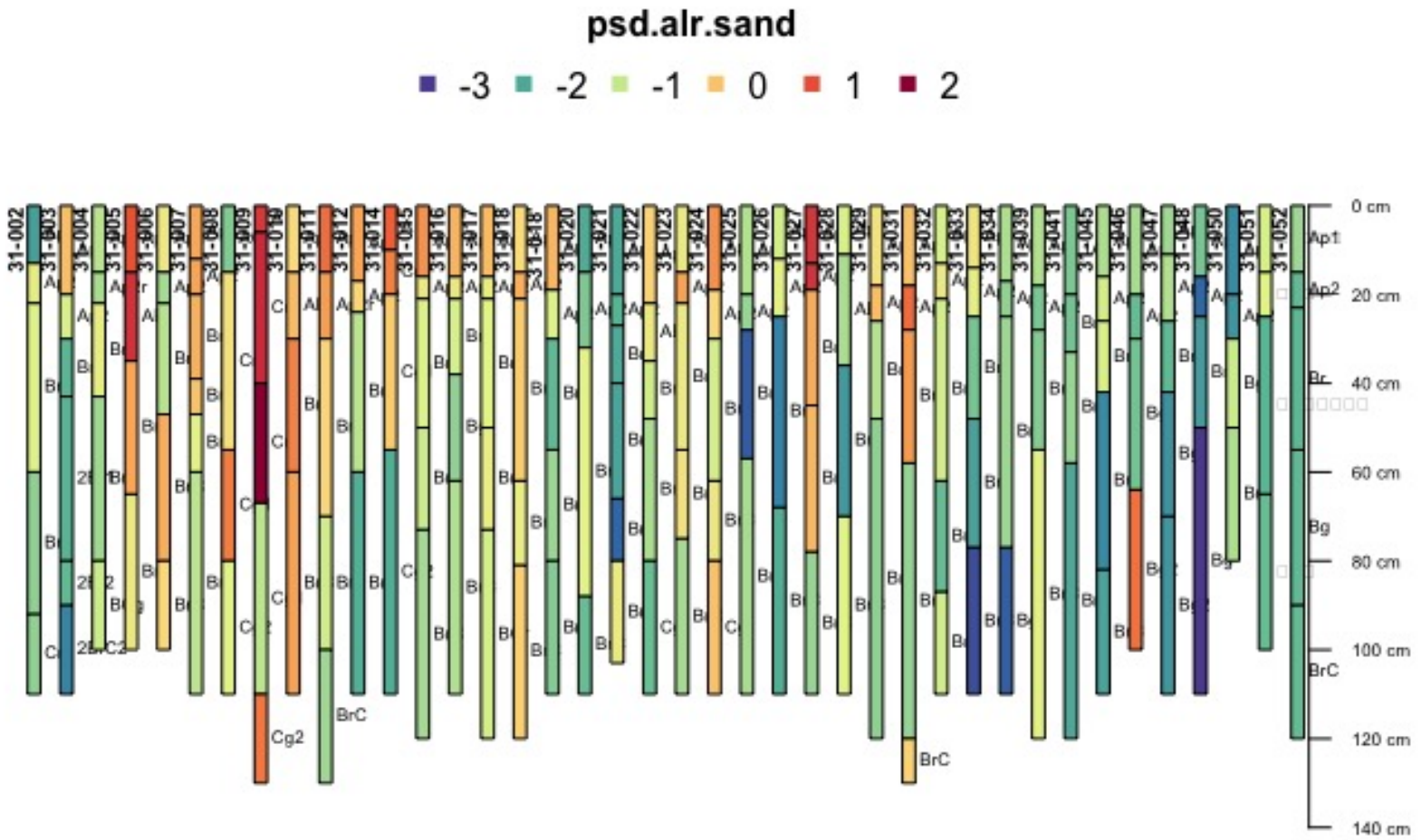
Property: bulk density by depth



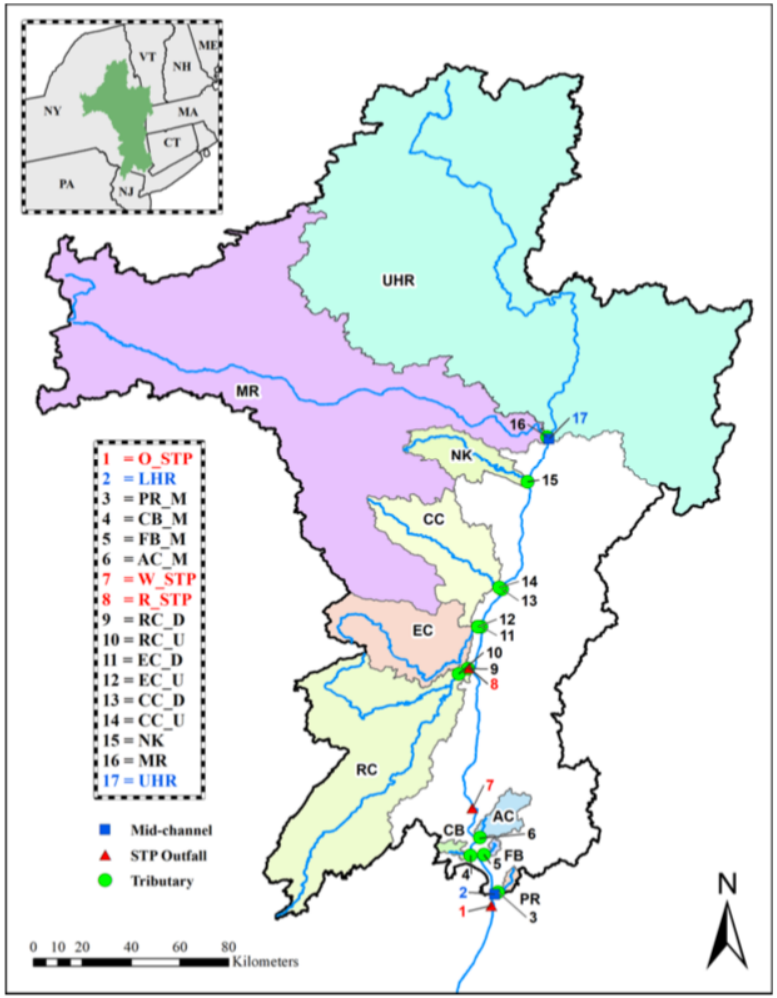
Property: free Fe by depth



Property: sand proportion (log ratio) by depth

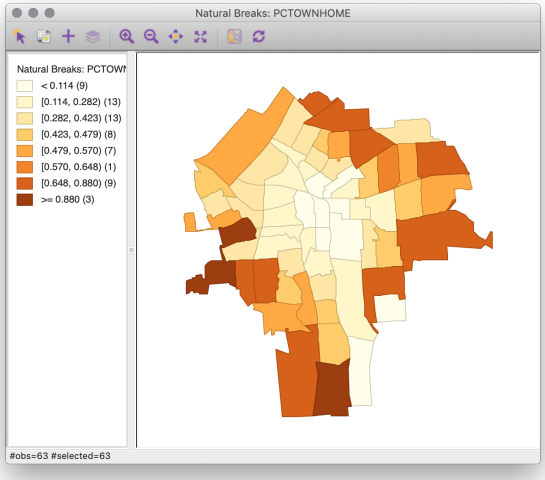


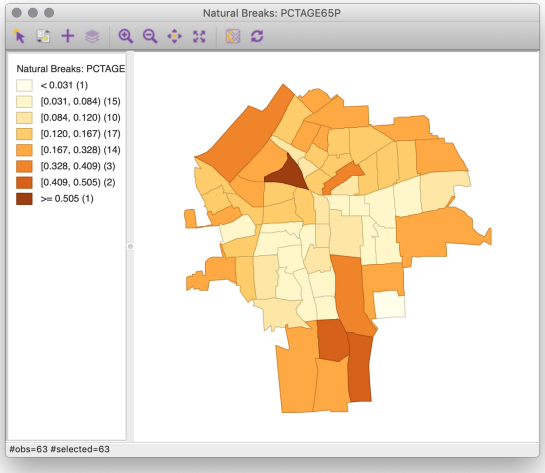
Example 2: spatio-temporal micropollutants

