Stability of Cluster Analysis

Matthias Templ & Peter Filzmoser
Vienna University of Technology

Vienna, June 16, 2006

Stability of Cluster Analysis

For real data sets without obvious grouping structure the stability of clusters depends on:

1. **Input data - the selection of variables**
2. **Preparation of the data**
3. **Distance measure used ***
4. **Clustering method**
5. **Number of clusters**

Changing one parameter may result in complete different cluster results.

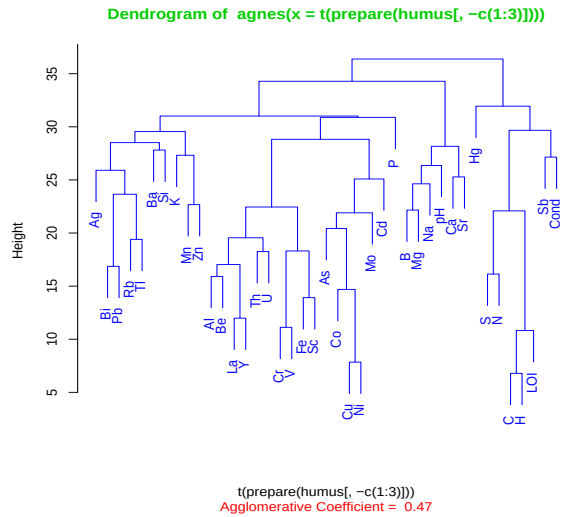
*if a distance measure must be chosen

1. Input Data - Variable Selection

```
> library(mvoutlier)
> library(cluster)
> data(humus)
> a <- agnes(t(prepare(humus[, -c(1:3)])))
> plot(a, which.plots = 2, col = c(4), col.main = 3, col.sub = 2)
```

1. Input Data - Variable Selection

```
> library(mvoutlier)
> data(humus)
> a <- agnes(t(prepare(humus[, -c(1:3)])))
> plot(a, which.plots = 2, col = c(4), col.main = 3, col.sub = 2)
```



A chemical process can be seen in more detail in a map (later) by choosing similar variables.

1. Input Data - Variable Selection

A selection of variables may be useful when clustering high-dimensional data because ...

- clustering with all variables may hide underlying processes
- we want to see some processes in more detail
- the inclusion of one irrelevant variable may hide the real clusters in the data

One easy way (amongst others) for variable selection can be done by graphical inspection of a dendrogram which results from hierarchical clustering of variables.

2. Data Preparation

Most of real data in practice can have some or all of these properties:

- neither normal nor log-normal
- strongly skewed
- often multi-modal distributions
- dependencies between observations
- weak clustering structures
- data includes outliers
- variables show a striking difference in the amount of variability

2. Data Preparation

If a good clustering structure for a variable exists we expect a distribution with two or more modes. A **transformation** (e.g with a box-cox transformation) will preserve the modes but remove large skewness.

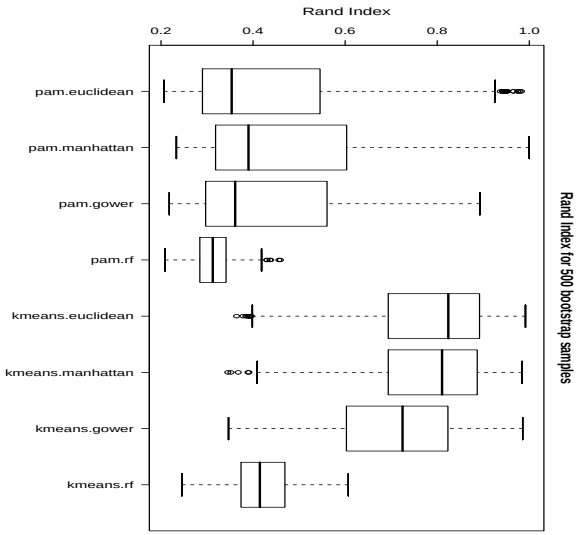
Standardisation of the variables is needed if the variables show a striking difference in the amount of variability.

Outliers can influence the clustering (depends on which clustering algorithms is chosen)

Removing outliers before clustering may be useful.

