Ward's method

It is based on the decomposition of the total deviance in deviance between groups and deviance within groups.

This method maximizes the deviance between the groups, minimizing the deviance within groups.

Creating, therefore, homogeneous internally groups, characterized by a low variability (variance).

This procedure tends to combine clusters with a small number of observations and it also biased toward the production of clusters with approximately the same number of observations.

Advantages and limitations

Single linkage: simple, little sensitive to outliers; the aggregation group is always the same (units join at each iteration to the first group).

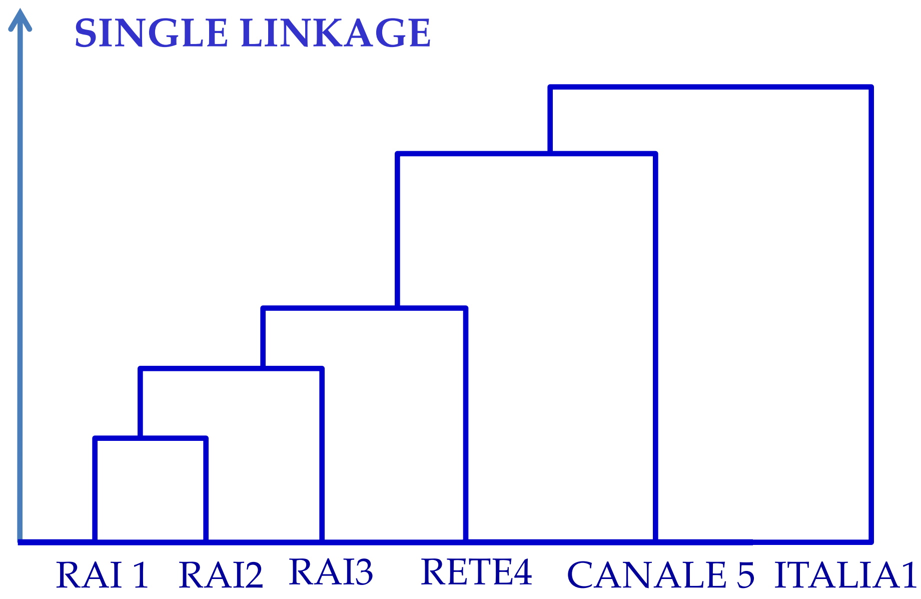
Complete linkage: simple; same dimension of the groups and sensitive to outliers.

McQuitty and average linkage: intermediate position between the first two.

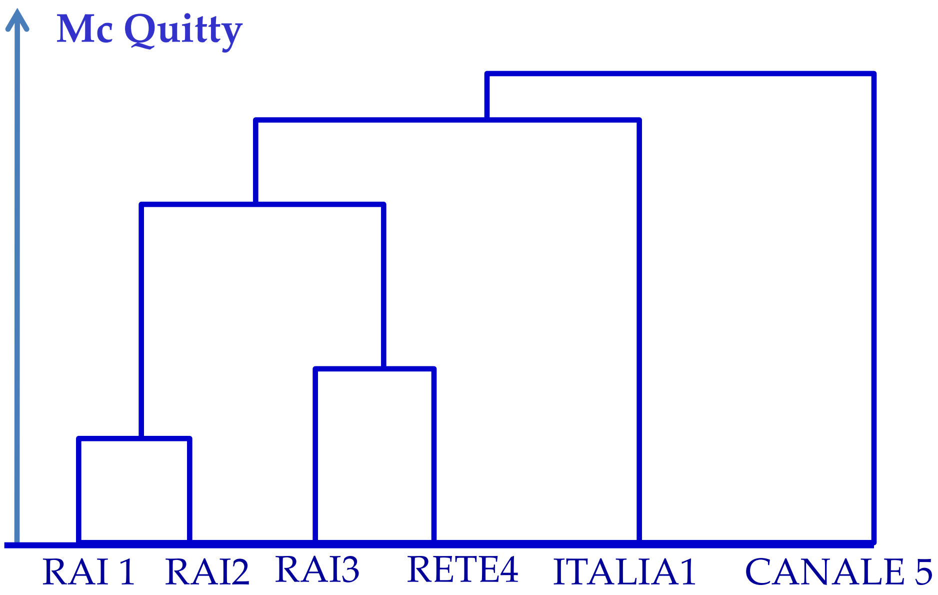
Centroid: not sensitive to outliers; only with Euclidean distances.