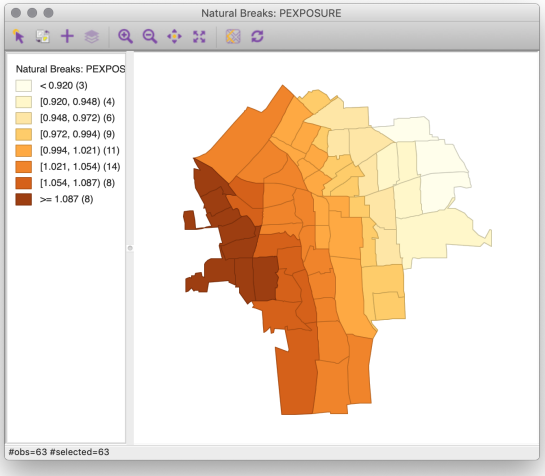


source: [1], Figure 1. 17 sites, 19 sampling times

Example 3: Syracuse (NY) census and health indicators







63 census tracts

demographic variables “proportion older than 65 years”, “proportion own their home”

health-related variable “potential exposure to carcinogenic chemicals”

Leukemia incidence and possible feature-space predictors

PEXPOSURE exposure to TCE (tricholoroethylene)[1] sources

- toxic chemical linked to cancer

PCTAGE65 % of residents $>$ 65 years old

- cancer incidence may increase with age

PCTOWNHOME % of homes owned

- wealthier =? better health care? less likely to have worked in a chemical plant?

[1] https://www.cdc.gov/niosh/topics/trichloroethylene/default.html

Centroid-based clustering

- Form groups around a **centroid** of the group
  - one of the objects in the group, the **exemplar**
  - anywhere in feature space: **mean** or **median** of the group
- Not hierarchical
- The number of clusters is selected by the analyst
  - external reason for that number, e.g., sampling strata
  - or, various methods to "optimize", see below
- Hastie et al. [2] Chapter 13 "Unsupervised Classification"
- James et al. [3] a simplified explanation

k-means

- Centroids are the **means** of a group of objects
  - Finds the approximate "centres" of the clusters in **univariate** or **multivariate** space
  - Criterion: *minimize* **within-cluster** variance; *maximize* between-cluster variance
  - No deterministic solution, must iterate; can get stuck in local optima
  - Allocates and then re-allocates individuals to clusters; re-computes centroids

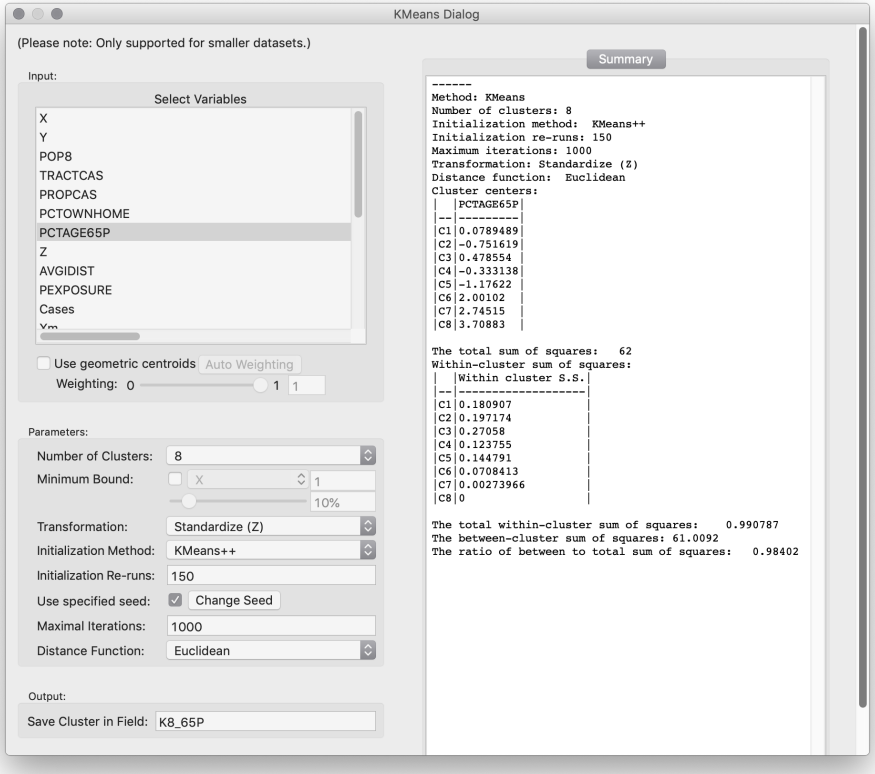
k-medians

- Like k-means, but optimizes the median of the cluster
- Non-parametric, avoids problems with extreme values

k-mediods

- like k-means or k-medians, but the centroid must be one of the observations
- this is called the **exemplar**
- it is “the most typical” representative of the cluster
- useful if you want to select a representative observation in a cluster

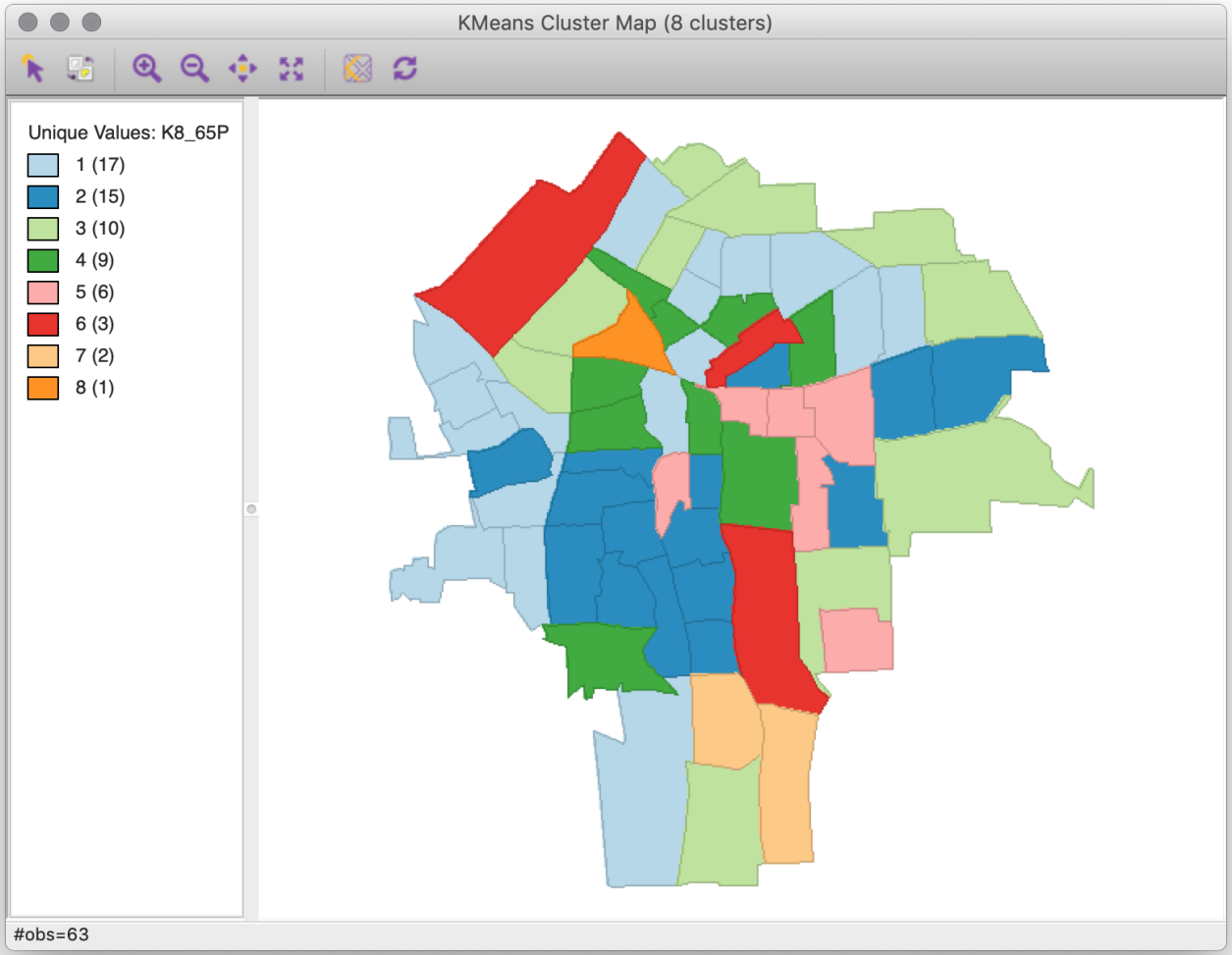
Clustering: univariate k-means, Syracuse (1)