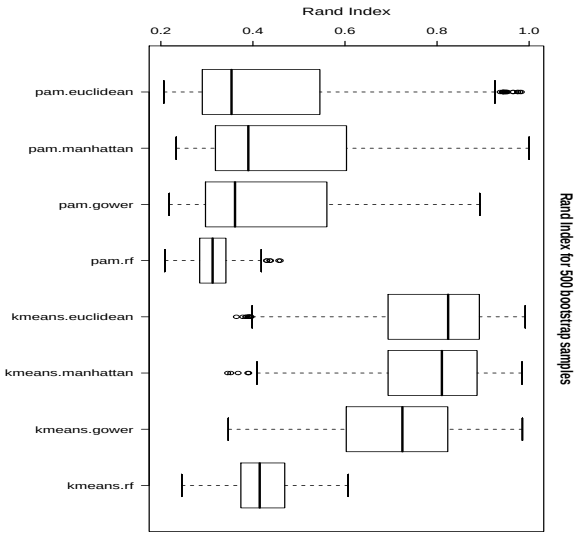
Finding outliers is not a trivial task, especially in high dimensions. (you can do this e.g. with Package `mvoutlier` from Filzmoser et al. (2005))

3. Distance Measure



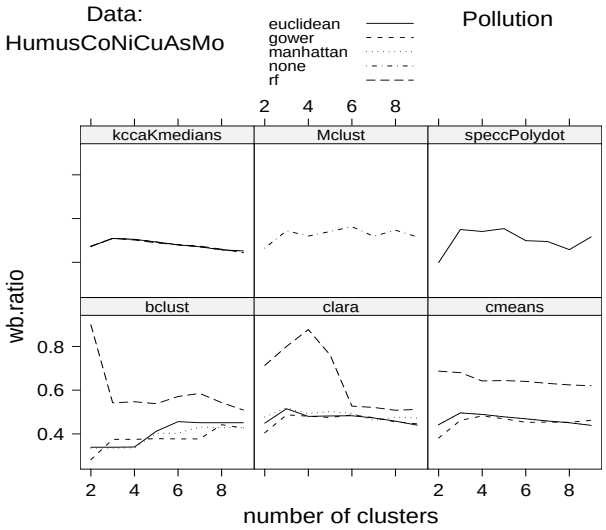
Comparing clustered data and clustered subsets of the data with Rand Index.
Distance measures which results in high Rand Indices should be chosen.

4. Clustering Method

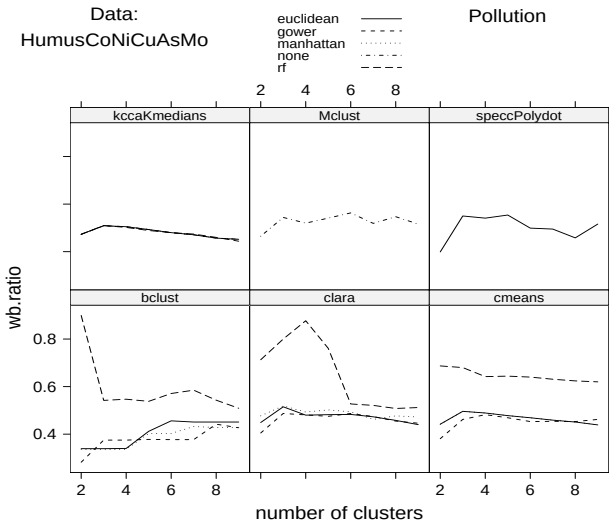


Comparing clustered data and clustered subsets of the data with Rand Index.
Algorithms which results in high Rand Indices may be chosen.

