Corso di Studio SM13 – CHIMICA

Introduzione alla chemiometria ed al disegno sperimentale (018CM)

17/11/2022 (lezione 4)

Pierluigi Barbieri

UNIVERSITÀ DEGLI STUDI DI TRIESTE

Dipartimento di Scienze Chimiche e Farmaceutiche

Gruppo di ricerca in Scienze Analitiche applicate alle InterazionI Uomo-Ambiente (ASHEI)

Analisi Esplorativa dei Dati

Vogliamo:

- Trovare strutture
- Trovare raggruppamenti
- Trovare dati anomali (outliers)

Metodi di Classificazione

- Ricerca di raggruppamenti (di campioni, molecole, etc.)= UNSUPERVISED classification
- I raggruppamenti sono noti = SUPERVISED classification
- Visualizzare i raggruppamenti
- Classificare
- Testare/validare

  la classificazione

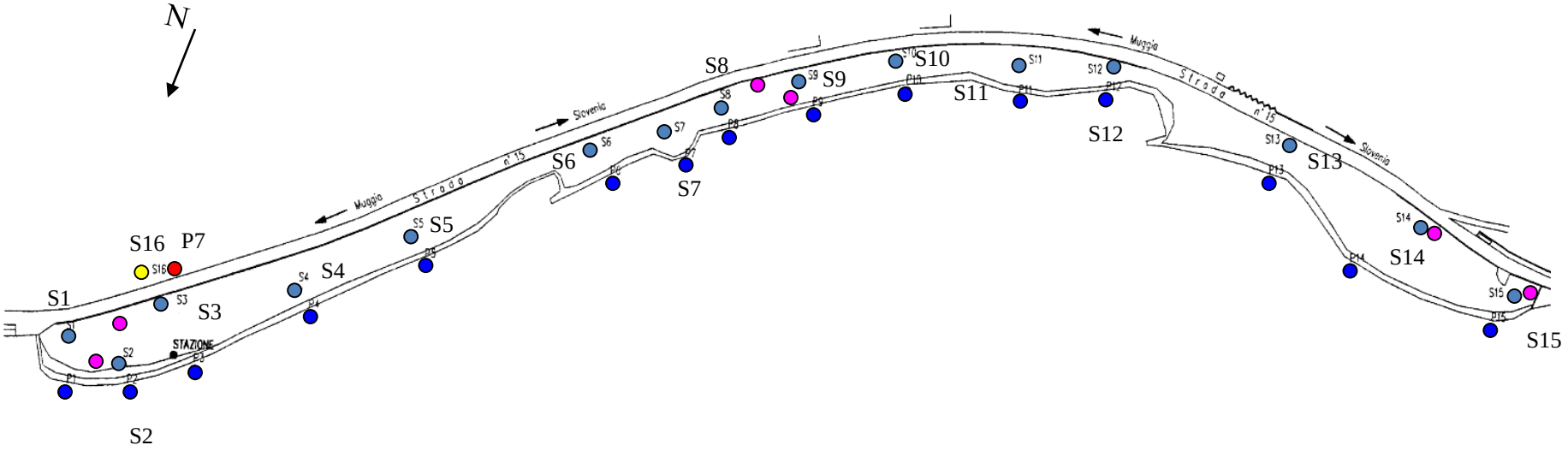
Metodi di raggruppamento (*clustering*)

https://www.researchgate.net/publication/261181646_Clustering_in_Analytical_Chemistry

ES: di applicazione ricerca di similitudine tra contaminazione e tipo suolo in diversi punti di campionamento per identificare tecnologia di bonifica/messa in sicurezza

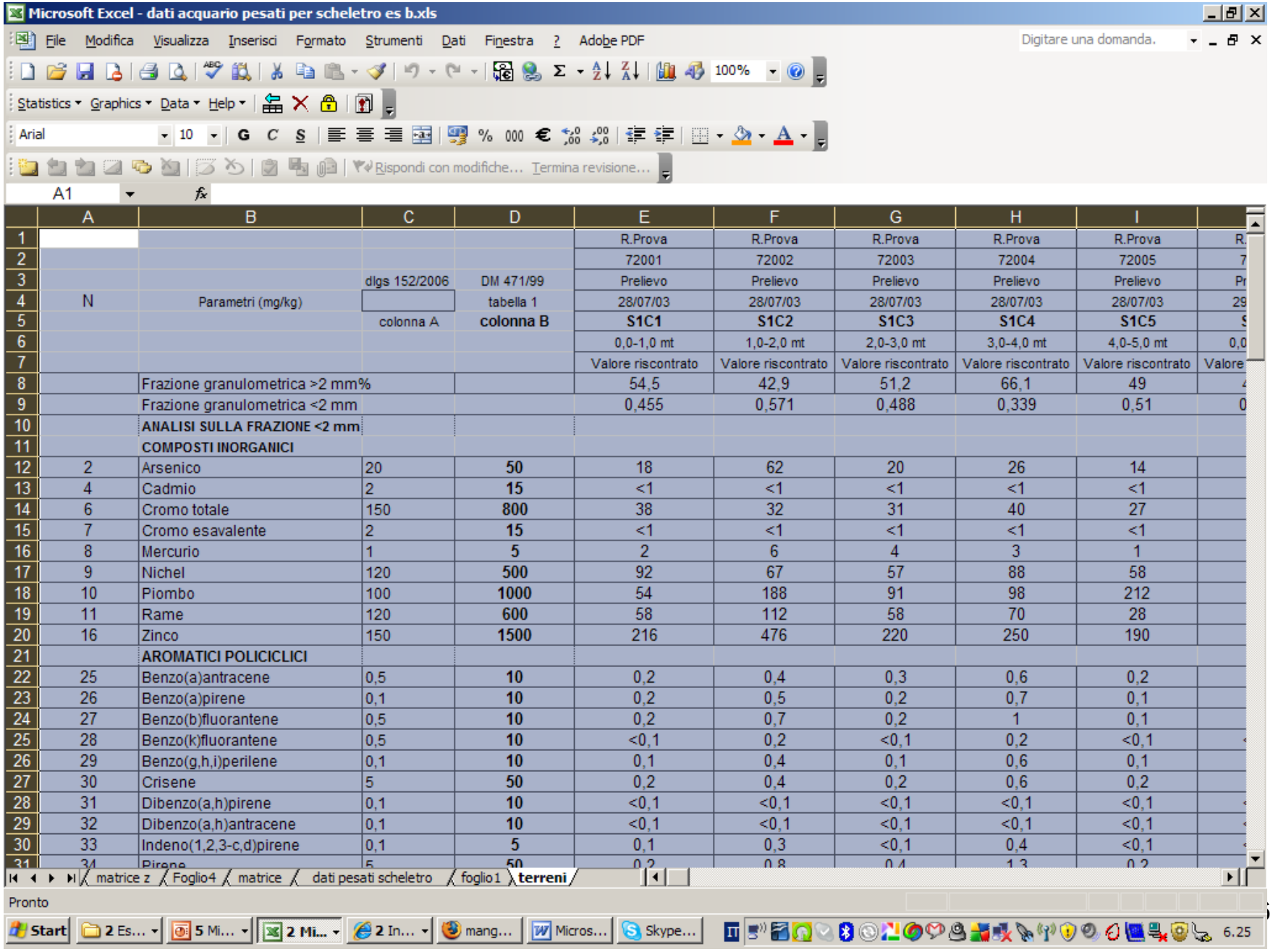
ricostruzione litostratigrafica dei terreni: S1÷S16

- prove di permeabilità
- installazione di un piezometro per individuare il fondo naturale e identificarlo come bianco: P7
- analisi di campioni di terreno
- analisi acqua di falda freatica, piezometri: P1 ÷ P7
- analisi di sedimento di fondale marino P1 ÷ P15



Scala 1:2500

The data



Microsoft Excel - dati acquario pesati per scheletro es b.xls

File Modifica Visualizza Inserisci Formato Strumenti Dati Finestra ? Adobe PDF

Digitare una domanda.

Statistics Graphics Data Help

Arial 10

Rispondi con modifiche... Termina revisione...

A1

	A	B	C	D	E	F	G	H	I
1					R.Prova	R.Prova	R.Prova	R.Prova	R.Prova
2					72001	72002	72003	72004	72005
3			dlgs 152/2006	DM 471/99	Prelievo	Prelievo	Prelievo	Prelievo	Prelievo
4	N	Parametri (mg/kg)		tabella 1	28/07/03	28/07/03	28/07/03	28/07/03	28/07/03
5			colonna A	colonna B	S1C1	S1C2	S1C3	S1C4	S1C5
6					0,0-1,0 mt	1,0-2,0 mt	2,0-3,0 mt	3,0-4,0 mt	4,0-5,0 mt
7					Valore riscontrato	Valore riscontrato	Valore riscontrato	Valore riscontrato	Valore riscontrato
8		Frazione granulometrica >2 mm%			54,5	42,9	51,2	66,1	49
9		Frazione granulometrica <2 mm			0,455	0,571	0,488	0,339	0,51
10		ANALISI SULLA FRAZIONE <2 mm							
11		COMPOSTI INORGANICI							
12	2	Arsenico	20	50	18	62	20	26	14
13	4	Cadmio	2	15	<1	<1	<1	<1	<1
14	6	Cromo totale	150	800	38	32	31	40	27
15	7	Cromo esavalente	2	15	<1	<1	<1	<1	<1
16	8	Mercurio	1	5	2	6	4	3	1
17	9	Nichel	120	500	92	67	57	88	58
18	10	Piombo	100	1000	54	188	91	98	212
19	11	Rame	120	600	58	112	58	70	28
20	16	Zinco	150	1500	216	476	220	250	190
21		AROMATICI POLICICLICI							
22	25	Benzo(a)antracene	0,5	10	0,2	0,4	0,3	0,6	0,2
23	26	Benzo(a)pirene	0,1	10	0,2	0,5	0,2	0,7	0,1
24	27	Benzo(b)fluorantene	0,5	10	0,2	0,7	0,2	1	0,1
25	28	Benzo(k)fluorantene	0,5	10	<0,1	0,2	<0,1	0,2	<0,1
26	29	Benzo(g,h,i)perilene	0,1	10	0,1	0,4	0,1	0,6	0,1
27	30	Crisene	5	50	0,2	0,4	0,2	0,6	0,2
28	31	Dibenzo(a,h)pirene	0,1	10	<0,1	<0,1	<0,1	<0,1	<0,1
29	32	Dibenzo(a,h)antracene	0,1	10	<0,1	<0,1	<0,1	<0,1	<0,1
30	33	Indeno(1,2,3-c,d)pirene	0,1	5	0,1	0,3	<0,1	0,4	<0,1
31	34	Pirene	5	50	0,2	0,8	0,4	1,3	0,2

matrice z | Foglio4 | matrice | dati pesati scheletro | foglio1 | terreni

Pronto

Start 2 Es... 5 Mi... 2 Mi... 2 In... mang... Micros... Skype... 6.25

Cluster analysis

Cluster analysis aims at grouping observations in *clusters*

Clusters should possibly be characterized by:

- High within homogeneity: observations in the same cluster should be *similar* (*not dissimilar*)
- High between heterogeneity: observations placed in different clusters should be *quite distinct* (*dissimilar*)

This means that we are interested in determining groups internally characterized by an high level of cohesion. Also, different clusters should describe different characteristics of the observations

Distanza e similarità

Campione/ osservazione	Variabile/ Proprietà $_1$	Variabile/ Proprietà $_2$	...	...	Variabile/ Proprietà $_m$
C_1					
C_2					
...					
...					
C_n					

Quanto si "assomigliano" queste due osservazioni?

Ovvero... quanto sono dissimili (distanti)?

DISTANZA = 1 / SIMILARITA'

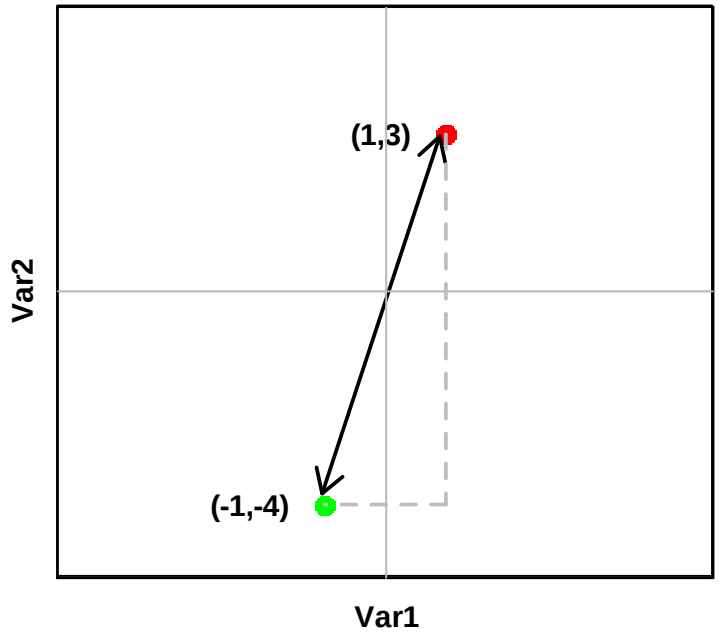
Distanza tra due osservazioni

Supponiamo di avere due osservazioni per cui sono state misurate due variabili:

Osservazione	Var1	Var2
C_1	1	3
C_2	-1	-4

$$\text{Distanza} = \sqrt{(\text{Var1}_{C1} - \text{Var1}_{C2})^2 + (\text{Var2}_{C1} - \text{Var2}_{C2})^2}$$

DISTANZA EUCLIDEA



$$\text{Distanza}_{C1\text{-}C2} = \sqrt{(1+1)^2 + (3+4)^2} = 7.3$$

Cluster analysis

A basic concept in cluster analysis is *dissimilarity* between observations and, also, between groups.

Let us first of all focus on observations.

In the context of cluster analysis, a measure of the dissimilarity between two cases, say the *i*-th and the *k*-th, satisfies the following:

$$d_{i,k} \geq 0 \qquad \text{for all } i,k$$

$$d_{i,i} = 0$$

$$d_{i,k} = d_{k,i}$$

WHAT IF $d_{ik} = 0$?

$d_{ik} = 0$ does not mean that two cases are identical. This only means that they are not dissimilar *with respect to the particular context under analysis*

Cluster analysis

Dissimilarities between observations are arranged in the so called dissimilarity matrix, a square ($n \times n$) matrix, where n is the number of observations.