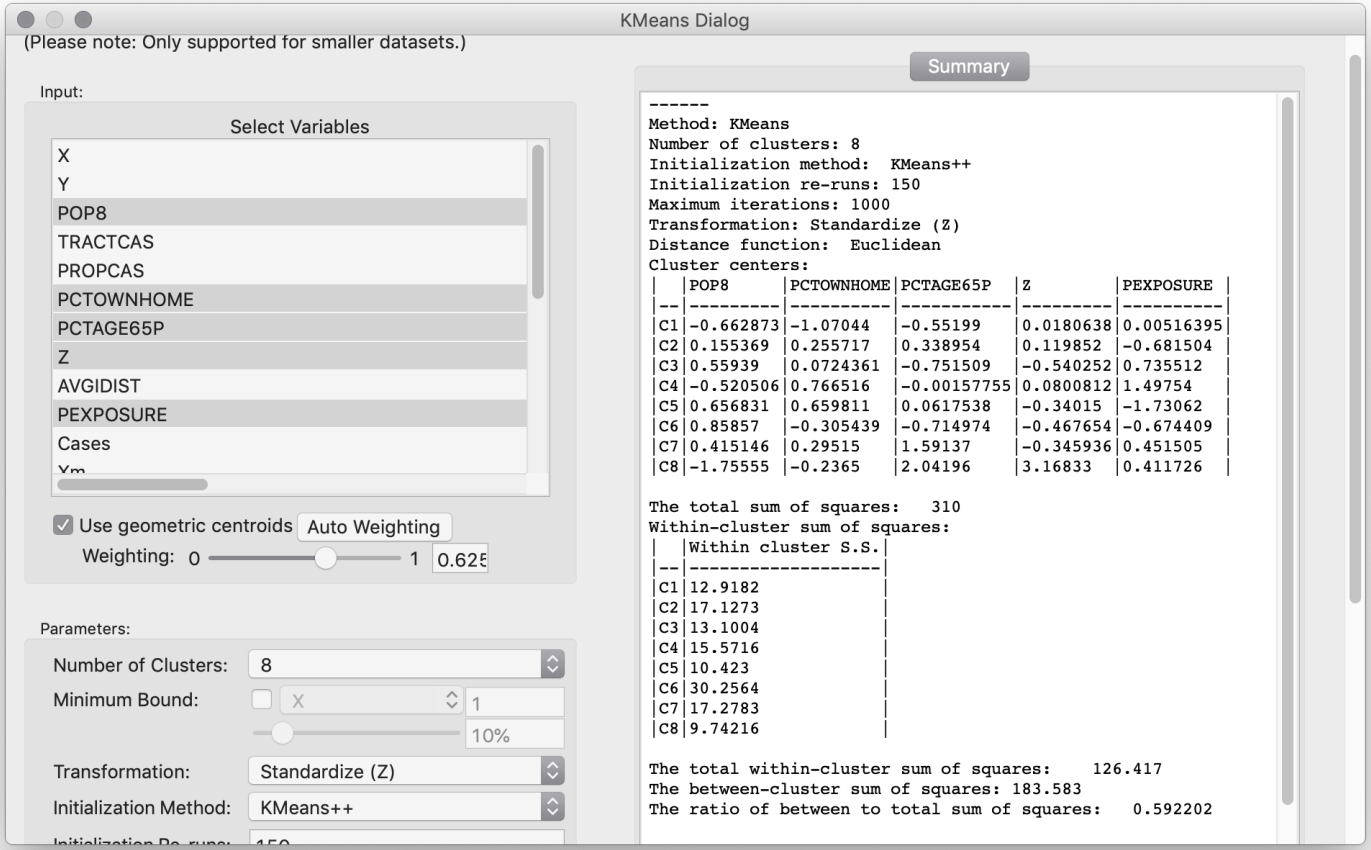
Minimize the within-class/maximize between-class variance of proportion 65+ years old

Cluster centres; within-cluster standard deviations; ratio of between to total sum of squares

Clustering: univariate k-means, Syracuse (2)



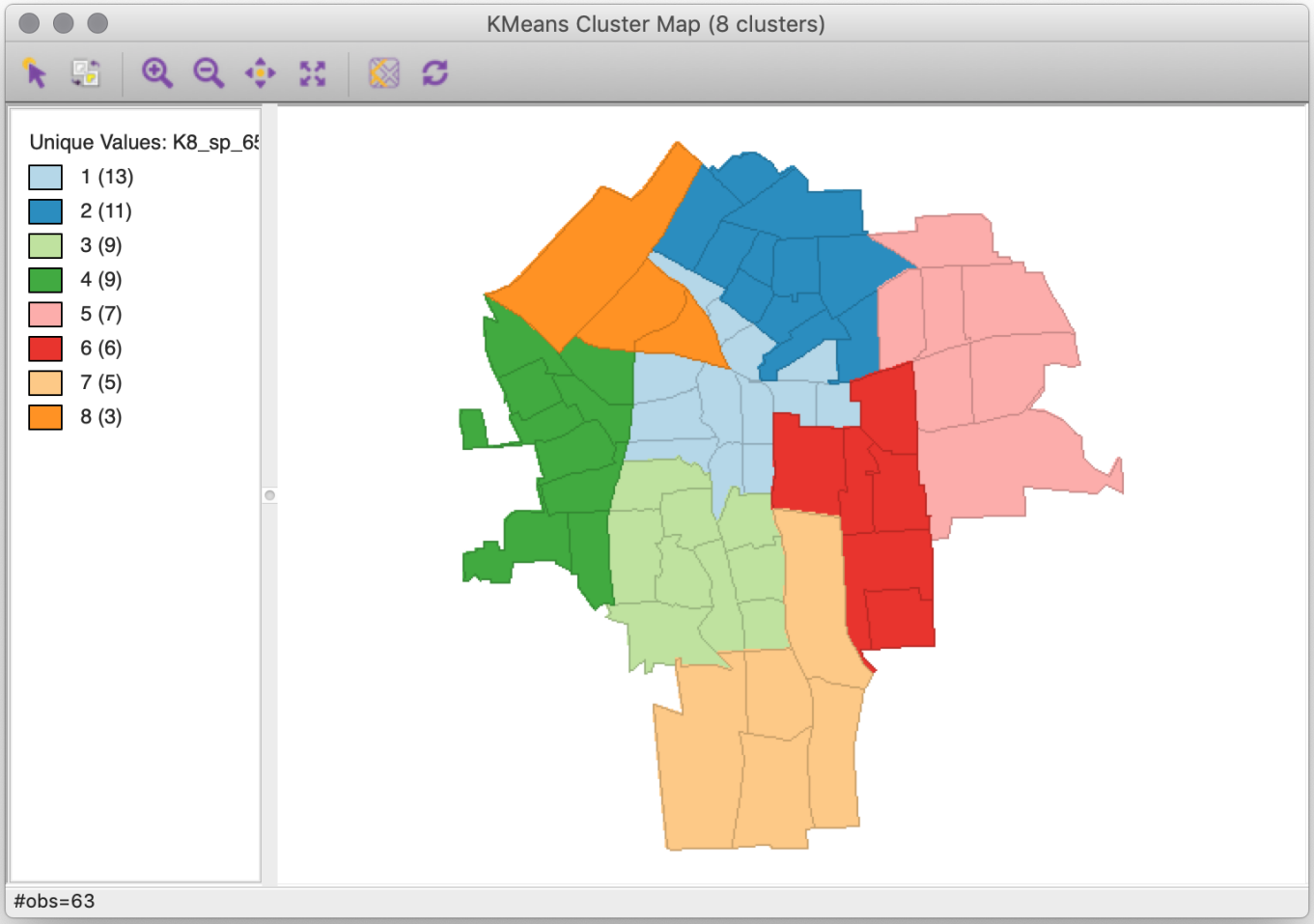
Clustering: multivariate geographic k-means (1)



Algorithm to minimize the within-class/between-class variance, while forcing clusters to be **spatially-continguous**

Also uses **multivariate** distances

Clustering: multivariate geographic k-means (2)



Hierarchical clustering

- increasingly-general groupings
- can form any number of classes from one hierarchy
- reveals “distances” in feature space between objects

Top-down vs. bottom-up

Top-down ("divisive") split the entire set into two groups, then these groups into two ...