Ward: with all distance measures; same dimension of the groups and sensitive to outliers.

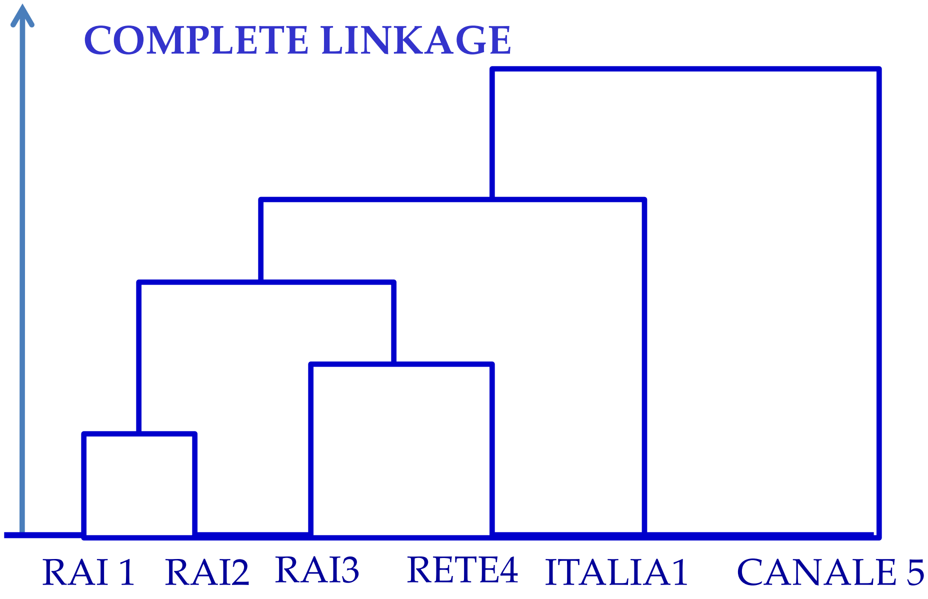
Dendrogram



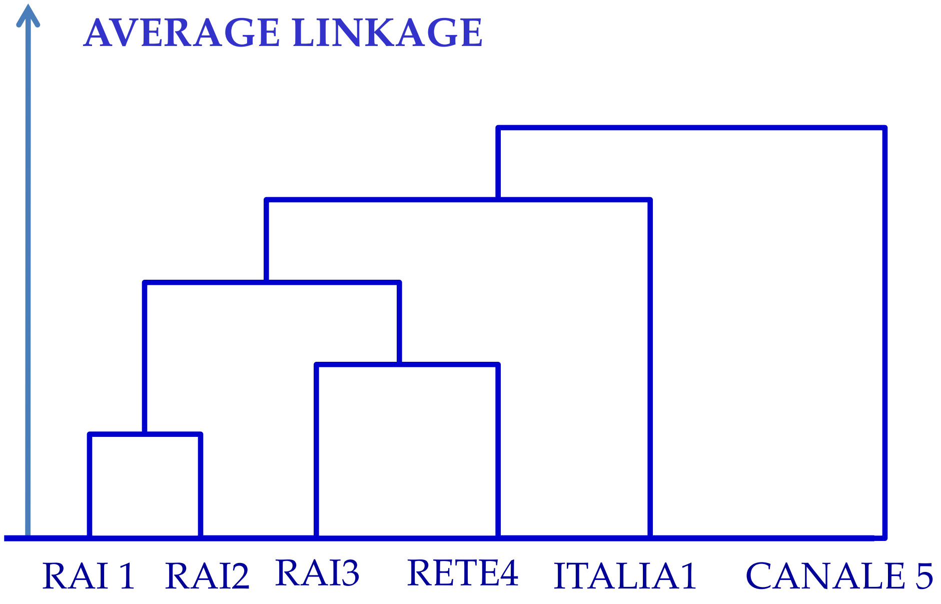
SINGLE LINKAGE



Mc Quitty



COMPLETE LINKAGE



AVERAGE LINKAGE

Cluster analysis

Methods:

- Hierarchical
  - agglomerative
  - divisive
- Non - Hierarchical

Divisive hierarchical methods

Characterized by a hierarchy in the group fusion, with a reverse path than agglomerative methods

1. 1 group
2. 2 groups
3. 3 groups…n-1 groups and so on until n-groups each formed by only 1 unit

Also in this case we have to select the stopping rule

Methods based on:

- centroids
- seed points

Centroids divisive method

Starting from a quantitative matrix X:

1. To choose among all possible subdivisions of n-units in 2 groups, it minimizes the sum of internal deviances of 2 groups

$$\boxed{SD = \sum_{g,i,h} ({}_g x_{ih} - {}_g x_i)^2}$$

$$\boxed{\begin{array}{l} {}_g x_{ih} = value\ of\ x_h\ for\ the\ i\text{-}th\ unit\ of\ the\ g\ group \\ g = 1,2;\ i = 1,\ldots n_g; h = 1,\ldots,p \end{array}}$$

2. At each step of the procedure the group with maximum internal deviance is subdivided in two groups and so on, starting from the step 1.

LIMITS:

Computational burden, same number of units for each group, it does not deal with mixed or qualitative variables

Seed points method

The method based on seed points starts from the distance matrix D, with variables of any type, also mixed (qualitative or quantitative):

1. 2 nodes are identified, more distant units, which are assigned to other units based on the minimum distance; They will thus form two groups,
2. Repeats step 1 on the 2 groups created, until to form n-groups.

Seed points method

D = DISTANCE MATRIX:

	RAI1	RAI2	RAI3	RETE4	CANALE5	ITALIA1
RAI1	0	864	933	1439	1863	2047
RAI2		0	1215	1591	2525	1886
RAI3			0	990	2370	2491
RETE4				0	2972	2062
CANALE5					0	3223
ITALIA1						0

Canale 5 and Italia 1 are the seed points

	RAI1	RAI3	CANALE5
RAI1	0	933	1863
RAI3		0	2370
CANALE 5			0

	RAI2	RETE4	ITALIA1
RAI2	0	1591	1886
RETE4		0	2062
ITALIA1			0

	RAI1	RAI3
RAI1	0	933
RAI3		0

CANALE 5

	RAI2	RETE4
RAI2	0	1591
RETE4		0

ITALIA 1

Cluster analysis

Methods:

- Hierarchical
  - agglomerative
  - divisive
- Non-Hierarchical

Non-hierarchical methods

These methods classify the n-units in a predetermined number of groups, without next agglomerations or resplitting.

They start from a quantitative data matrix X, generally standardized, and follow iterative algorithms that proceed by modifications of provisional groupings to obtain the configuration that optimizes an objective function.