Cases \ Cases				
	d_{11}	d_{12}	….	d_{1n}
	d_{21}	d_{22}	….	d_{2n}
	….	….	….	….
	d_{n1}	d_{n2}	….	d_{nn}

The (i,k)-th element of the matrix is the dissimilarity between the i-th and the k-th case. The matrix is symmetric, since we assumed that $d_{i,k} = d_{k,i}$

In some applications the dissimilarity matrix is obtained by taking into account measurements on a set of variables. Different measures of dissimilarities have been introduced in literature depending on the characteristics of the involved variables. Hence, different dissimilarity matrices can be obtained.

In other situations, the dissimilarity matrix may contain other kind of information, for example *judgements* about the dissimilarity between cases. In this case, the dissimilarity matrix is given.

Cluster Analysis for numerical variables

Dissimilarity measures for numerical variables

In the case when clusters have to be obtained on the basis of a vector of measurements on *p* variables (data matrix), the dissimilarity between two cases may be calculate by referring to the standard Euclidean distance or to the statistical distance

Euclidean distance

$$d_{i,k}^{E} = \sqrt{(x_{i1} - x_{k1})^2 + (x_{i2} - x_{k2})^2 + \ldots + (x_{ip} - x_{kp})^2}$$

Statistical distance

$$d_{i,k}^{S} = \sqrt{(z_{i1} - z_{k1})^2 + (z_{i2} - z_{k2})^2 + \ldots + (z_{ip} - z_{kp})^2}$$

Where z_{ij} is the standardized value corresponding to x_{ij}

Notice that the squared deviations are considered. As a consequences, extreme values on a given variable will have a great influence on the resulting dissimilarity. Moreover, extreme observations will be very dissimilar from the others, and hence regular observations will possibly be clustered together independently on their differences (clusters of regular vs clusters of extreme obs)

Cluster Analysis for numerical variables

Dissimilarity measures for numerical variables

An alternative criterion based on absolute rather than squared deviations is the

Manhattan (or City block) distance

$$d_{i,k}^{CB} = | x_{i1} - x_{k1} | + | x_{i2} - x_{k2} | + ... + | x_{ip} - x_{kp} |$$

Also in this case a transformation may be applied similar to standardization.

The absolute deviation relative to the *j*-th variable may be divided by:

The range, R_j = (highest value– lowest value) for the *j*-th variable

The MAD, the median of the absolute deviations from the median.

The second criterion is less sensible to outliers (outliers may strongly influence the range) and it is similar to the standardization (dividing by a ‘standard’ deviation, in this situation a synthesis (median) of the deviance from the median)

In statistics, **Mahalanobis distance** is a distance measure introduced by P. C. Mahalanobis in 1936.[1] It is based on correlations between variables by which different patterns can be identified and analyzed. It is a useful way of determining *similarity* of an unknown sample set to a known one. It differs from Euclidean distance in that it takes into account the correlations of the data set and is scale-invariant, i.e. not dependent on the scale of measurements.

Contents [hide]

Definition [edit]

Formally, the Mahalanobis distance from a group of values with mean $\mu = (\mu_1, \mu_2, \mu_3, \ldots, \mu_N)^T$ and covariance matrix S for a multivariate vector $x = (x_1, x_2, x_3, \ldots, x_N)^T$ is defined as:

$$D_M(x) = \sqrt{(x-\mu)^T S^{-1} (x-\mu)}.$$[2]

Mahalanobis distance (or "generalized squared interpoint distance" for its squared value[3]) can also be defined as dissimilarity measure between two random vectors $\vec{x}$ and $\vec{y}$ of the same distribution with the covariance matrix S :

$$d(\vec{x}, \vec{y}) = \sqrt{(\vec{x}-\vec{y})^T S^{-1} (\vec{x}-\vec{y})}.$$

If the covariance matrix is the identity matrix, the Mahalanobis distance reduces to the Euclidean distance. If the covariance matrix is diagonal, then the resulting distance measure is called the *normalized Euclidean distance*:

$$d(\vec{x}, \vec{y}) = \sqrt{\sum_{i=1}^{N} \frac{(x_i - y_i)^2}{\sigma_i^2}},$$

where σ_i is the standard deviation of the x_i over the sample set.

Cluster analysis

Example (synthetic data). Dissimilarity matrix

label	Dist 1	Dist 2	Dist 3	Dist 4	Dist 5	Dist 6	Dist 7	Dist 8	Dist 9	Dist 10	Dist 11	Dist 12	Dist 13	Dist 14	Dist 15	Dist 16
obs1	0.00	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
obs2	0.85	0.00	.	.	.	.	.	.	.	.	.	.	.	.	.	.
obs3	1.53	0.70	0.00	.	.	.	.	.	.	.	.	.	.	.	.	.
obs4	2.12	1.93	1.80	0.00	.	.	.	.	.	.	.	.	.	.	.	.
obs5	5.22	4.57	3.94	3.50	0.00	.	.	.	.	.	.	.	.	.	.	.
obs6	2.60	1.81	1.12	1.94	2.88	0.00	.	.	.	.	.	.	.	.	.	.
obs7	4.10	3.31	2.61	3.03	1.79	1.50	0.00	.	.	.	.	.	.	.	.	.
obs8	4.10	3.30	2.60	3.08	1.88	1.50	0.10	0.00	.	.	.	.	.	.	.	.
obs9	4.84	4.01	3.31	3.91	2.11	2.26	0.89	0.83	0.00	.	.	.	.	.	.	.
obs10	6.07	5.31	4.62	4.66	1.45	3.50	2.03	2.06	1.67	0.00	.	.	.	.	.	.
obs11	3.23	2.59	2.00	1.71	1.99	1.12	1.38	1.45	2.27	2.98	0.00	.	.	.	.	.
obs12	4.93	5.05	5.00	3.20	4.74	4.92	5.42	5.50	6.24	6.19	4.12	0.00	.	.	.	.
obs13	5.42	5.46	5.33	3.55	4.55	5.12	5.43	5.52	6.22	5.99	4.22	0.67	0.00	.	.	.
obs14	6.14	5.95	5.65	4.05	3.64	5.12	4.97	5.07	5.61	4.99	4.04	2.21	1.60	0.00	.	.
obs15	5.79	5.19	4.58	3.95	0.72	3.56	2.51	2.60	2.77	1.73	2.59	4.69	4.39	3.28	0.00	.
obs16	4.96	4.72	4.39	2.84	2.78	3.86	3.83	3.93	4.55	4.22	2.81	2.06	1.77	1.26	2.64	0

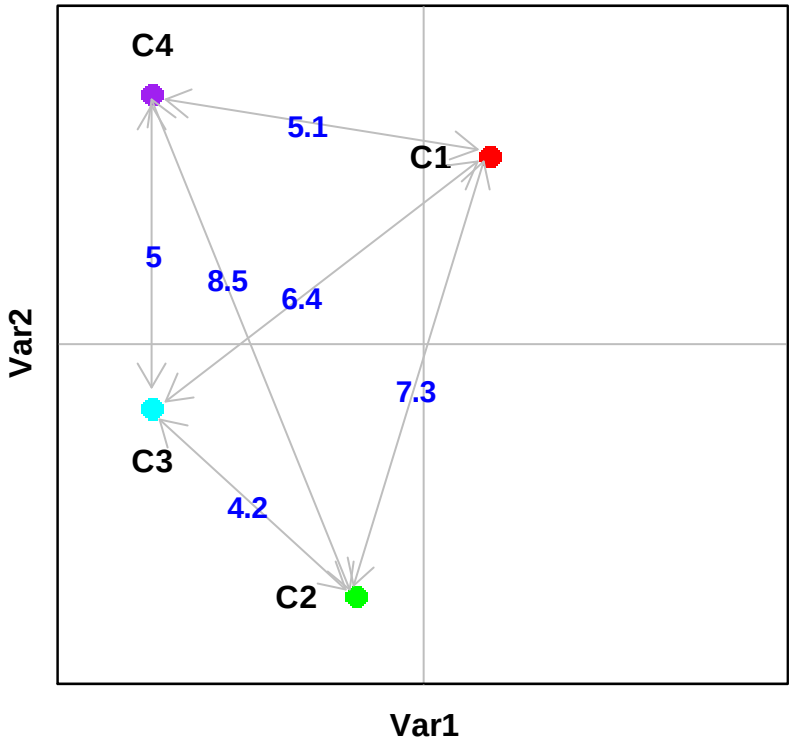
How should we choose groups? We can individuate some close pairs (for example, obs 7 and obs 8 are closest). But how many groups should we consider? How can we properly assign each observation to a given group?

Distanza tra più osservazioni: tabella delle distanze

Osservazione	Var1	Var2
C_1	1	3
C_2	-1	-4
C_3	-4	-1
C_4	-4	4

In R: **dist(nomematrice)**

```
> originale
   Var1 Var2
C1    1    3
C2   -1   -4
C3   -4   -1
C4   -4    4
>
> Dist<-round(dist(originale),digit=1)
>
> Dist
    C1  C2  C3
C2 7.3
C3 6.4 4.2
C4 5.1 8.5 5.0
>
```



Normalizzazione

Il risultato del calcolo della distanza ha significato se le variabili hanno unità di misura diverse?

Esempio:

Osservazione	Altezza (m)	Peso (kg)
P_1	1.65	81
P_2	1.95	80
P_3	1.60	70
P_4	1.85	80
P_5	1.95	95
P_6	1.55	60
P_7	1.80	80

```
> round(dist(AP))
   P1 P2 P3 P4 P5 P6
P2  1
P3 11 10
P4  1  0 10
P5 14 15 25 15
P6 21 20 10 20 35
P7  1  0 10  0 15 20
>
```

```
> APnorm<-scale(AP)
>
> round(dist(APnorm))
   P1 P2 P3 P4 P5 P6
P2  2
P3  1  2
P4  1  1  2
P5  2  1  3  2
P6  2  3  1  3
P7  1  1  2  0  2  2
```

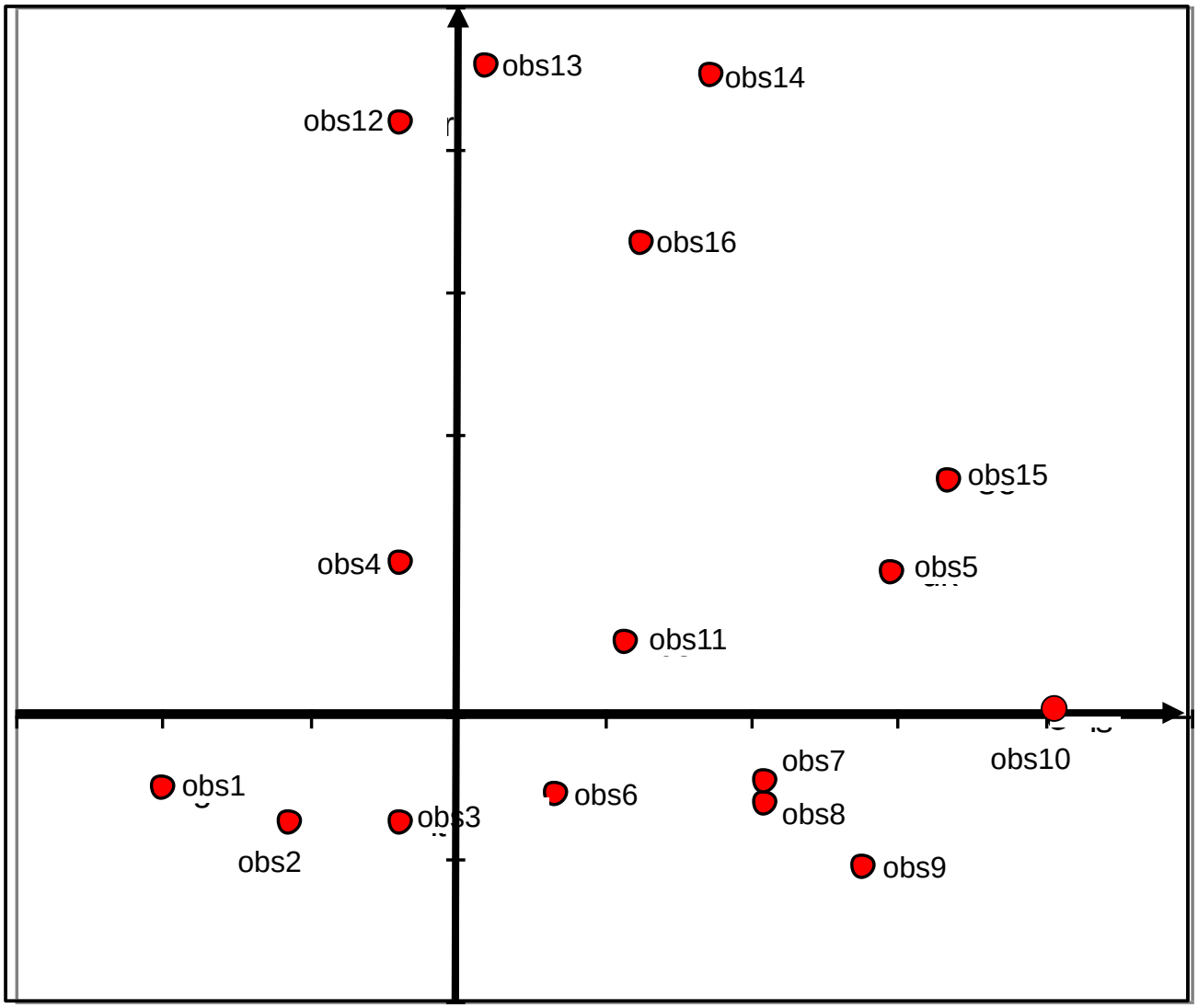
Quindi bisogna normalizzare i dati! (almeno usando la media per variabile)

Normalizzazione Z-score:

$$Z_{Var(i,j)} = \frac{Var_{(i,j)} - \overline{Var_j}}{\sigma_{Varj}}$$

Cluster analysis

Example (synthetic data). Simple example, 2 dimensions – graphical analysis



2 groups are clearly identifiable

But maybe also 3 groups may be considered. Which cluster should obs11 and obs6 be assigned to?

Types of clustering

Data clustering algorithms can be hierarchical. Hierarchical algorithms find successive clusters using previously established clusters. Hierarchical algorithms can be agglomerative ("bottom-up") or divisive ("top-down"). Agglomerative algorithms begin with each element as a separate cluster and merge them into successively larger clusters. Divisive algorithms begin with the whole set and proceed to divide it into successively smaller clusters.

Partitional algorithms typically determine all clusters at once, but can also be used as divisive algorithms in the hierarchical clustering.

Density-based clustering algorithms are devised to discover arbitrary-shaped clusters. In this approach, a cluster is regarded as a region in which the density of data objects exceeds a threshold. DBSCAN and OPTICS are two typical algorithms of this kind.