Bottom-up ("agglomerative") group two individuals into a group, then build larger groups

- clusters at each level of the hierarchy are created by merging clusters at the next lower level.
- several "linkage" methods to merge lower-level clusters
- must specify a measure of **pairwise dissimilarities** among any two observations
- must specify a measure of **group dissimilarity** between (disjoint) groups of observations,

Pairwise dissimilarity

- "Distance" in multivariate attribute space
  - Euclidian
  - Mahalanobis (takes into account variance/covariance of attributes)
- Generally standardize all attributes to mean 0, standard deviation 1, to give equal weight
- But can use centred original values or some other weighting method

Linkage methods

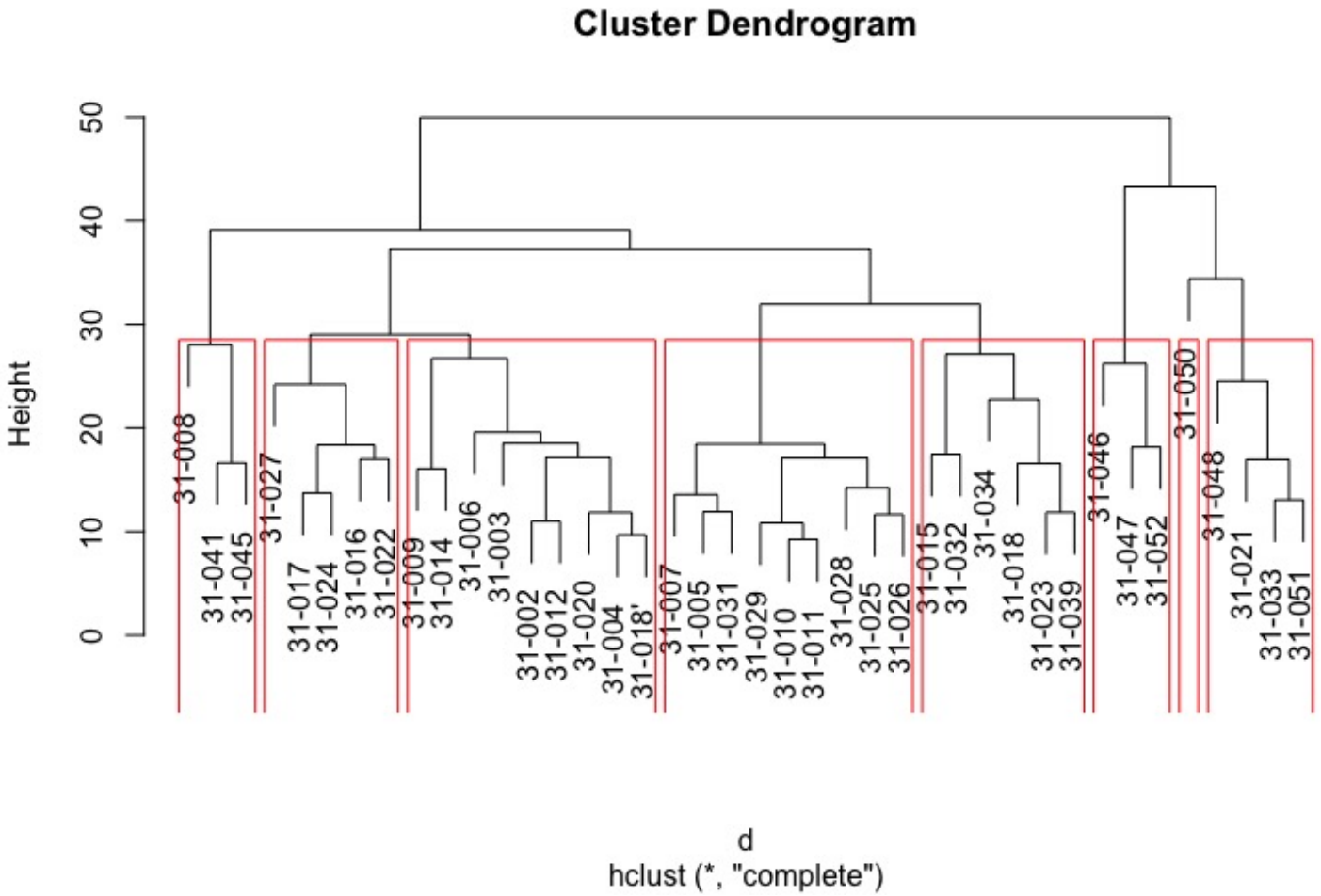
Aim: find two groups to merge, considering all groups aleady formed.
Note that "groups" here include single observations

single linkage "nearest neighbour": most similar individuals within the two groups

complete linkage "furthest neighbour": most dissimilar pair of individuals within the two groups

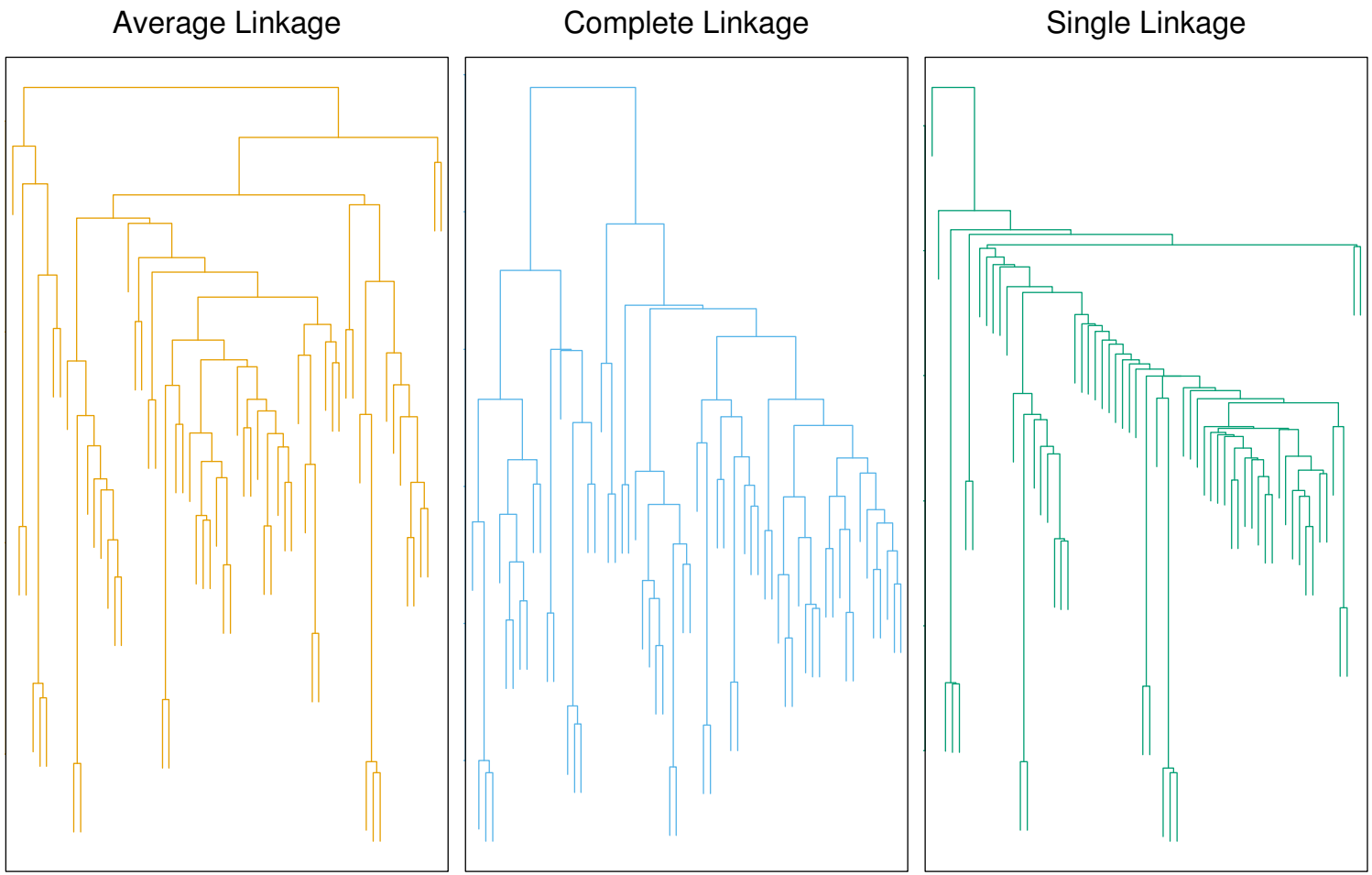
group average linkeage average dissimilarity between all observations in one group with all observations in the other