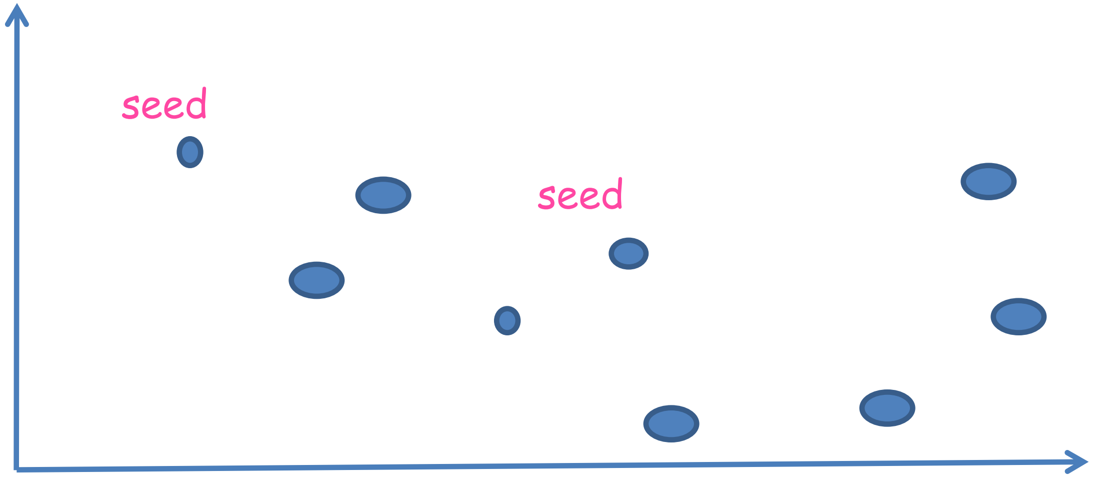
K-MEANS method:

1. Specify k-points (*seed*), in the space of p-variables: a seed for each group to be formed.
2. Each unit is assigned to the seed according to the minimum distance from the centroid of each provisional group.
3. Recalculate the centroids of provisional group.
4. reallocate the units into new groups based on centroids closer
5. Proceed iteratively until to get a stable configuration

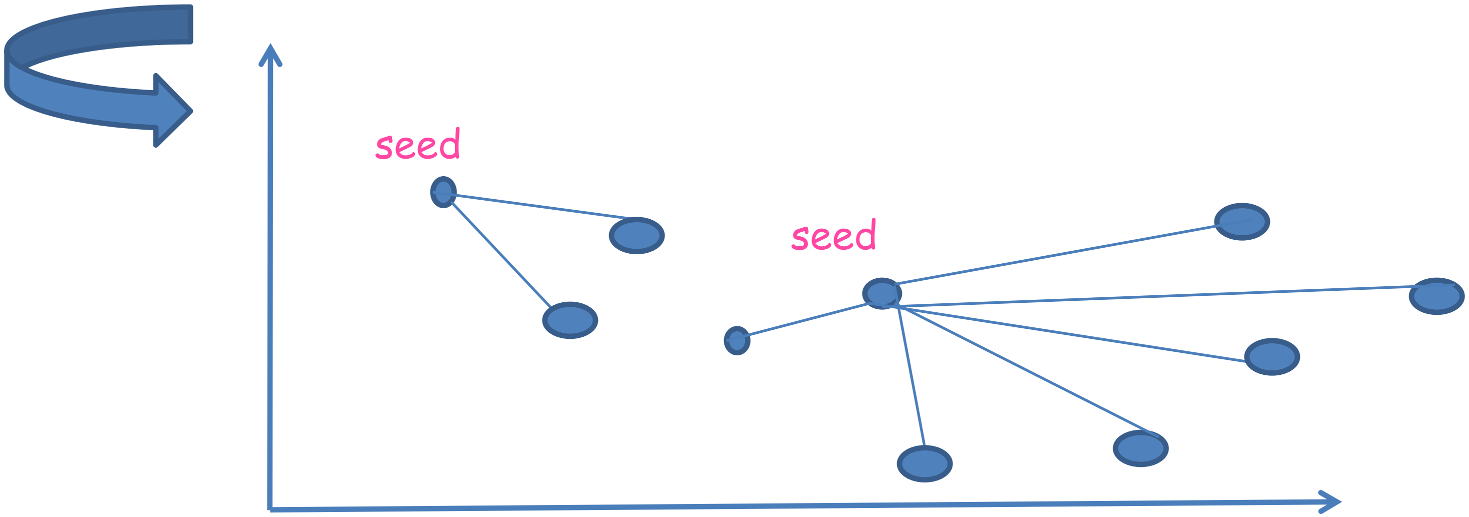
Limit: the arbitrary choice of the seeds.