Two-way clustering, *co-clustering* or biclustering are clustering methods where not only the objects are clustered but also the features of the objects, i.e., if the data is represented in a data matrix, the rows and columns are clustered simultaneously.

Another important distinction is whether the clustering uses symmetric or asymmetric distances. A property of Euclidean space is that distances are symmetric (the distance from object *A* to *B* is the same as the distance from *B* to *A*). In other applications (e.g., sequence-alignment methods, see Prinzie & Van den Poel (2006)), this is not the case.

Many clustering algorithms require specification of the number of clusters to produce in the input data set, prior to execution of the algorithm. Barring knowledge of the proper value beforehand, the appropriate value must be determined, a problem for which a number of techniques have been developed.

Cluster analysis: methods

Hierarchical (agglomerative) algorithms

Sequential procedures.

At the first step, each observation constitutes a cluster. At each step, the two closest clusters are joined to form a new cluster. Thus, the groups at each step are nested with respect to the groups obtained at the previous step.

Once an object has been assigned to a group it is never *removed* from the group later on in the clustering process.

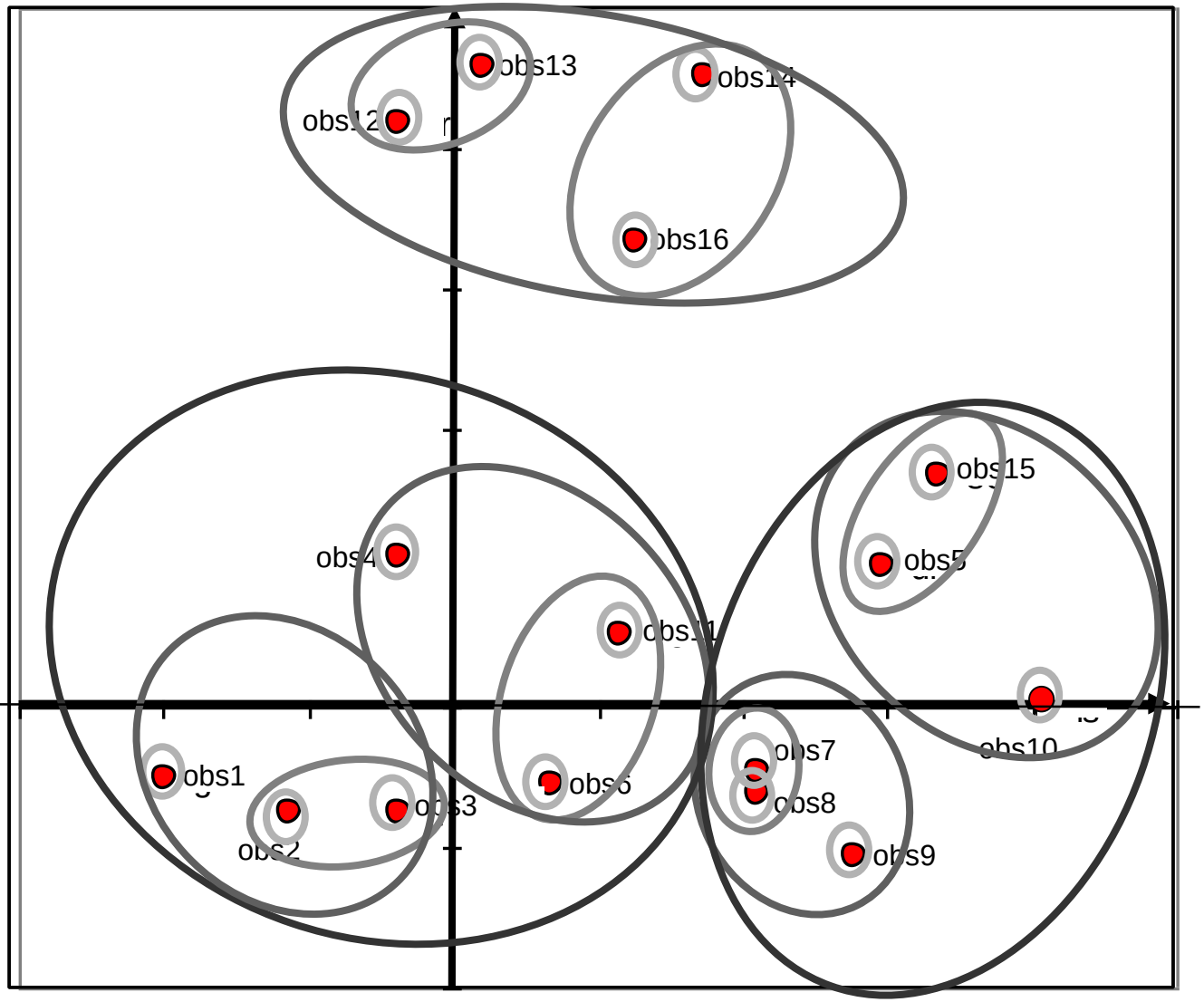
The hierarchical method produce a complete sequence of cluster solutions beginning with *n* clusters and ending with one clusters containing all the *n* observations.

In some application the set of nested clusters is the required solution whereas in other applications only one of the cluster solutions is selected as the solution, i.e., the proper number of clusters has to be selected.

Cluster analysis: **hierarchical algorithms**

Initial solution: *n clusters* (one for each observation)

At each step: the two closest (lowest dissimilarity) clusters are joined to form a new cluster

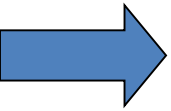


Hierarchical agglomerative algorithms

At each step, we should join the two closest clusters.

Our starting point is the dissimilarity matrix. It is almost easy to determine which are the two closest observations.

Nevertheless, now a problem arises: how do we calculate the dissimilarity between one observation and one cluster or between two clusters?



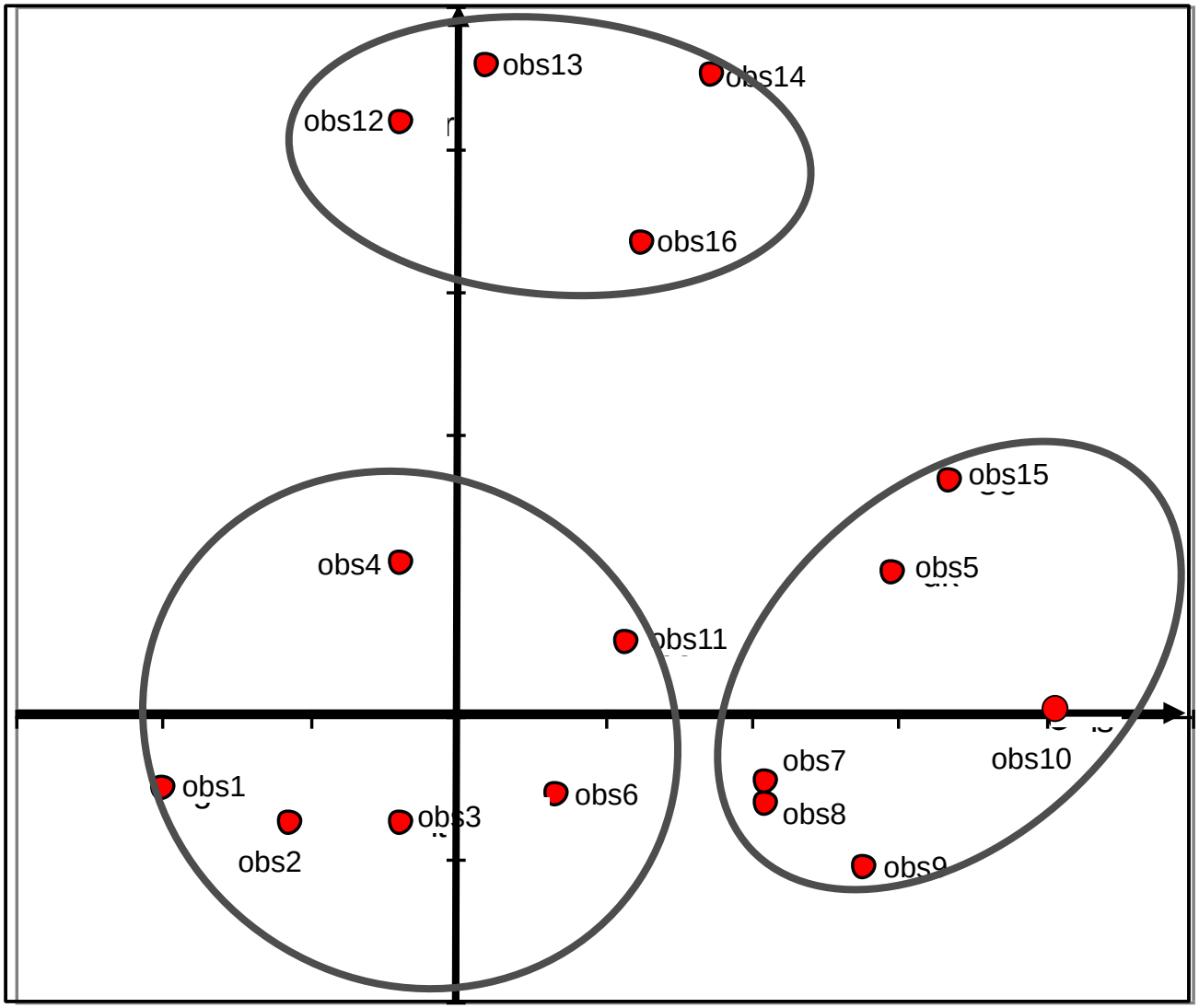
Definition of criteria to measure the

Dissimilarity between groups of observations (clusters)

Cluster analysis: hierarchical algorithms

We limit attention to two approaches to measure dissimilarity between clusters

1. Criteria based on the dissimilarity between two properly chosen observations
2. Criteria based on syntheses of the dissimilarities or on dissimilarities between syntheses.



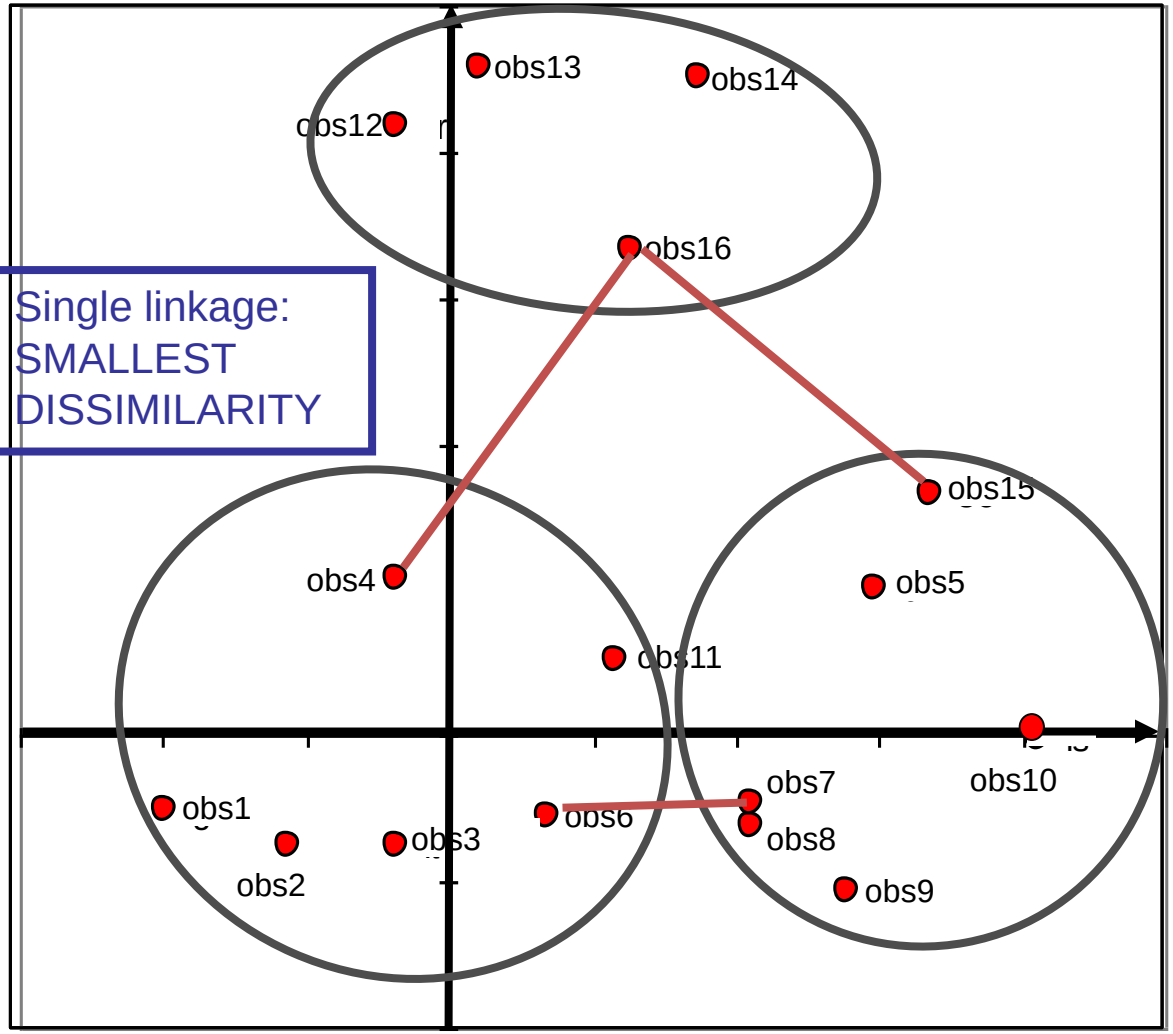
For the sake of clarity, we illustrate the proposals by referring to a simplified 2-dimensional plot (synthetic data) may be applied also when a dissimilarity matrix is available (regardless of how it was obtained).

We consider a 3-clusters partition and show how to measure the dissimilarity between 2 clusters.