Dendrogram – complete linkage



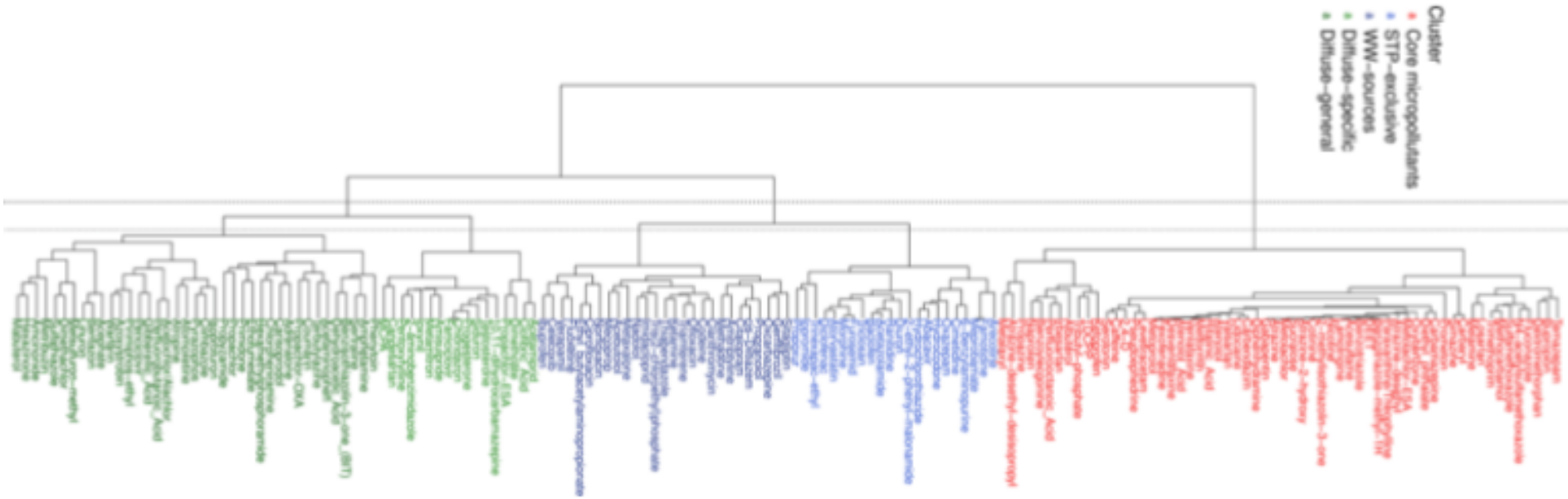
Vertical scale is the dissimilarity measure

Results of different linkage strategies



source: [3]

Spatio-temporal hierarchical clusters



source: [1], Figure 2

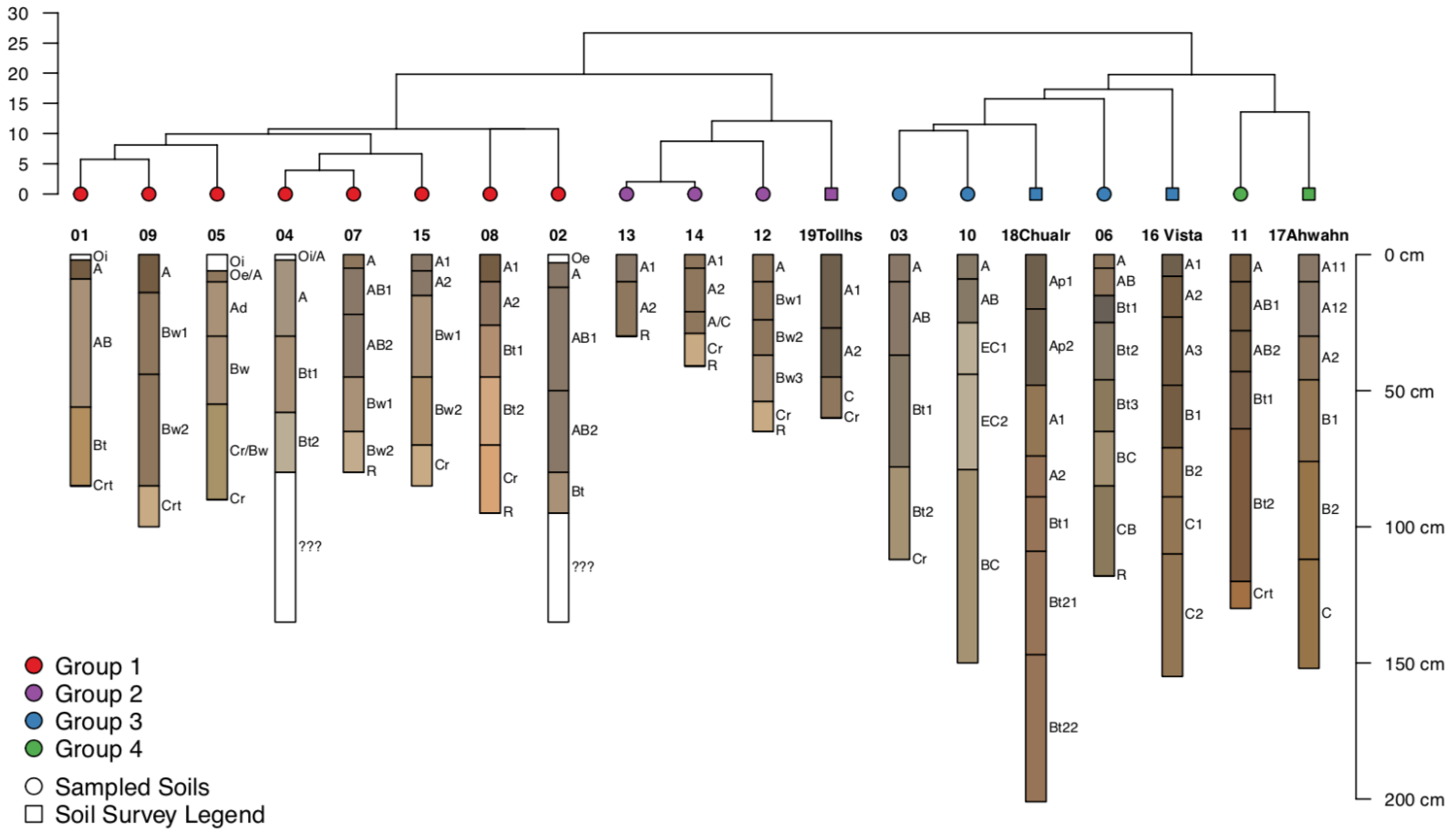
core micropollutants

sewage treatment plant source

diffuse

Numerical Soil Classification (aqp R package)