- It is usual to repeat the analysis, trying to choose different starting points, and test the stability of the final configuration!

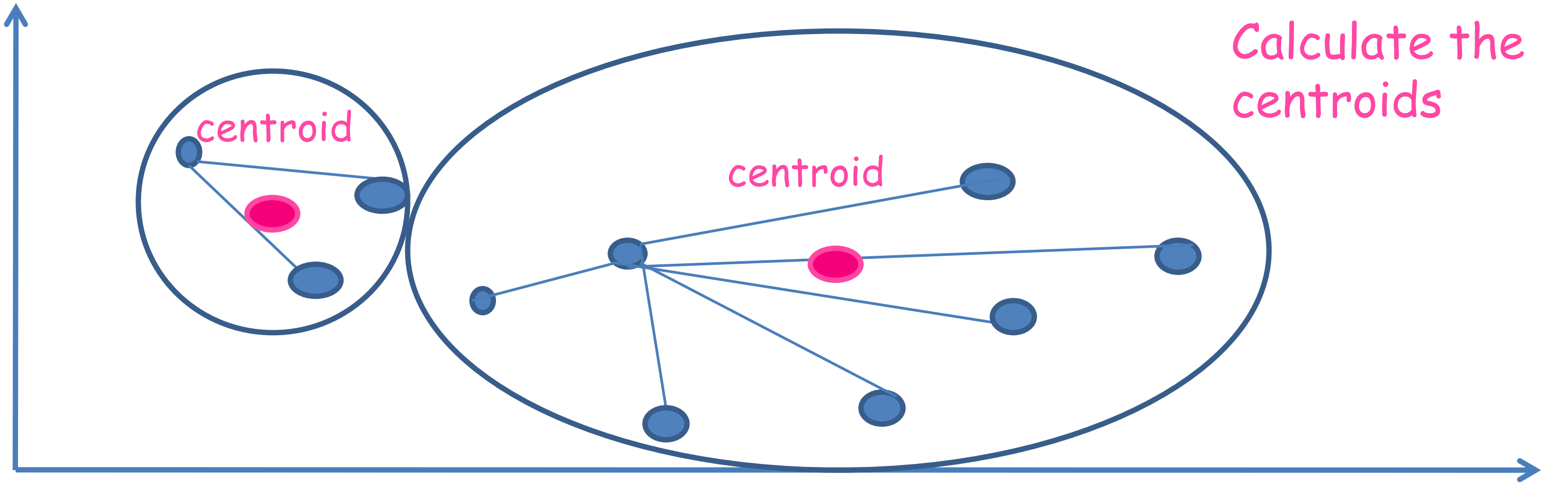
K-Means



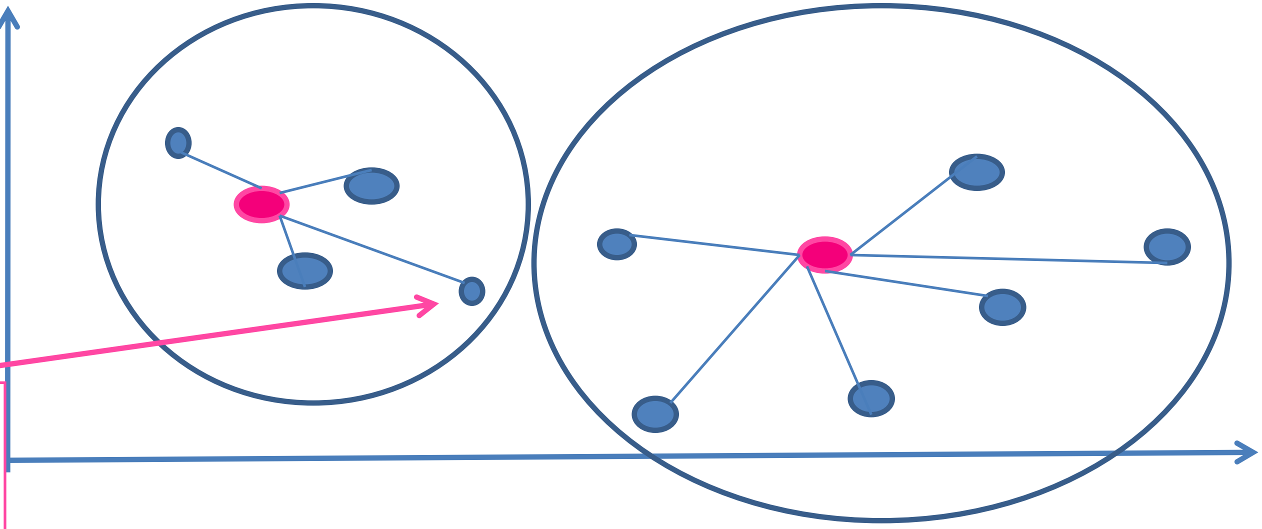
We have to form only two groups then we choose only two seeds