Cluster analysis: hierarchical algorithms – dissimilarity/clusters

Single linkage: the dissimilarity between two clusters is measured by the smallest possible dissimilarity between cases in the two clusters (dissimilarity between the two closest cases)

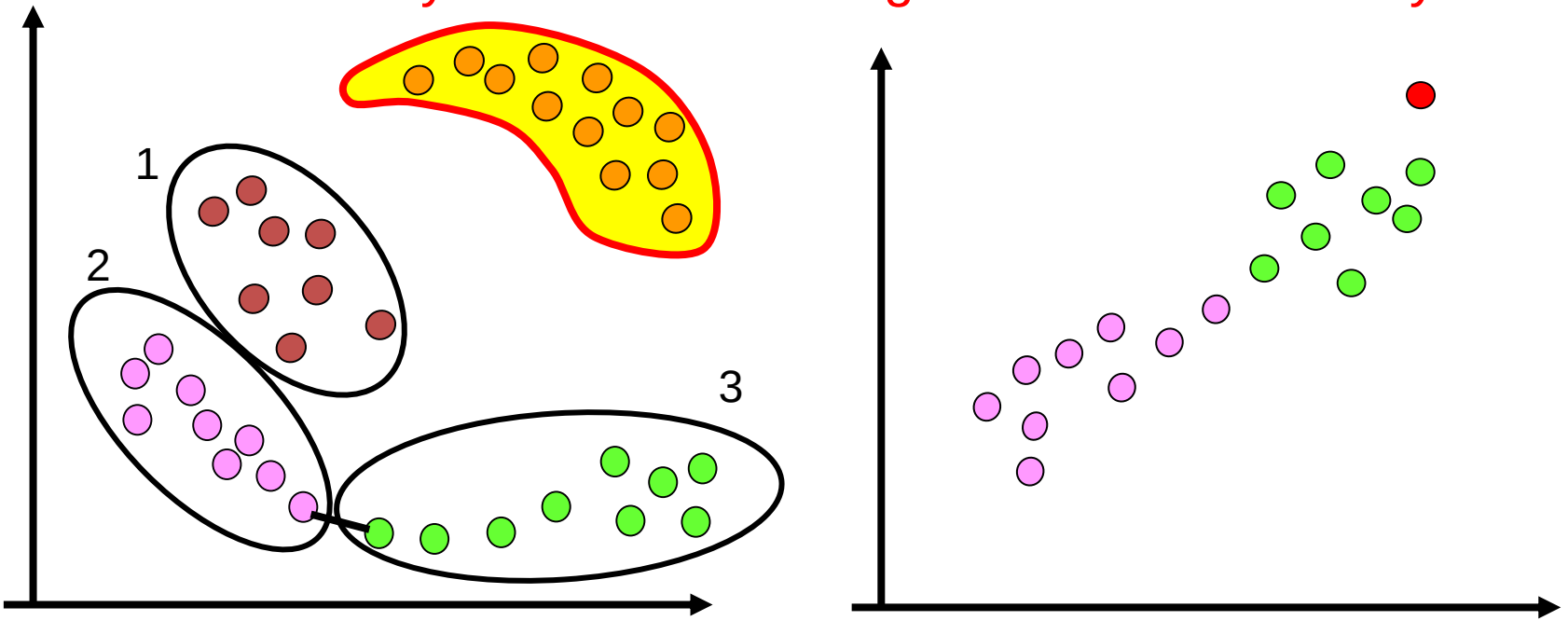
The two clusters with minimum single linkage are joined



Single linkage: which clusters should be joined?

The dissimilarity between two clusters is based on one of the possible pairs of observations

Cluster analysis: hierarchical algorithms – dissimilarity/clusters



Single linkage: It is a flexible method and it can individuate also clusters with particular shapes (elongated, elliptical)

However, in cases when clusters are not well separated this method may lead to unsatisfactory solutions due to the so called chaining effect.

Consider the three clusters 1-3 in the left panel. Clusters 1 and 2 are (“globally”) closer.

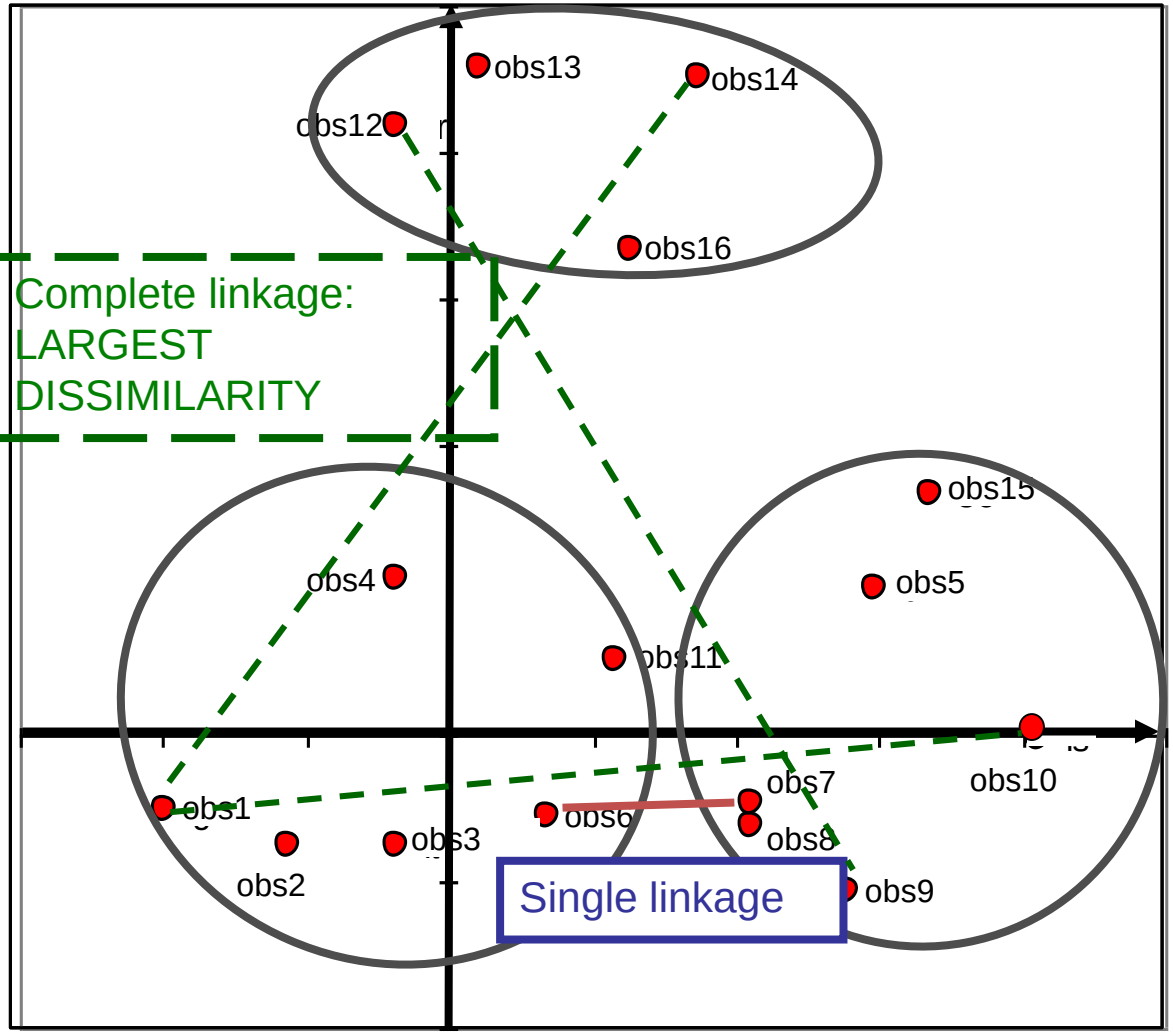
Nevertheless, due to the presence of two very close cases in clusters 2 and 3, they will be joined instead.

Another example is in the right panel. This last example evidences that this method may be useful in outliers detection.

Cluster analysis: hierarchical algorithms – dissimilarity/clusters

Complete linkage: the dissimilarity between two clusters is measured by the smallest possible dissimilarity between cases in the two clusters (dissimilarity between the two furthest cases)

The two clusters with minimum complete linkage are joined



Complete linkage: which clusters should be joined?

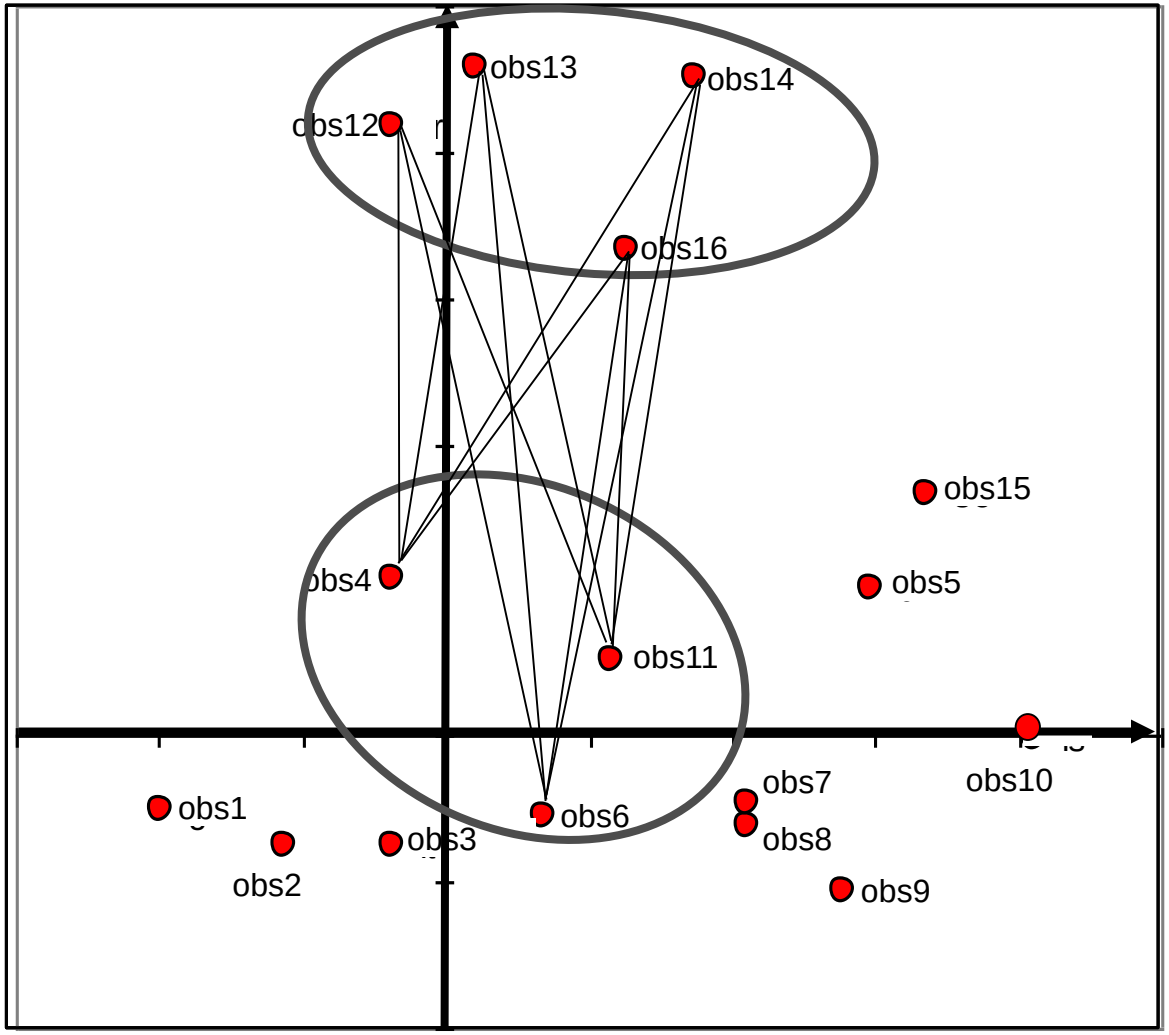
The dissimilarity between two clusters is based on one of the possible pairs of observations

Usually clusters with similar diameters are obtained

Depending on the criterion chosen to measure dissimilarity between clusters, different clusters are joined

Cluster analysis: hierarchical algorithms – dissimilarity/clusters

Average linkage: The dissimilarity between two clusters is given by the average the dissimilarities between all the possible pairs of cases

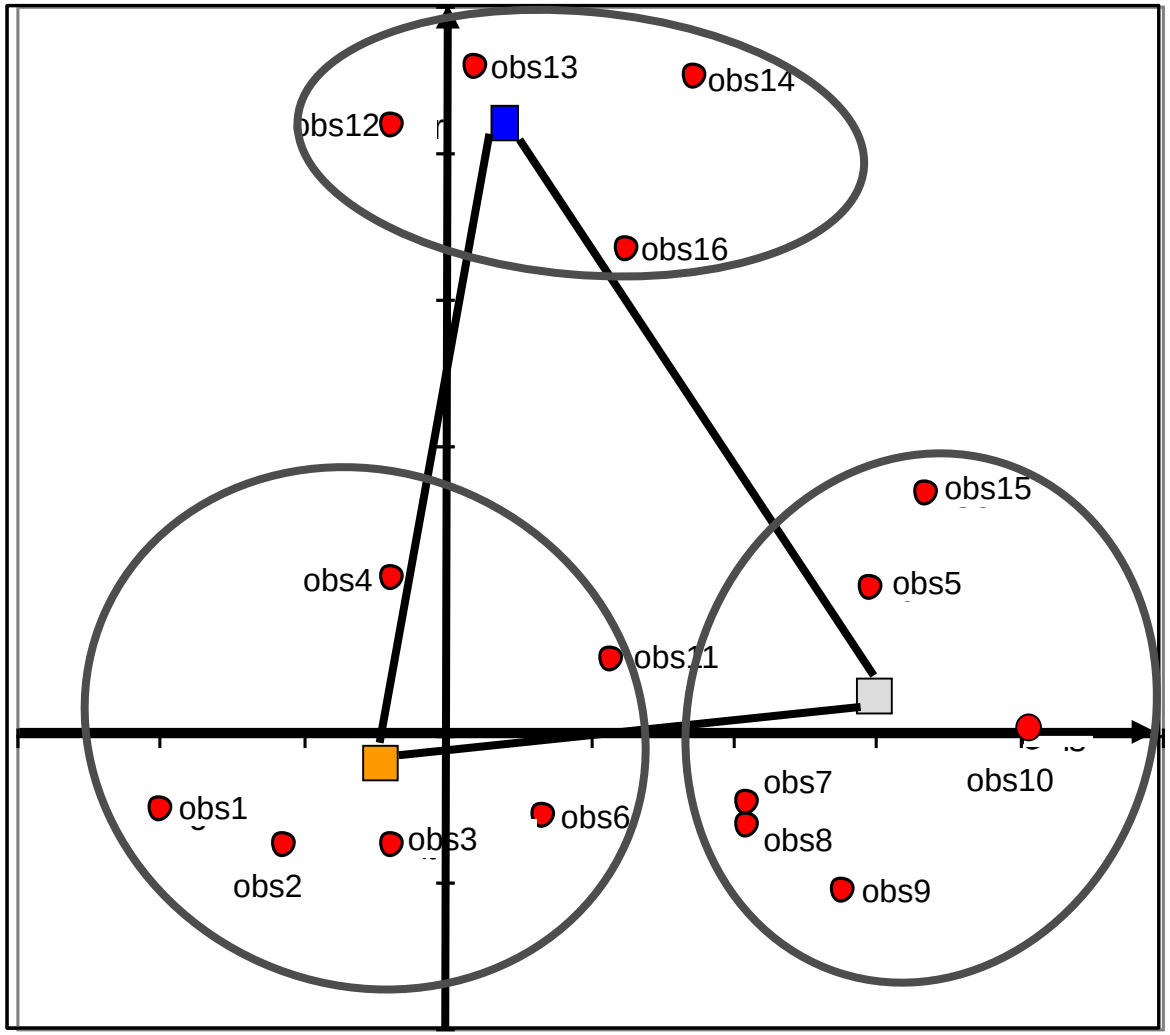


The dissimilarity is based upon a synthesis of all the dissimilarities

Usually clusters with similar variances are obtained

Cluster analysis: hierarchical algorithms – dissimilarity/clusters

Dissimilarity between centroids: The dissimilarity between two clusters is given by the dissimilarities between the centroids (Important: this quantity may also be evaluated when only the dissimilarity matrix is available)