Numerical soil classification

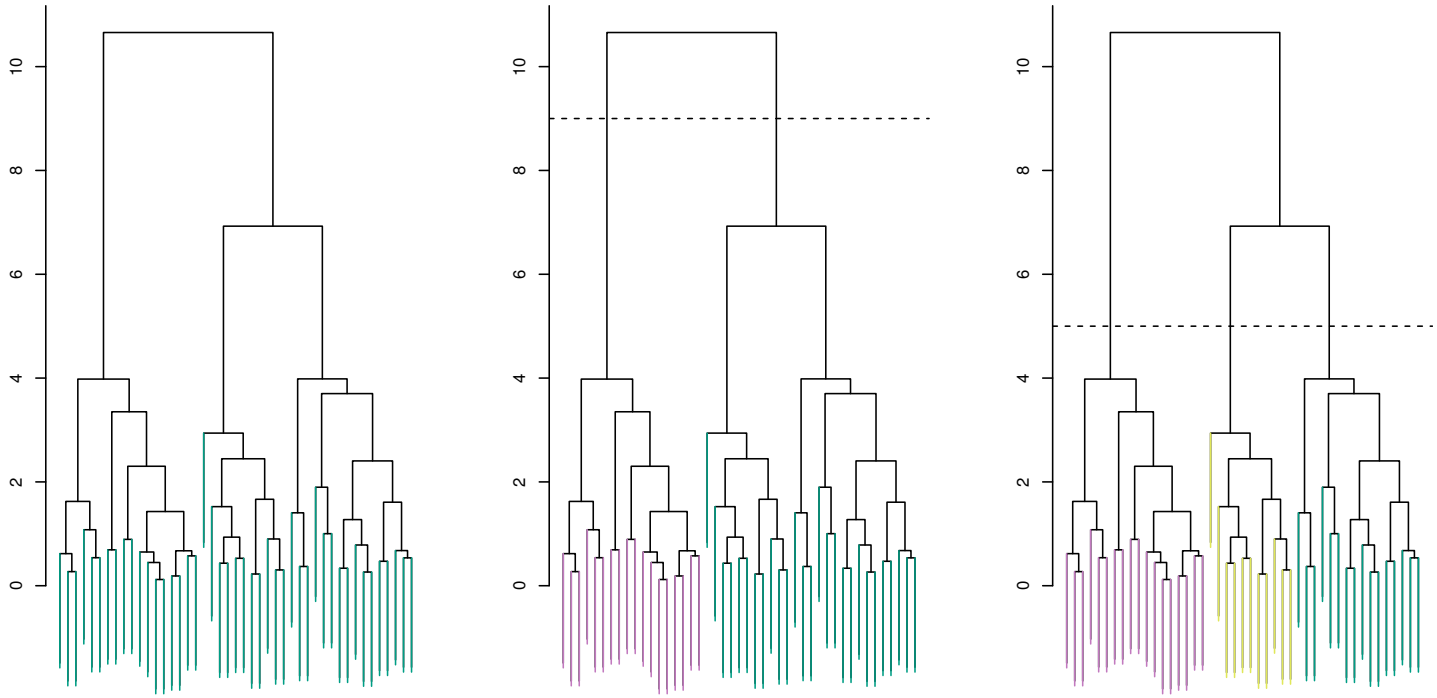


Multivariate hierarchical, shows "distance" between objects

Forming groups

- Can decide how many groups, and cut the dendrogram at the level to produce that number
- Or, can decide how disimilar groups must be, and cut the dendrogram at that level

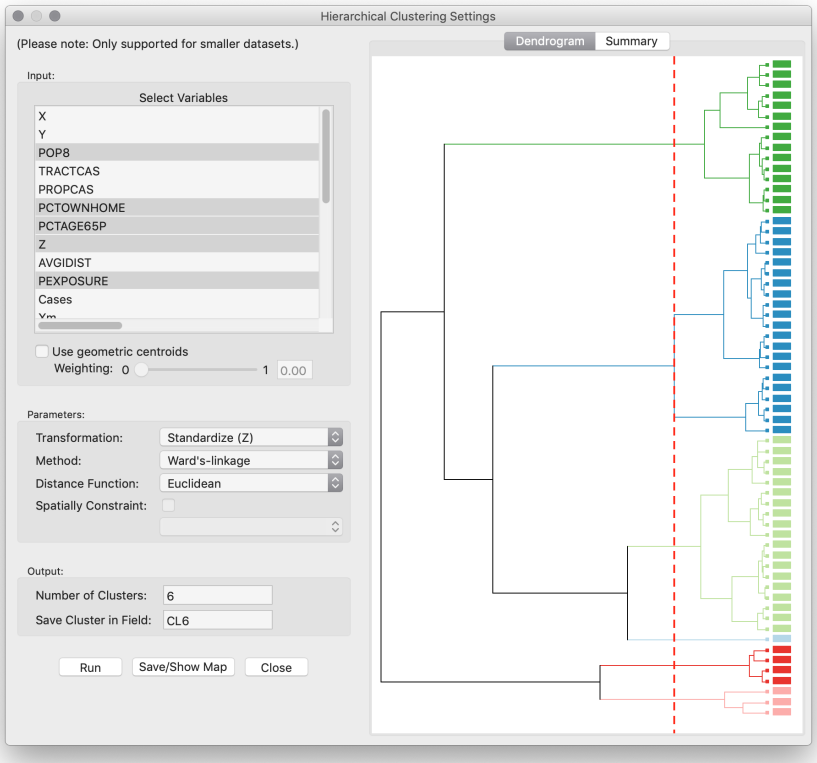
Forming different numbers of groups from one dendrogram

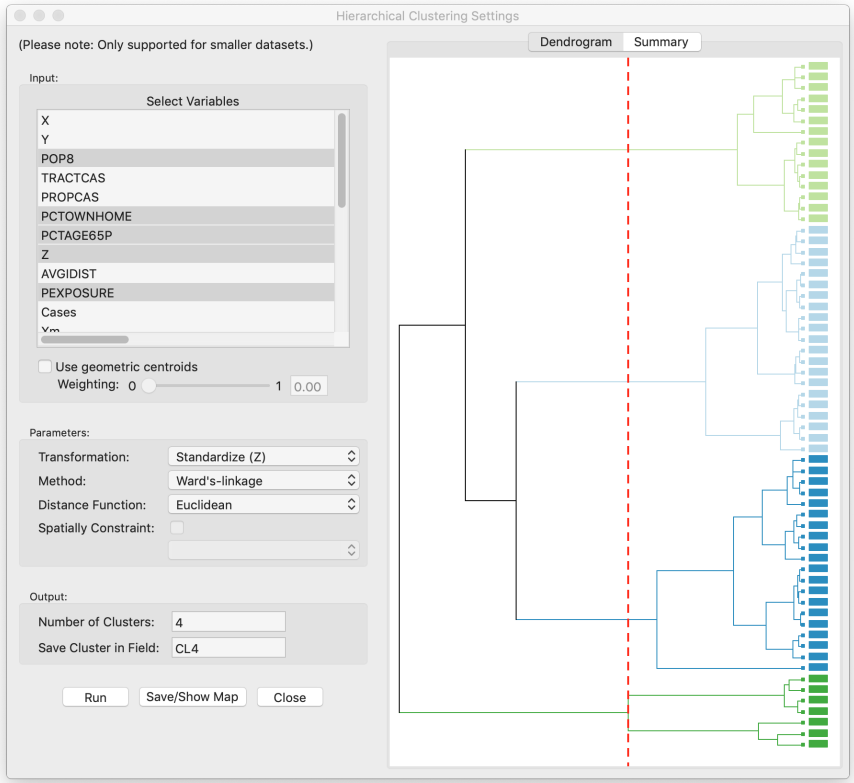


source: [3]

Example: Syracuse census and health

Clustering: multivariate hierarchical: specification and dendrogram





Group at any level of detail; see “distance” between groups in multivariate attribute space