K-Means



Calculate the centroids

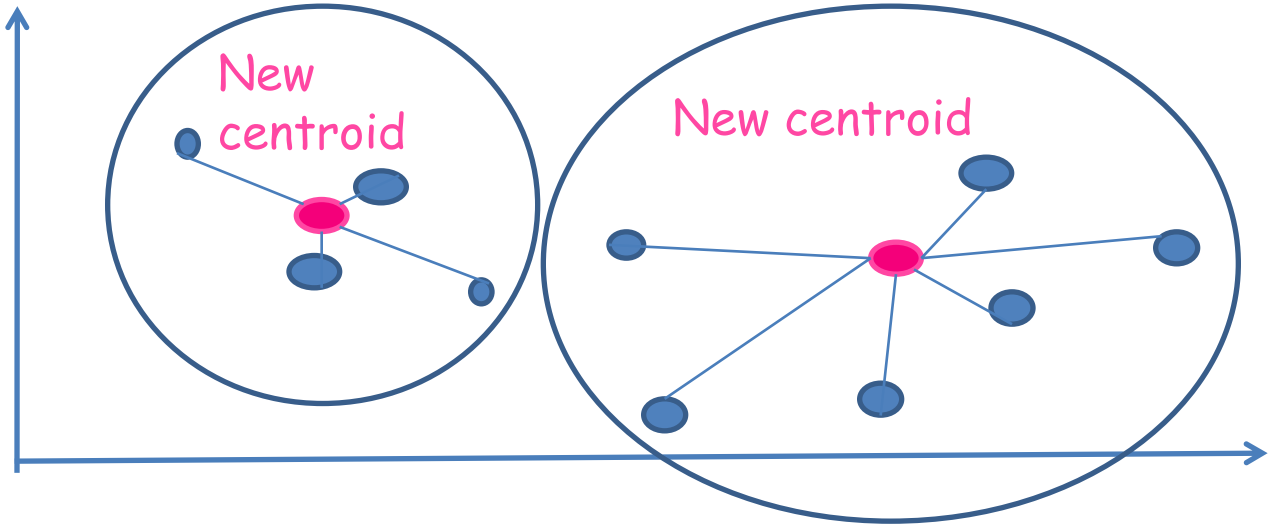


We assign the units on the minimum distance basis from the centroids building two groups

This unit in the last step was in the other group!!!

K-Means

Recalculate the centroids of the two new groups and reassign the units according to the minimum distance from the centroid and reform again the 2 groups



The difference with the agglomerative hierarchical method consists into the possiblity to separate group previously formed

Stable configuration: the units are collocated into the same groups of the last step

Hierarchical and non-hierarchical methods: a comparison

- The hierarchical methods are not very flexible: if two units were aggregated can not be divided in the next iteration.
- Generally applies a first hierarchical analysis, to find the optimal number of groups, to assign, then, as input in a subsequent non-hierarchical analysis which allows to obtain the final configuration.