The dissimilarity is based upon a synthesis of all the dissimilarities

Cluster analysis: hierarchical algorithms – dissimilarity/clusters

Ward's method: Let us focus only on two of the three clusters considered before, and let us consider the case when a data matrix is available (even if the procedure can be extended to the case when we only have a dissimilarity matrix). This method is based upon the concept of *within sum of squares.*

Within sum of squares for a cluster *c*

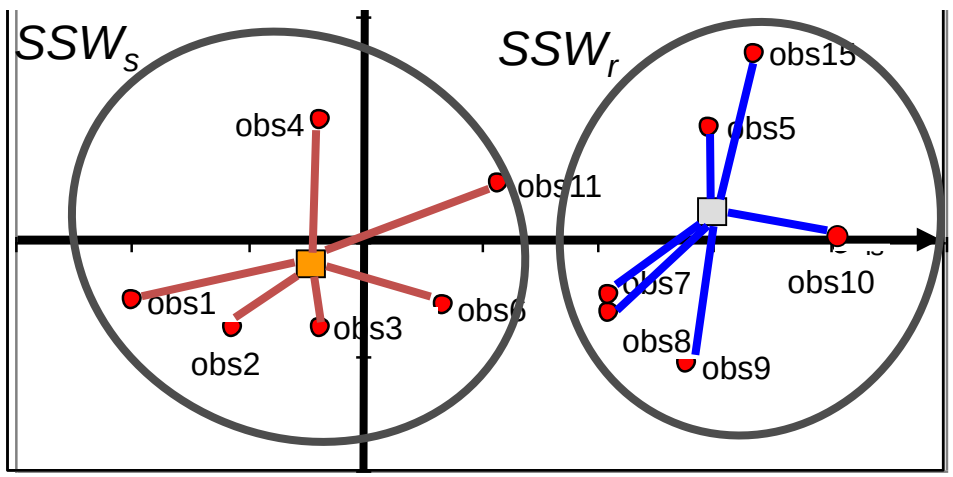
$$SSW_c = \sum_{i=1}^{n_c} \sum_{j=1}^{p} (x_{ijc} - \bar{x}_{jc})^2$$

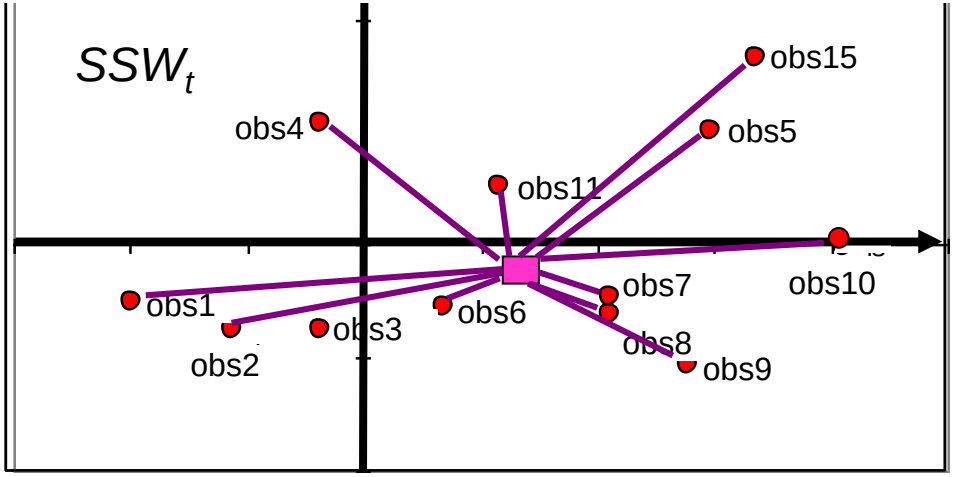
Suppose now that the two clusters *r* and *s* are joined to form cluster *t.*

It will be $SSW_t > SSW_r + SSW_s$ (the two original centroids will explain better cases within clusters). The increase consequent to the joining of *r* and *s* will be quite small if the two clusters are very close, and high if they are very different.

The quantity $SSW_t - (SSW_r + SSW_s)$ is called between sum of squares (SS).

Ward's method: the two clusters with the smallest Between SS are joined.





Cluster analysis: hierarchical algorithms

Given a dissimilarity matrix, based on a certain measure of the dissimilarity between cases, there are different methods to measure the dissimilarity between *clusters.* These criteria often lead to different partitions.

Single Linkage Cluster Analysis

NCL	Clusters Joined		FREQ	Min Dist
15	obs7	obs8	2	0.1
14	obs12	obs13	2	0.6708
13	obs2	obs3	2	0.7
12	obs5	obs15	2	0.7211
11	CL15	obs9	3	0.8322
10	obs1	CL13	3	0.8544
9	CL10	obs6	4	1.118
8	CL9	obs11	5	1.118
7	obs14	obs16	2	1.2649
6	CL8	CL11	8	1.3793
5	CL12	obs10	3	1.45
4	CL14	CL7	4	1.6031
3	CL6	CL5	11	1.6651
2	CL3	obs4	12	1.7088
1	CL2	CL4	16	2.6401

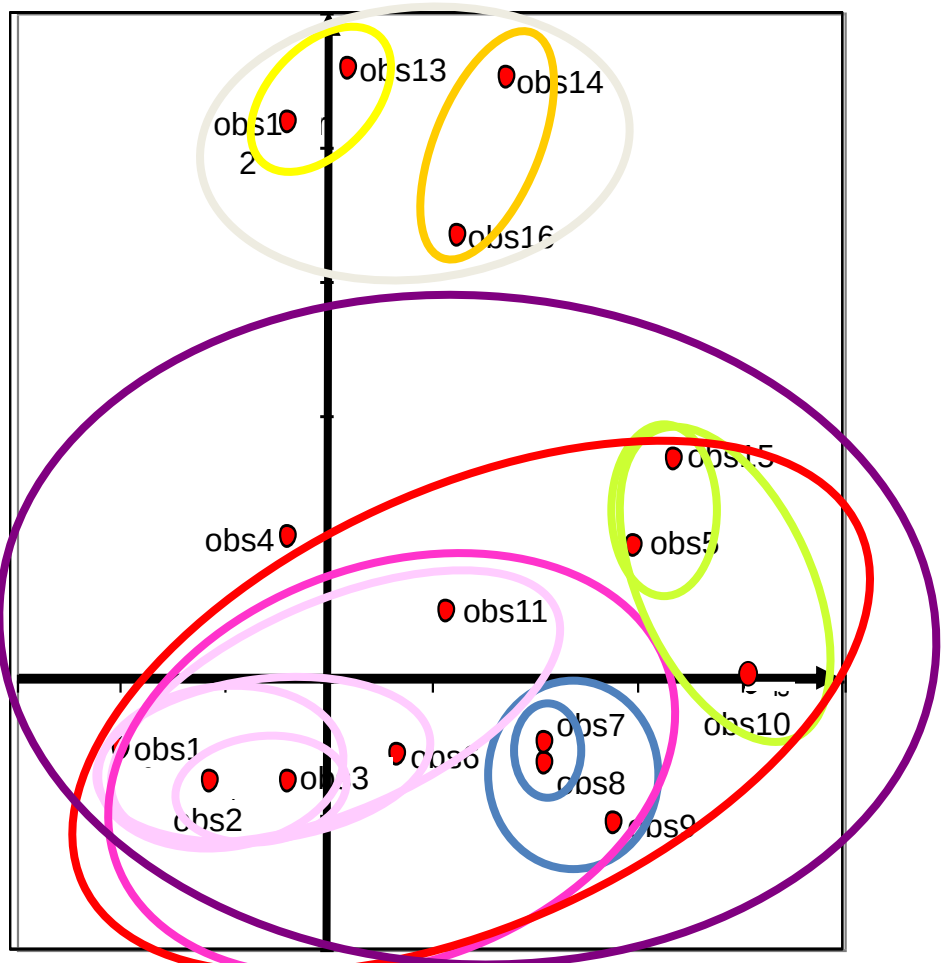
Complete Linkage Cluster Analysis

NCL	Clusters Joined		FREQ	Max Dist
15	obs7	obs8	2	0.1
14	obs12	obs13	2	0.6708
13	obs2	obs3	2	0.7
12	obs5	obs15	2	0.7211
11	CL15	obs9	3	0.8902
10	obs6	obs11	2	1.118
9	obs14	obs16	2	1.2649
8	obs1	CL13	3	1.5297
7	CL12	obs10	3	1.727
6	obs4	CL10	3	1.9416
5	CL14	CL9	4	2.2091
4	CL7	CL11	6	2.7659
3	CL8	CL6	6	3.228
2	CL3	CL4	12	6.0706
1	CL2	CL5	16	6.2434

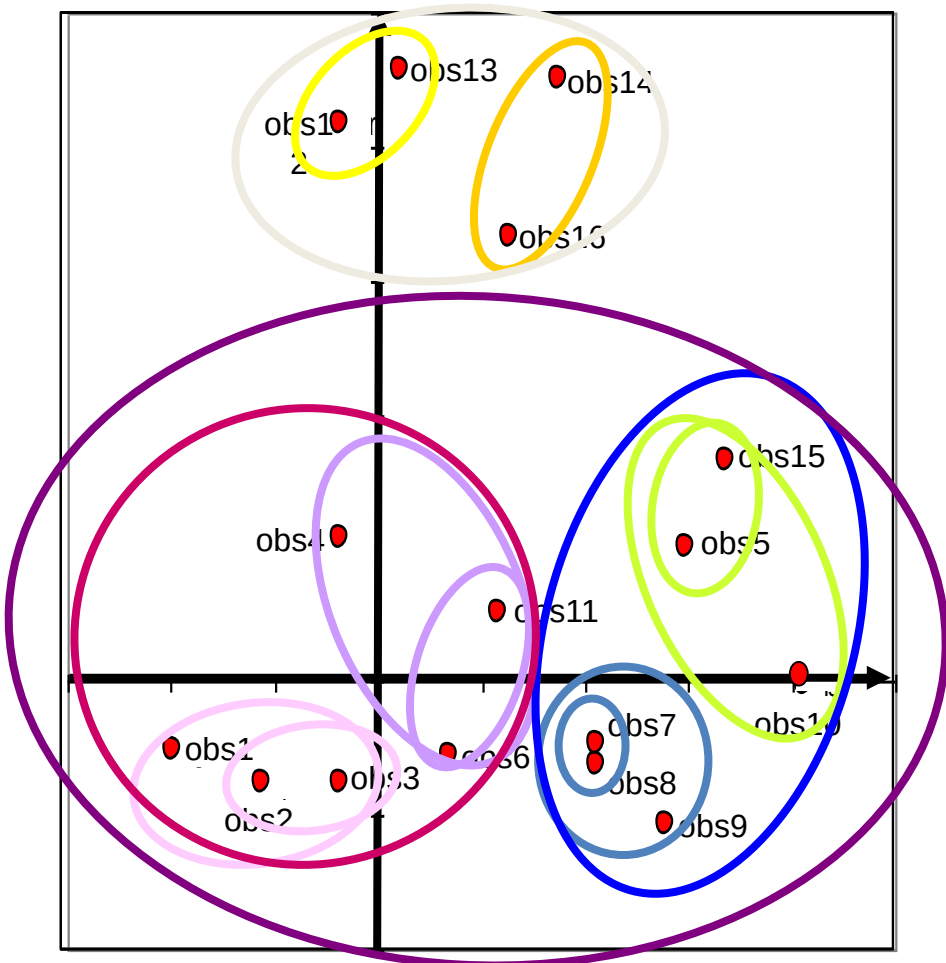
Cluster analysis: hierarchical algorithms

Given a dissimilarity matrix, based on a certain measure of the dissimilarity between cases, there are different methods to measure the dissimilarity between *clusters.* These criteria often lead to different partitions.

Single Linkage Cluster Analysis



Complete Linkage Cluster Analysis



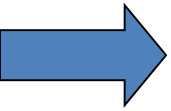
Cluster analysis: hierarchical algorithms

To apply a hierarchical agglomerative algorithm we have to:

1. Obtain the dissimilarity matrix containing the dissimilarities between all the possible pairs of observations (as we will see later, different criteria may be referred to)
2. Choose a method to measure the dissimilarity between clusters

These choices have an impact on the *sequence of nested partitions* obtained as an output. So we usually have different sequences of nested partitions.

But, also, for a given sequence of nested partitions the following problem arises:



How should we select a suitable number of clusters?

Cluster analysis: hierarchical methods/choosing the nr of clusters

We consider first the problem of choosing one out of the clustes solutions obtained with one hierarchical clustering process.