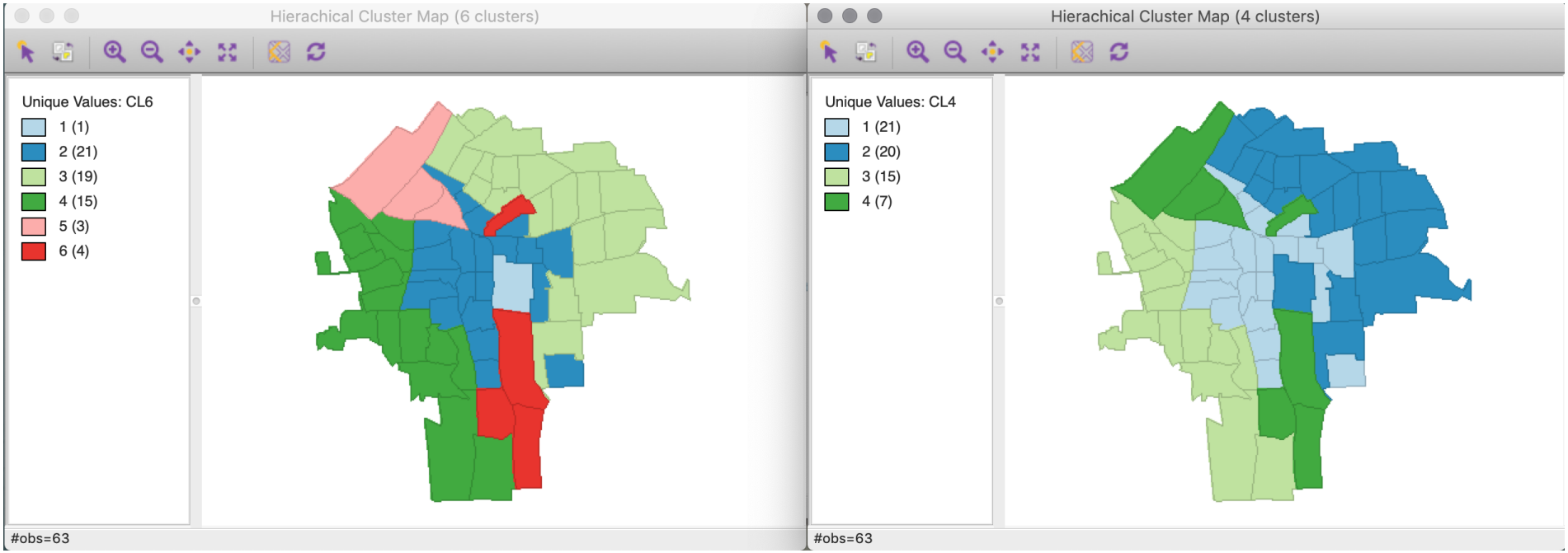
Clustering: cluster statistics

```
Number of clusters: 6
Transformation: Standardize (Z)
Method: Ward's-linkage
Distance function:  Euclidean
Cluster centers:
|   |POP8      |PCTOWNHOME|PCTAGE65P|Z          |PEXPOSURE   |
|---|----------|----------|---------|-----------|------------|
|C1|4.49078   |-1.46223   |-0.430876|-0.637778  |-0.309463   |
|C2|-0.205116|-0.867063  |-0.709461|-0.2613    |0.0199737   |
|C3|0.345393  |0.56543    |0.0892871|-0.13715   |-1.08593    |
|C4|-0.173746|0.847444   |-0.137512|-0.0483164|1.28624     |
|C5|-1.75555  |-0.2365    |2.04196  |3.16833    |0.411726    |
|C6|0.281756  |-0.768694  |2.39247  |-0.012324  |-0.00152647|

The total sum of squares:   310
Within-cluster sum of squares:
|   |Within cluster S.S.|
|---|-------------------|
|C1|0                   |
|C2|36.8377             |
|C3|33.667              |
|C4|27.6518             |
|C5|9.74216             |
|C6|4.68514             |

The total within-cluster sum of squares:    112.584
The between-cluster sum of squares: 197.416
The ratio of between to total sum of squares:   0.636827
```

Clustering: multivariate hierarchical: maps



Different levels of detail.

“Optimum” number of clusters

internal based on the between- and within-cluster variances

- How well can clustering partition the dataset?
- When does too many clusters result in partitioning “noise”, not structure?

external based on the match of proposed clusters with some external classification

- How well does numerical clustering match predefined clusters?
- e.g., predefined soil classes

Internal optimization

- R package `NbClust`; 30 indices
- e.g. "silhouette" index: $\frac{1}{n}\sum_{i=1}^{n} S(i), \in [-1, 1]$ where:
  - $S(i) = \frac{b(i)-a(i)}{\max\{a(i);b(i)\}}$ where:
    - $a(i) = \frac{\sum_{j \in \{C_r \setminus i\}}}{n_r - 1} d_{ij}$: the average dissimilarity of the ith object to all *other* objects of cluster C_r
    - $b(i) = \min_{s \neq r} \frac{\sum_{j \in C_S} d_{ij}}{n_s}$: the average dissimilarity of the ith object to all objects of cluster C_s
    - i is a single object, of n total, that has been clustered into cluster C_r
    - j is a single object, of n total, that has been clustered into class C_s
    - d_{ij} is the distance in attribute space between two objects
- Choose the maximum value of the index.

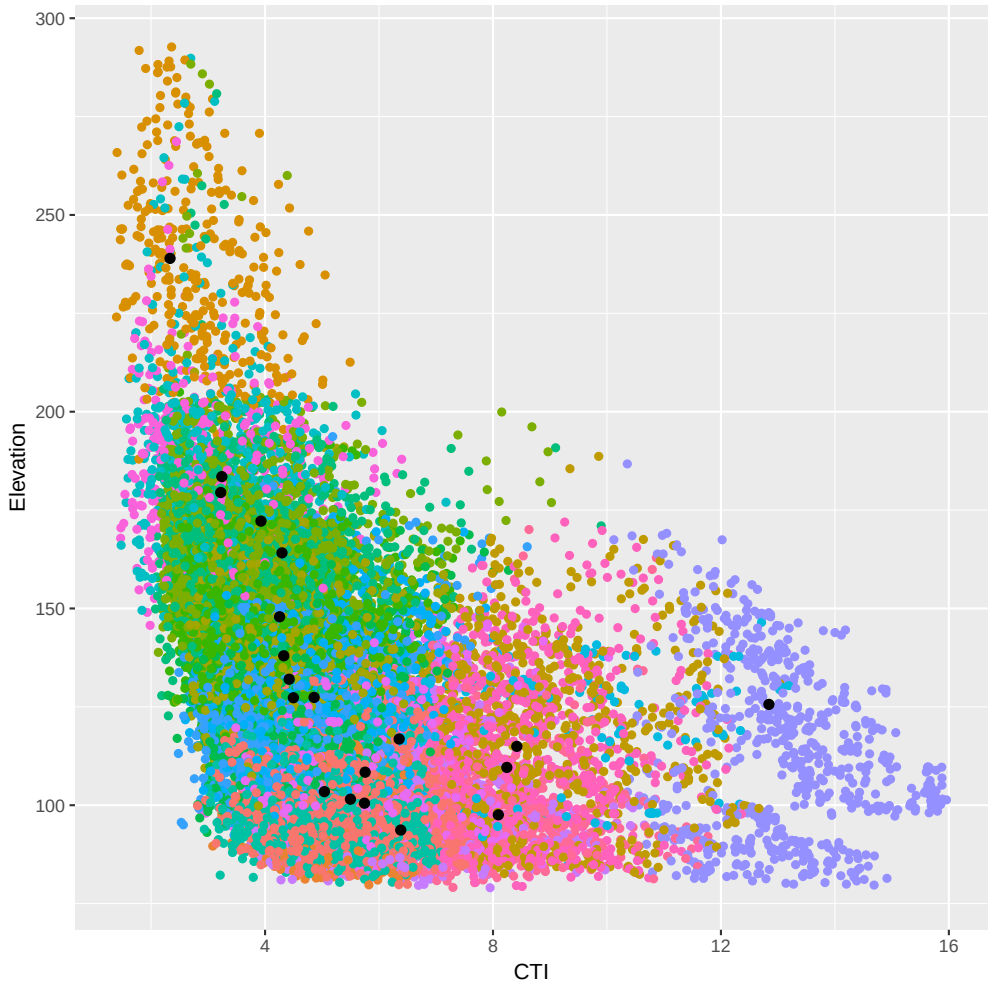
External optimization

- R package fpc "flexible procedures for clustering"
- adjusted Rand index (ARI)
  - range from -1 = random assignment to +1 = perfect agreement

$$\text{ARI} = \frac{\sum_{ij}\binom{n}{2} - \left[\sum_i \binom{n_i}{2} \sum_j \binom{n_j}{2}\right] / \binom{n}{2}}{\left[\sum_i \binom{n_i}{2} + \sum_j \binom{n_j}{2}\right] / 2 - \left[\sum_i \binom{n_i}{2} \sum_j \binom{n_j}{2}\right] / \binom{n}{2}} \quad (1)$$