5. Number of Clusters

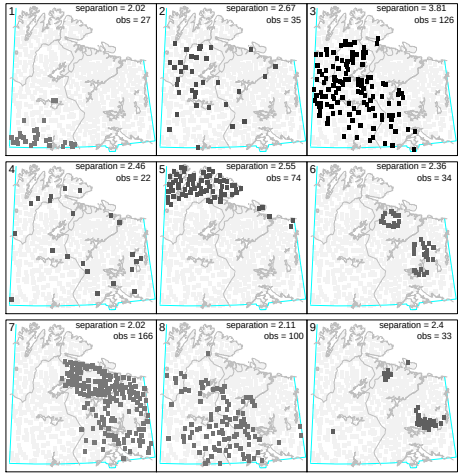


5. Number of Clusters

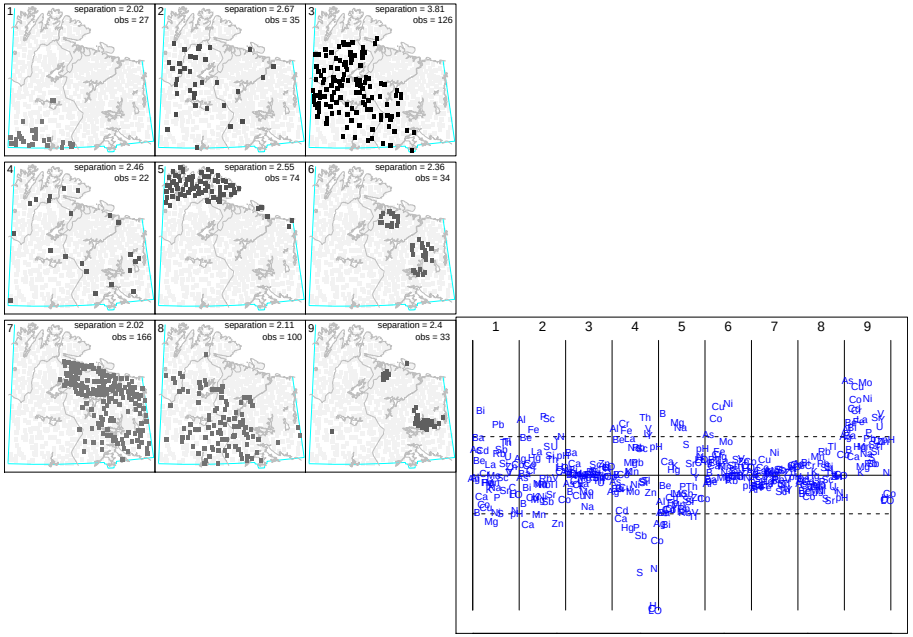


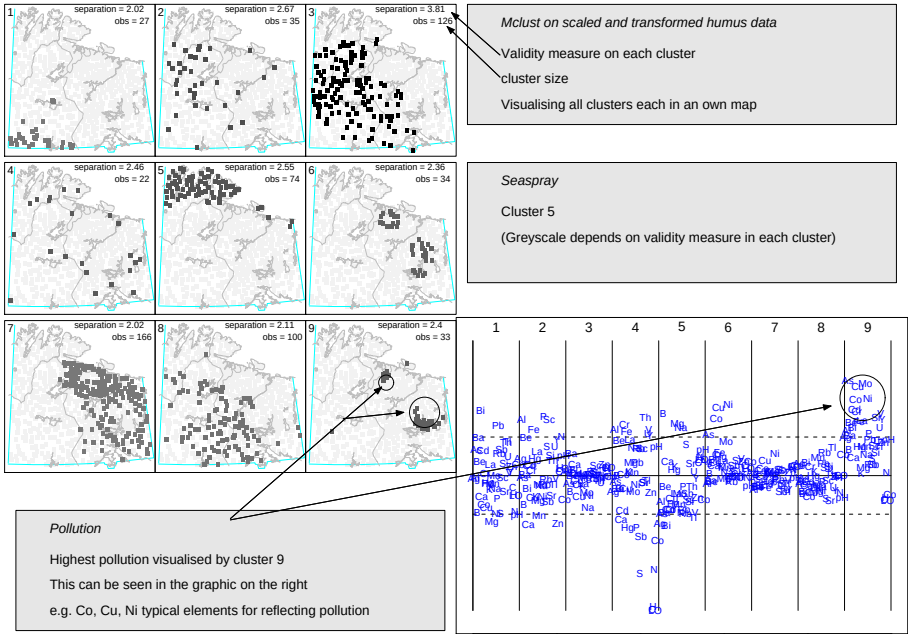
→

Example



→





Mclust on scaled and transformed humus data

Validity measure on each cluster

cluster size

Visualising all clusters each in an own map

Seaspray

Cluster 5

(Greyscale depends on validity measure in each cluster)

Pollution

Highest pollution visualised by cluster 9

This can be seen in the graphic on the right

e.g. Co, Cu, Ni typical elements for reflecting pollution

Conclusions

- Applying cluster analysis on real data results in highly non-stable results for many reasons
- The selection of variables and the selection of the optimal number of clusters on real data is a non-trivial task.
- Cluster analysis can be seen as explorative data analysis to get ideas about your data
- Interactive tools which allow for various methods are very helpful