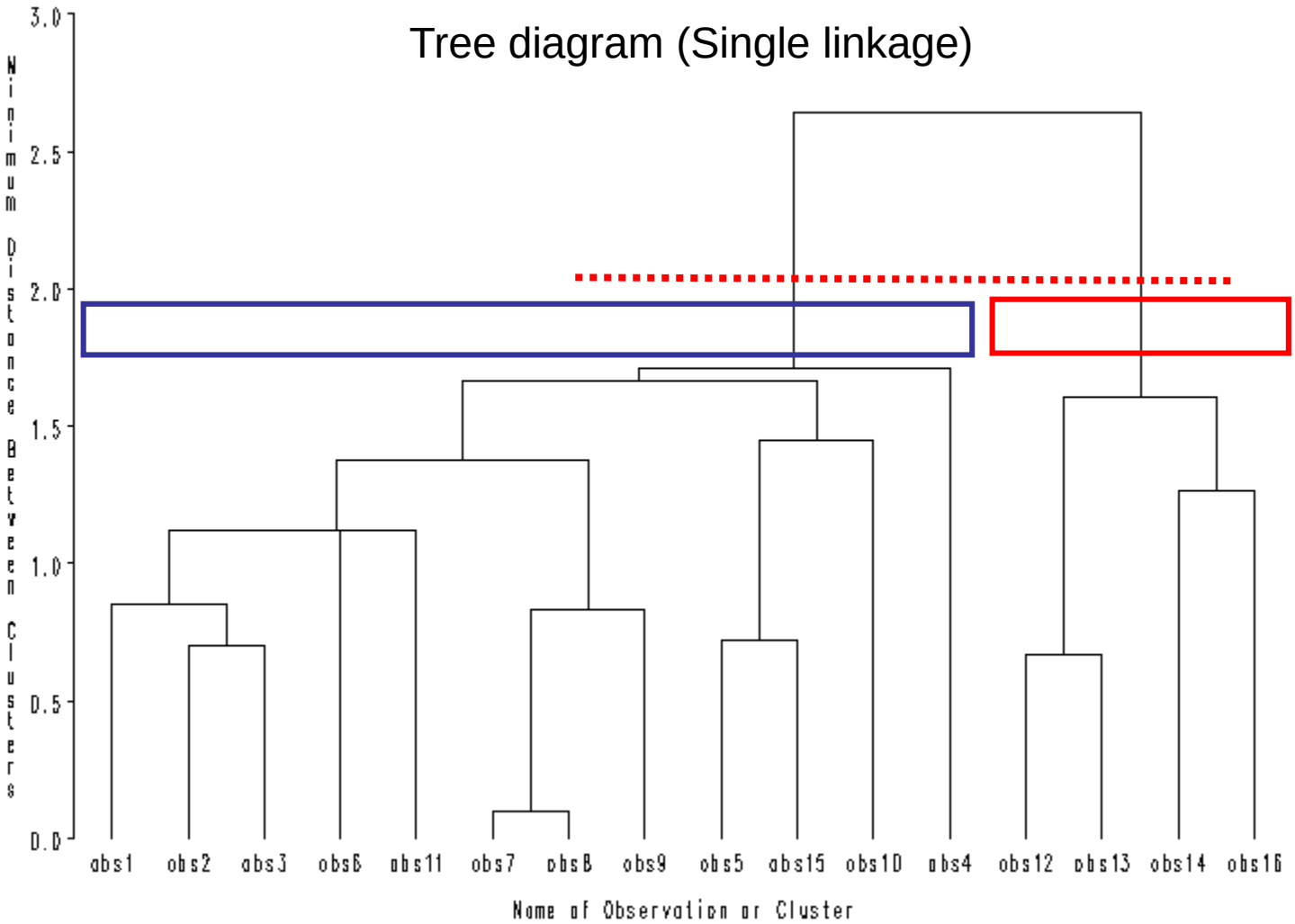
At this aim, the agglomeration process is monitored as the number of clusters declines from *n* to 1, and some quality of clustering criteria are evaluated.

1. Internal criteria. The simplest approach to cluster choice consists in the evaluation of the *dissimilarity between the two clusters joined at each step.* In the first steps of the procedure, similar cases/groups will be joined to form new clusters. At subsequent steps, we can expect an increasing of this dissimilarity, and this increase will tend to grow exponentially in the last aggregation phases, i.e. when very dissimilar clusters are joined.
2. External criteria. Another possibility consists in the evaluation of some statistics – not related to the criterion used to measure the dissimilarity between clusters – which are solely based upon the R^2, the within and the between sum of squares characterizing partition of different degree (different number of clusters)

Cluster analysis: hierarchical methods/choosing the nr of clusters

Internal criteria: Tree diagram (dendrogram) and its height

The agglomerative process can be graphically represented using a tree diagram, also called dendrogram, with cases on the horizontal axis and the dissimilarity between the clusters joined at each step on the vertical axis (the dissimilarity is normalized).



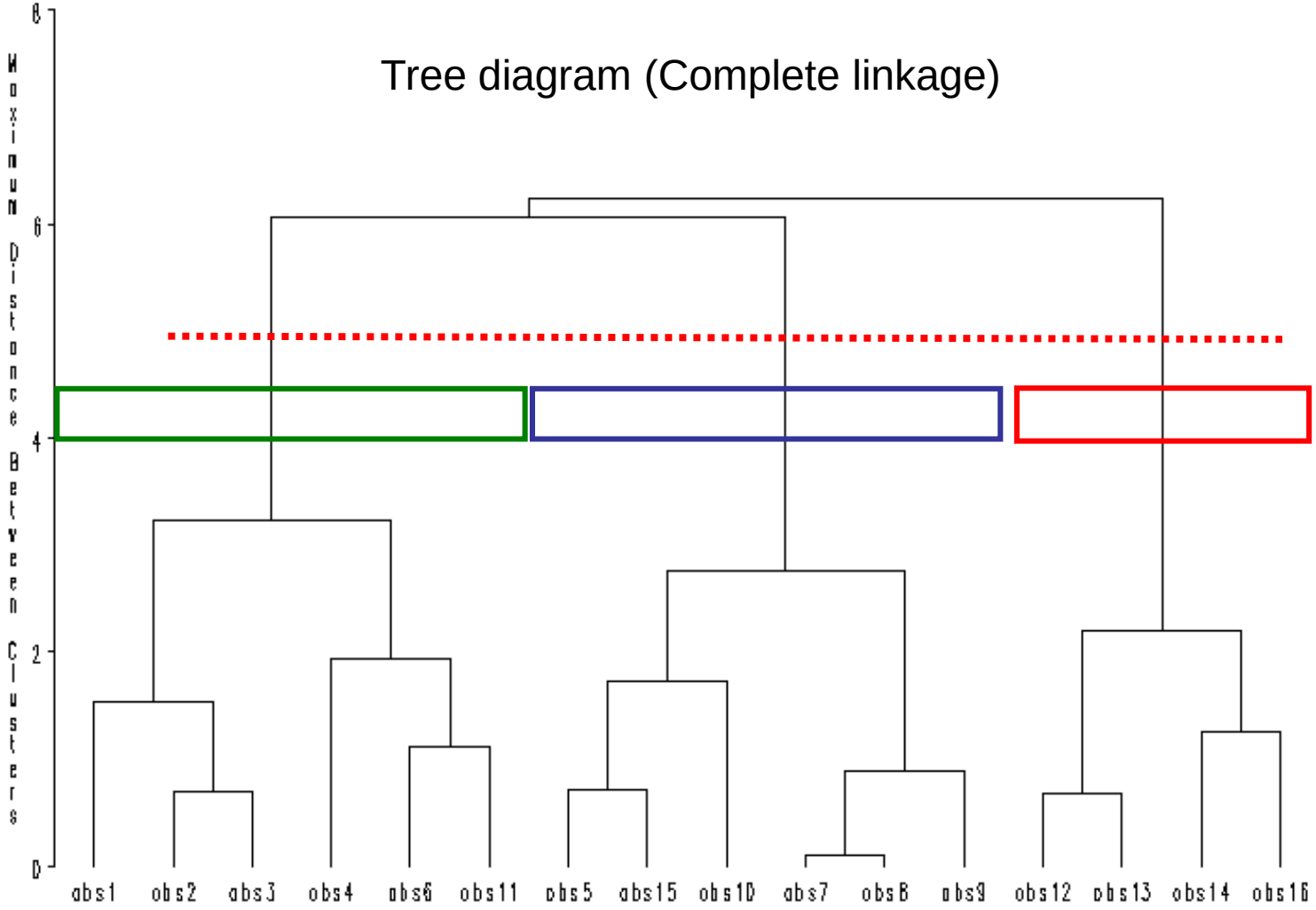
If a large change in the height occurs consequently to an aggregation at step *C* then the (*C* + 1) solution immediately prior to this step should be chosen.

A 'cut' in the dendrogram, corresponding to a given aggregation defines the groups to be considered (obs connected to the branches).

Cluster analysis: hierarchical methods/choosing the nr of clusters

Internal criteria: Tree diagram (dendrogram) and its height

Where would you 'cut' the dendrogram? I.e., which aggregation would you avoid?



Remember that the height of the dendrogram is normalized. Observe that the dissimilarity *between observations is different from one tree to another.* (consider for example the distance between obs 5 and 15)

Cluster analysis: hierarchical methods/choosing the nr of clusters

External criteria: R^2, Pseudo F, Pseudo t^2 .

Criteria based upon the within and between sum of squares. Consider for simplicity the situation when cluster analysis is based upon a vector of measurements (the concepts can be extended to the case when only a dissimilarity matrix is available)

$$\mathbf{T} = \sum_{c=1}^{C}\sum_{i=1}^{n_c}\sum_{j=1}^{p}(x_{ijc} - \bar{x}_j)^2$$

Total sum of squares (SS)

Consider a partition of the dataset into C clusters

$$\mathbf{W}_C = \sum_{c=1}^{C}\sum_{i=1}^{n_c}\sum_{j=1}^{p}(x_{ijc} - \bar{x}_{jc})^2$$

Within SS

$$\mathbf{B}_c = \mathbf{T} - \mathbf{W}_C = \sum_{c=1}^{C}\sum_{j=1}^{p} n_c(\bar{x}_{jc} - \bar{x}_j)^2$$

Between SS

The Within SS sum of squares is a synthesis of the squared errors incurred when using the clusters to “make predictions”/explain the variables at the basis of the clustering procedure. Instead, the Total SS is the synthesis of the squared errors when the general means are used (no external information – clusters)

Cluster analysis: hierarchical methods/choosing the nr of clusters

External criteria: R^2, Pseudo F, Pseudo t^2 .

$$R_C^2 = 1 - \frac{\mathbf{W}_C}{\mathbf{T}} = \frac{\mathbf{B}_C}{\mathbf{T}}$$

R square: quality of a partition. It is related to the proportion of total variation among cases explained by clusters

$R^2 = 1$ when $C = n$ (each case constitute a cluster – no within SS)

$R^2 = 0$ when $C = 1$ (all cases placed in a single cluster –Within SS=Total SS) As the number of clusters decreases the R^2 also decreases. A sudden decrease of the R^2 would indicate the joining of clusters which are really dissimilar

$$\Delta R^2 = R_C^2 - R_{C-1}^2$$

Semi-partial R square: decrease of the R^2 when moving from C clusters to (C – 1) clusters.

Cluster analysis: hierarchical methods/choosing the nr of clusters

External criteria: R^2, Pseudo F, Pseudo t^2 .

Pseudo F statistic $$F_C = \frac{\mathbf{B}_C /(C-1)}{\mathbf{W}_C /(n-C)}$$

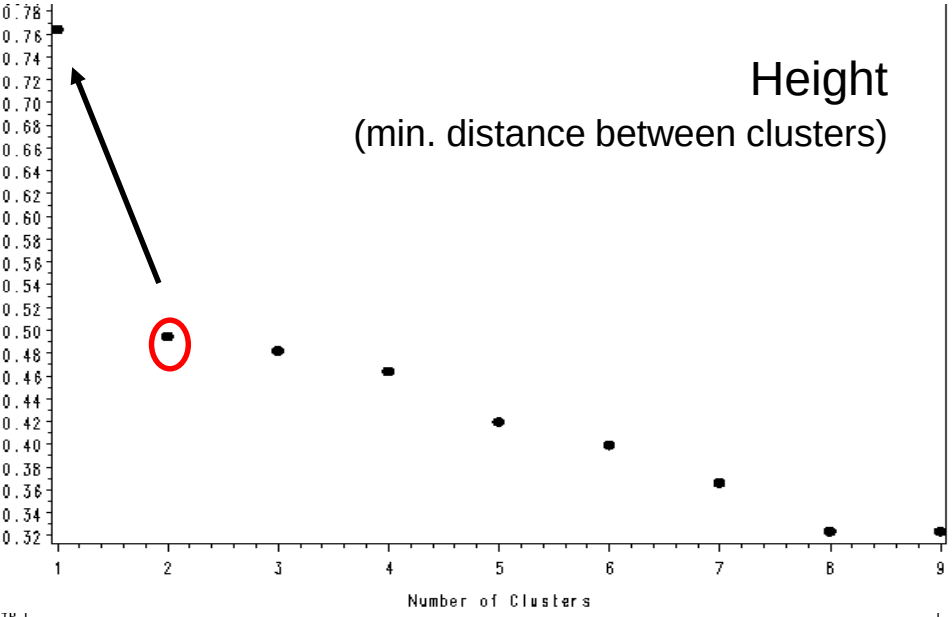
In the initial steps of agglomeration, as n decreases, B decreases and W increases, so F_C gradually decreases. A sudden relatively high decrease of F_C consequent to an aggregation indicates the joining of two quite distinct clusters. The (C + 1) cluster solution immediately prior to this decrease should be selected.