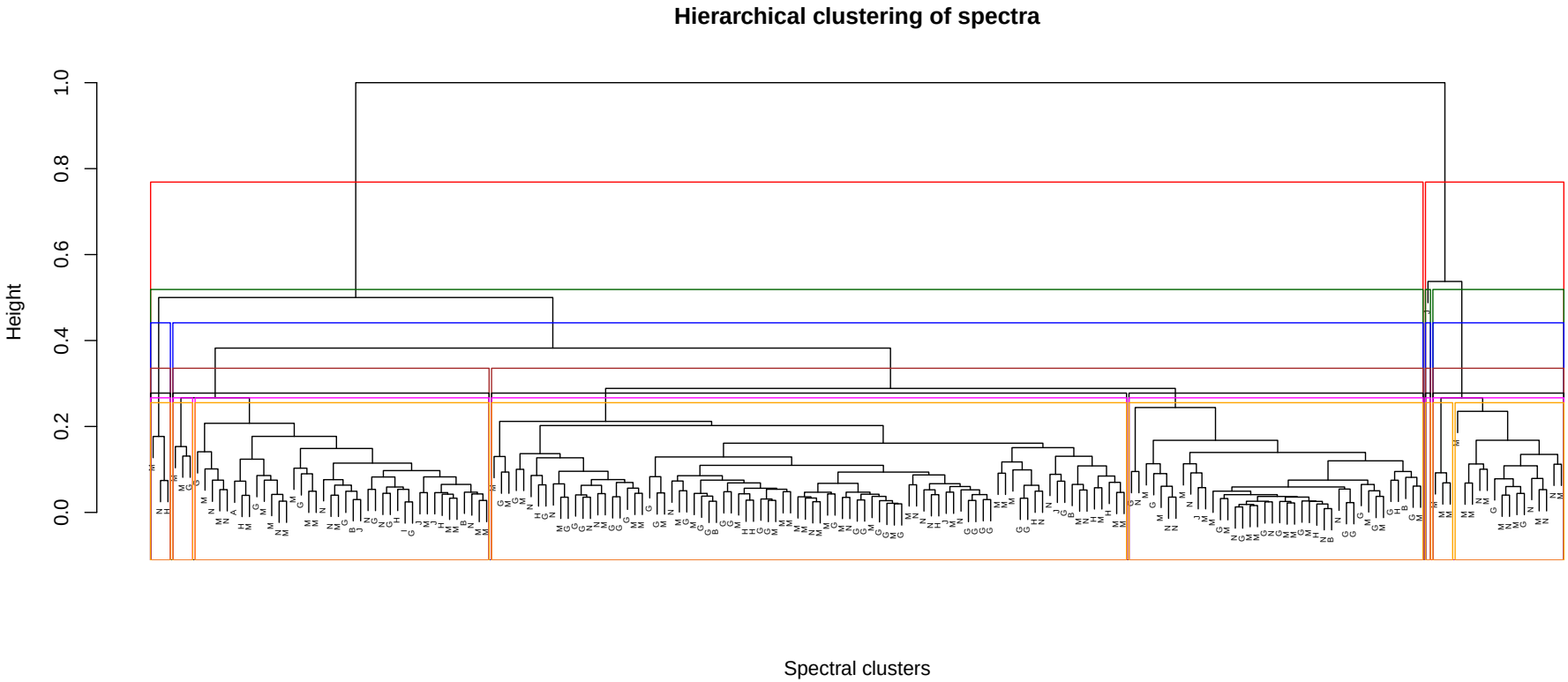
where n_{ij} is the number of observations in cluster i and class j, n_i is the total observations in cluster i and n_j is the total observation in class j.

Example: k-means clustering



Hunter Valley (NSW) landscape clustered by covariates related to soil geography. Colours are clusters.
Example of over-clustering – overlap in feature space

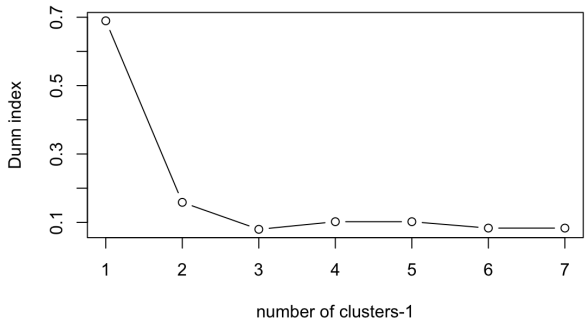
Example: Hierarchical clustering

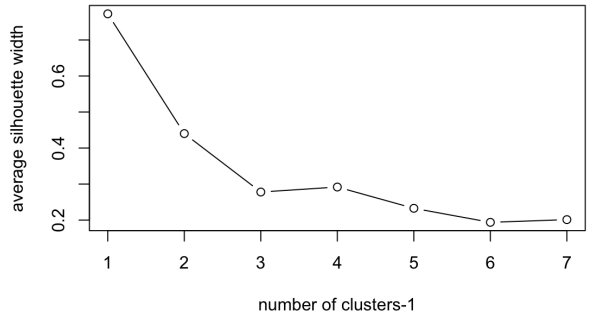


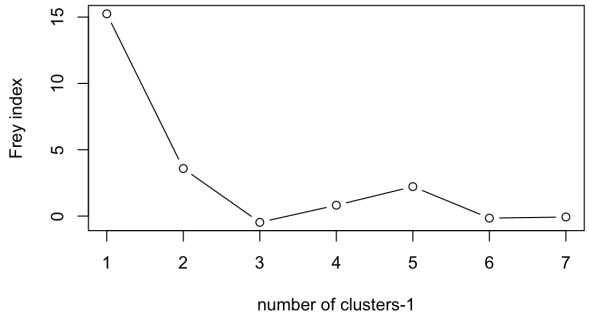
source: [4]

2–8 clusters, depending on where the dendrogram is cut – which is "optimum"?

Internal optimization indices







Dunn Silhouette Frey

2 to 4 clusters are “optimal” by these internal measures

External optimization

	B	J	G	N	A	I	H	M
1	0	0	9	7	0	0	0	7
2	32	11	199	84	0	3	61	252
3	0	27	0	19	5	0	2	79
4	0	0	22	10	0	0	9	0

Cross-classification, e.g., 4 spectral clusters vs. 8 soil orders

Adjusted Rand Index: 0.002; 0.047; 0.069; 0.068; 0.063; 0.092; 0.091
for 2 – 8 spectral classes

References I

[1] Corey M. G. Carpenter and Damian E. Helbling. Widespread micropollutant monitoring in the Hudson River estuary reveals spatiotemporal micropollutant clusters and their sources. *Environmental Science & Technology*, 52(11):6187–6196, Jun 2018. doi: 10.1021/acs.est.8b00945.

[2] Trevor Hastie, Robert Tibshirani, and J. H Friedman. *The elements of statistical learning data mining, inference, and prediction*. Springer series in statistics. Springer, 2nd ed edition, 2009. ISBN 978-0-387-84858-7.

[3] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning: with applications in R*. Springer texts in statistics. Springer, 2013. ISBN 978-1-4614-7137-0.

[4] R. Zeng, D. G. Rossiter, and G. L. Zhang. How compatible are numerical classifications based on whole-profile vis–NIR spectra and the Chinese Soil Taxonomy? *European Journal of Soil Science*, 70(1):54–65, 2019. doi: 10.1111/ejss.12771.