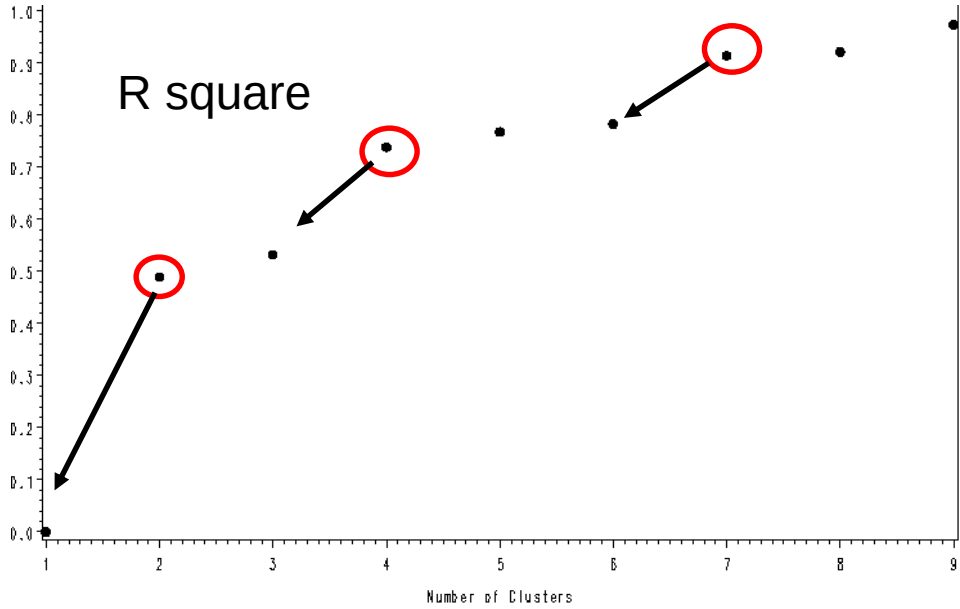
Pseudo t statistic $$t_C = \frac{(SSW_t - SSW_r - SSW_s)(n_r + n_s - 2)}{SSW_r + SSW_s}$$

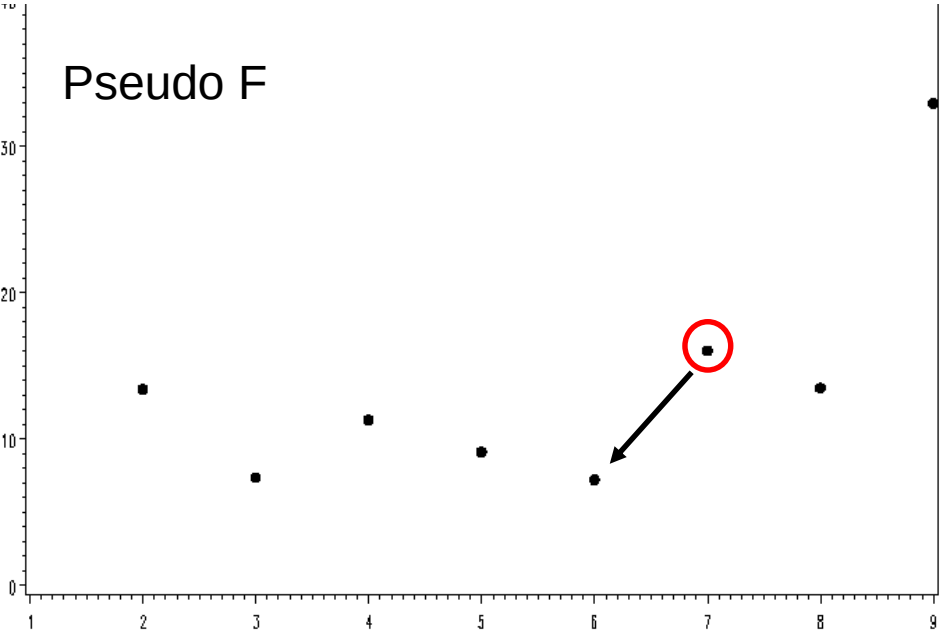
Numerator = increase in the Within SS resulting from joining r and s to form a new cluster. Denominator=sum of the within SS of the two joined clusters. A sudden increase of the statistics indicates the joining of two distinct clusters (high relative increase of the within consequent to aggregation).

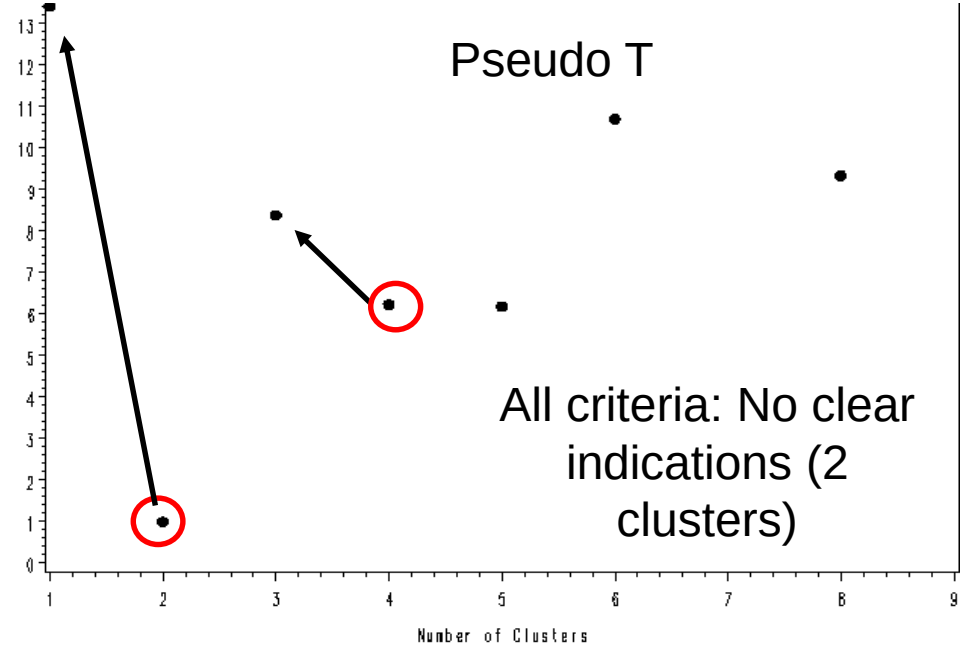
Cluster analysis: hierarchical methods/choosing the nr of clusters

Monitoring internal and external criteria: Single linkage



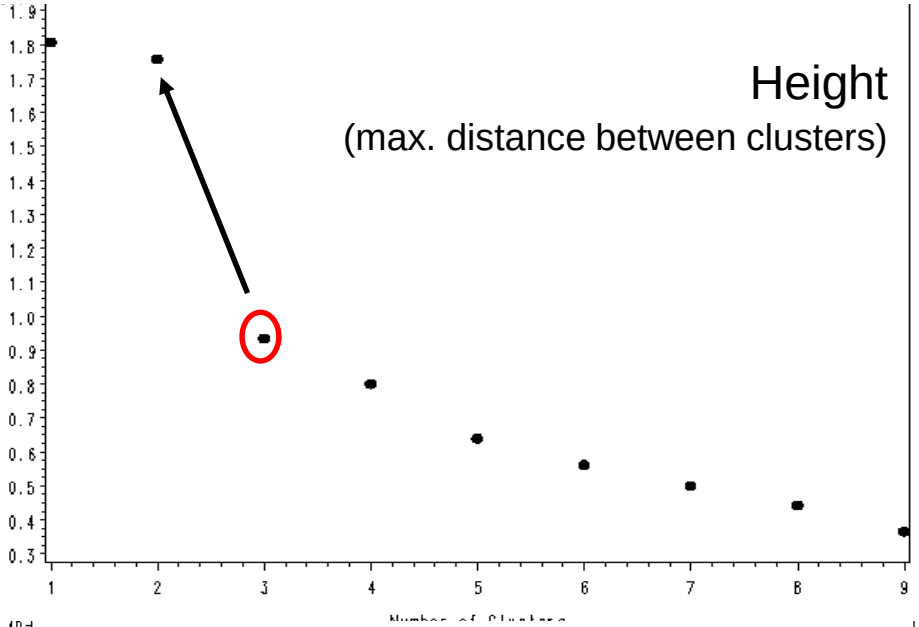


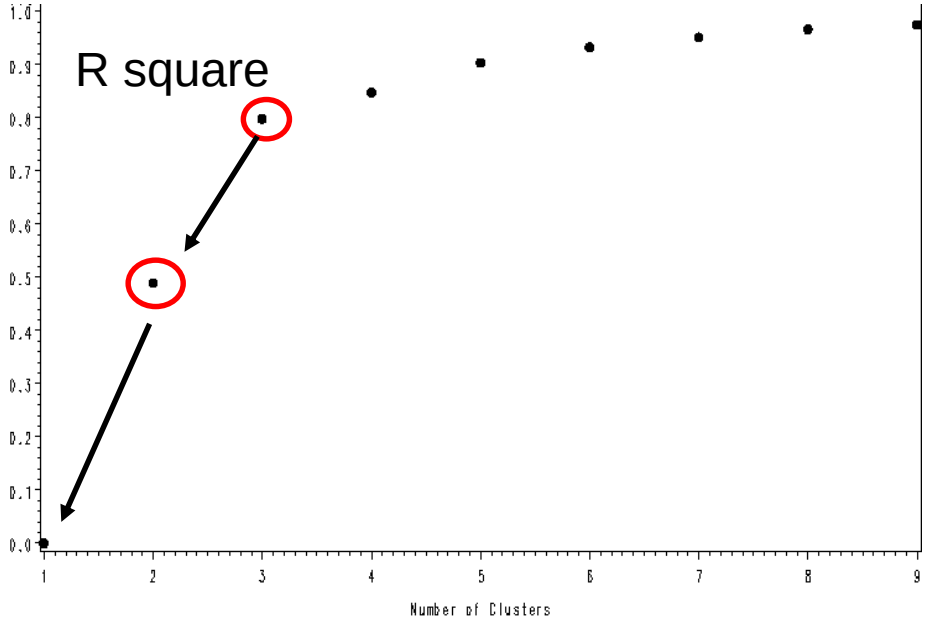


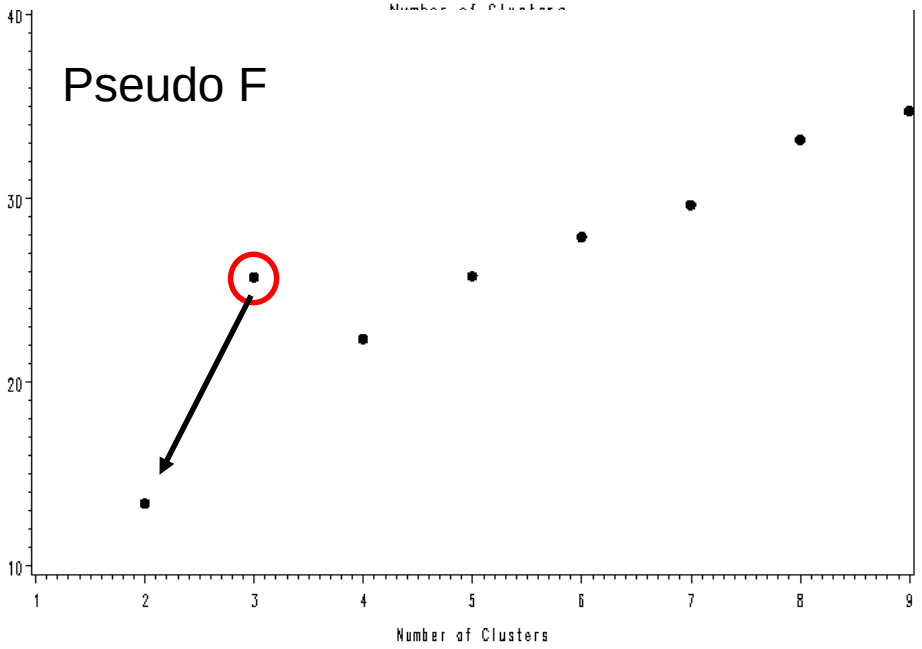


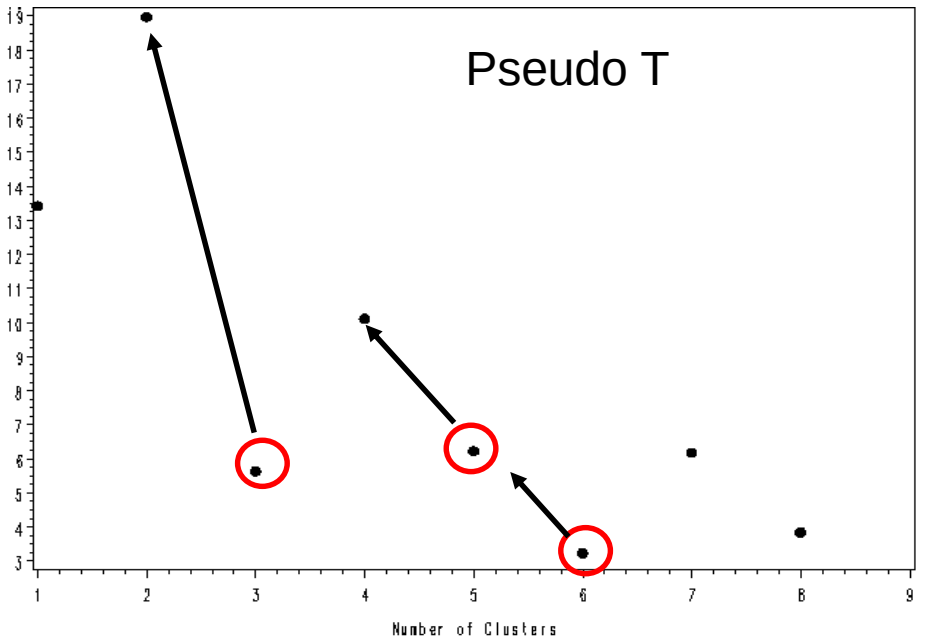
Cluster analysis: hierarchical methods/choosing the nr of clusters

Monitoring internal and external criteria: Complete linkage









Cluster analysis – **partitioning algorithms**

In partitioning algorithms, the number of clusters has to be specified. The algorithm usually starts with an initial allocation of the objects into *G* groups. Then observations are placed in the cluster they are closest to. Alternatively, observations are assigned to one cluster so as to maximize an objective function. The procedure iterates until all objects belong to the closest group (the objective function is maximized) or until a convergence criterion is satisfied .

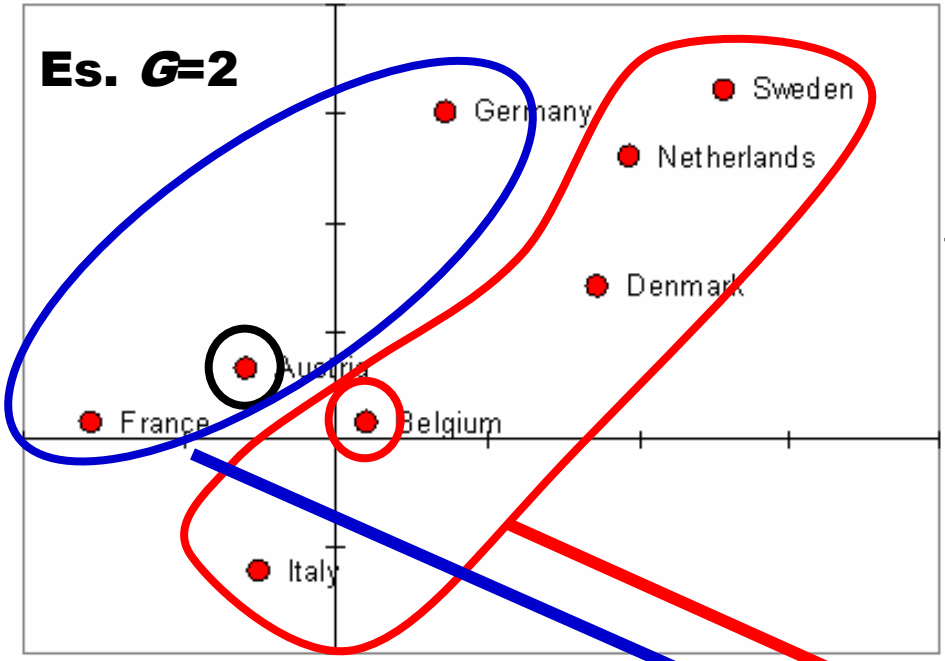
Usually partitioning methods are based upon measurements on a set of variables rather than on a dissimilarity, and on Euclidean distances.

One of the most important partitioning algorithms is the *k-means* algorithm. In this algorithm, the distance from one observation to a cluster is measured as the distance between the observation and the centroid of the cluster.

It can be easily shown that in this case the algorithms attempts to find the partition characterized by the minimum Within SS, i.e., by the maximum R^2.

In this sense, Ward's and the *k*-means algorithms are two R^2-maximizing algorithms. The former is based upon a hierarchical solution to the optimization problem. The latter is instead based on an iterative search of the optimum.

Cluster analysis – partitioning algorithms



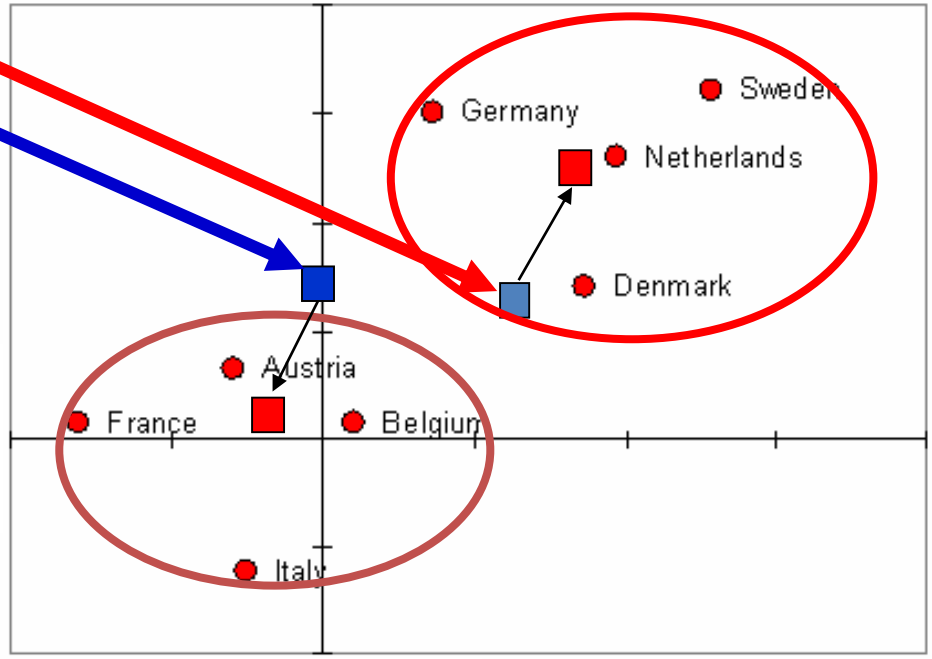
Step 1: Select the initial partition

(this partition may also be defined on the basis of a preliminary cluster analysis / hierarchical procedure)

Usually, *G seeds* are selected

Step 2: Allocation

Each case is allocated to the closest cluster (closest centroid)



Step 3: Seeds update

Seeds are updated: centroid of the obtained clusters

Step 2 and 3 are iterated until convergence:

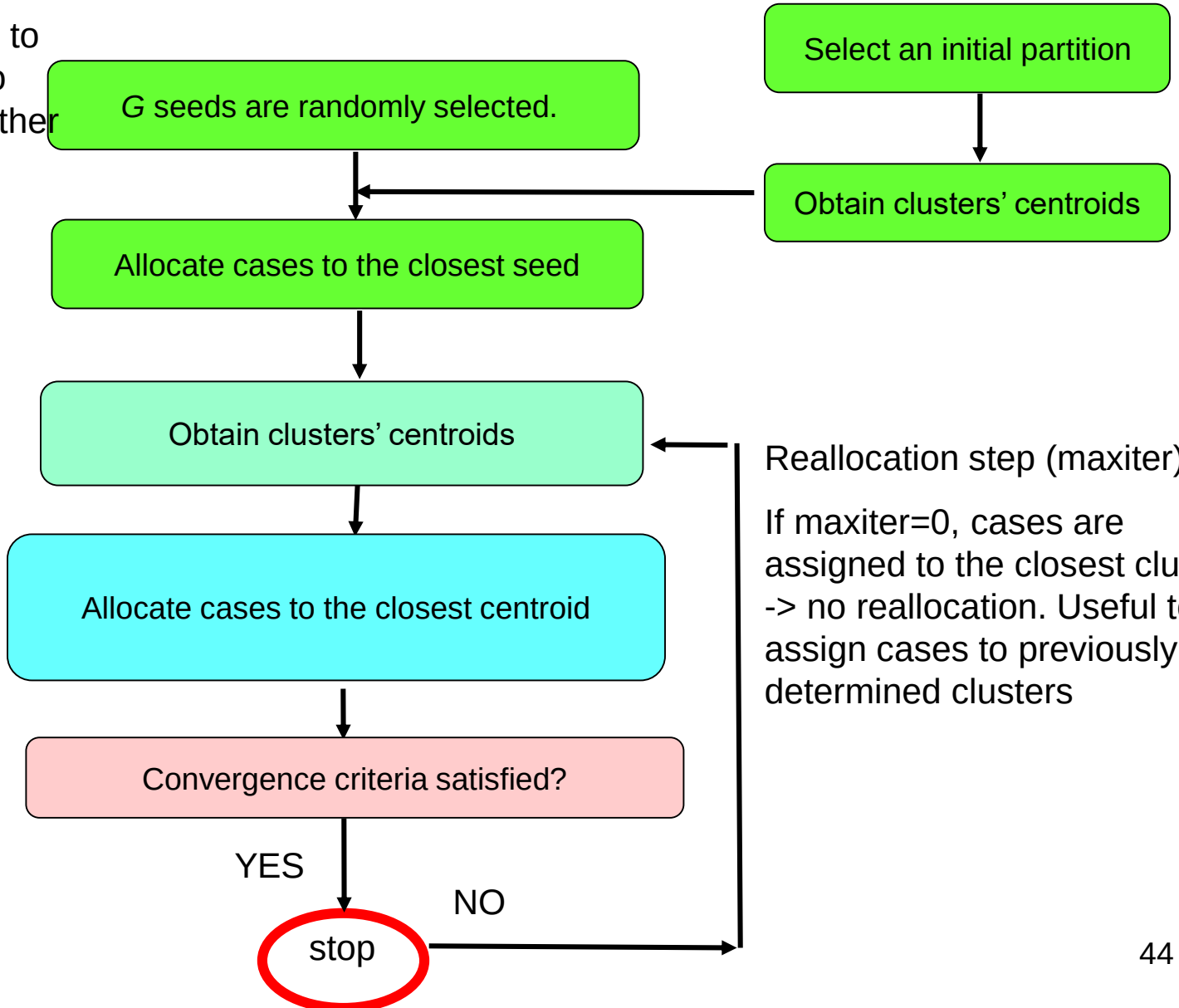
2. Re-allocation

3. Seeds update

Cluster analysis – partitioning algorithms

k-means algorithm

subroutines defined to select seeds not too close one to each other – higher group separation



Reallocation step (maxiter)

If maxiter=0, cases are assigned to the closest cluster -> no reallocation. Useful to assign cases to previously determined clusters

Cluster analysis

Agglomerative algorithms:

- OK Many solutions – monitoring of the process
- OK Flexibility in choosing the measure of dissimilarity (both for obs – see later – and for clusters). The case when only the dissimilarity matrix is available can be handled.
- KO Problems with large datasets (difficulty in handling very large dissimilarity matrices)
- KO Hierarchy is not flexible: once obs are joined they are no longer split. Possible distortions in the case when there are outliers

Partitioning algorithms

- OK Large dataset
- OK Flexible with respect to the aggregation of cases. Groups may change.
- KO Choice of the number of clusters: difficult
- KO Partitions with a different number of clusters are not nested and consequently it may be difficult to analyze the relationship between them.

In some applications combinations of the algorithms are considered

Large databases: a sample of cases is selected. A hierarchical algorithm is applied to select the number of clusters. Then a partitioning algorithm is applied (optionally, the initial seeds are the centroids of the clusters obtained with the hierarchical algorithm)

Flexibility: A hierarchical algorithm is applied. A partition is selected. The centroids of the obtained clusters are used as seeds in a partitioning algorithm. The aim is to evaluate if and to which extent the initial solution changes and, also, to evaluate the possible influence of outliers on results.

Cluster analysis

Whatever the algorithm used to obtain clusters:

1. Number of clusters

- Choice – agglomerative methods
- Guess – partitioning methods

2. Evaluation of the quality of the obtained partition

3. Interpretation of clusters

Internal evaluation:

- Analysis of cases grouped together. Sensible only if obs are identifiable (meaningful labels).
- Analysis of cluster syntheses (means in the case when cluster analysis is based upon numerical variables, other measures – medians, modes – in other cases). This is possible only when measurements on variables are available
- Visualization of clusters in factorial maps

External evaluation

- Evaluation of the characteristics of the clusters (same as before) by referring to variables which were not used to obtain clusters

Cautions

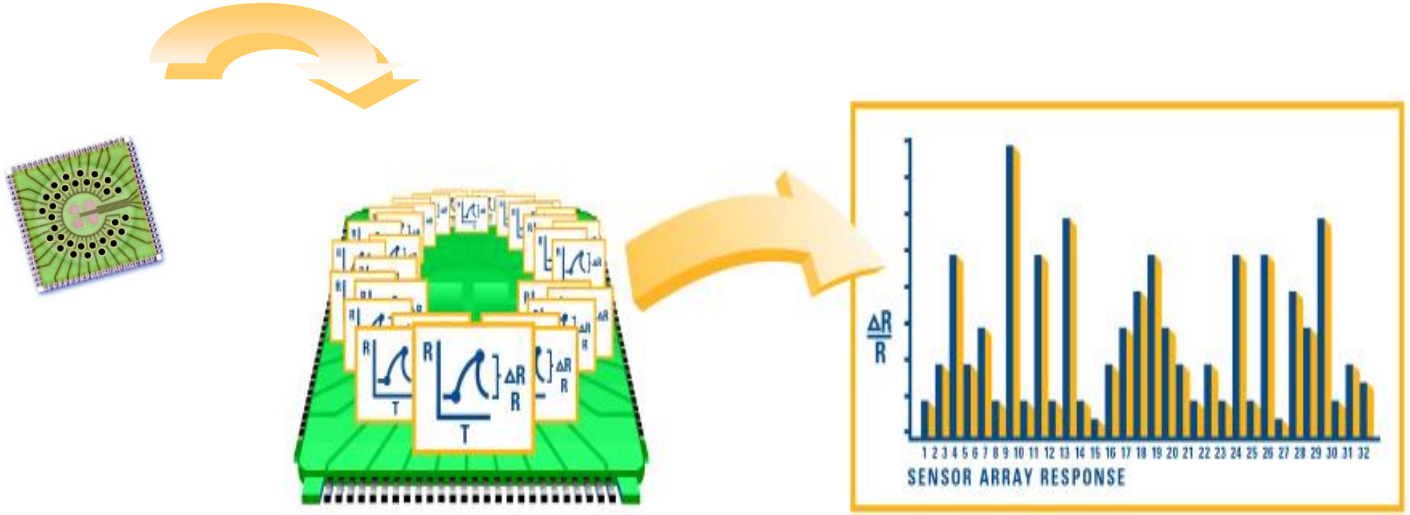
1. **Cluster analysis (as we described it) is a descriptive technique.** The solution is not unique and it strongly depends upon the analyst's choices. We will describe how it is possible to combine different results in order to obtain stable clusters, not depending too much on the criteria selected to analyze data.
2. **Cluster analysis always provide groups, even if there is no group structure.** When applying a cluster analysis we are *hypothesizing* that groups exist. But this assumption may be false or weak.
3. **Cluster analysis results' should not be generalized**. Cases in the same cluster are (hopefully) *similar* only with respect to the information cluster analysis was based on (i.e., dimensions/variables inducing the considered dissimilarities).

9 Esempio IOMS produzione di vettori di dati sperimentali ad alta frequenza, a seguito di somministrazione a array di sensori chimici

Tecnologia Sensoristica –

Rilevazione di composti volatile basata su un insieme organizzato (array) di sensori

Buona soddisfazione con polimeri nanocompositi (Nano Composite Array - NCA)



Nathan S. Lewis
Comparisons between Mammalian and Artificial Olfaction Based on Arrays of Carbon Black−Polymer Composite Vapor Detectors
Acc. Chem. Res., 2004, 37 (9), pp 663–672 DOI: 10.1021/ar030120m

9 Esempio di IOMS – instrumental odour monitoring system

Un sistema di sensori ibrido con a bordo diverse tipologie di sensori (NC

coo

atti

Produce dati con frequenza al minuto e consente di raccogliere campioni d'aria da analizzare in laboratorio in caso di segnalazioni di molestia.

10. L’addestramento del sistema / training

Il nostro approccio è inizialmente non supervisionato (unsupervised - NTA 9065:2012 Air quality – Electronic air monitoring – Odour (nuisance) and safety)

si procede alla rilevazione di segnali del sistema multisensore per un periodo rappresentativo (anche mesi) per mappare variabilità della composizione dei vapori rilevabili nel sito

si identificano quindi vettori/pattern ricorrenti ch evengono poi raggruppati (algoritmo Self Organizing Map)

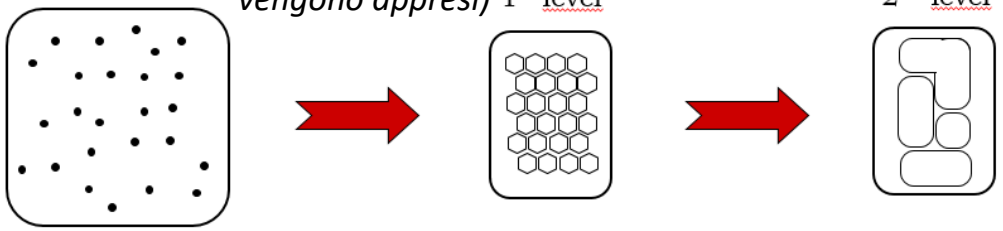
Unsupervised data elaboration by Artificial Neural Networks:

Self Organizing Maps (SOM)

DATO SENSORIALE, INFORMAZIONE MEMORIZZATA, PRIMA GENERALIZZAZIONE

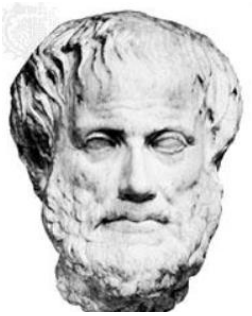
Sensazione *(dato empirico, sperimentale)*

Memoria *(«Pattern ricorrenti» vengono appresi)*



Esperienza: *non è singola sensazione ma l'accumularsi di sensazioni grazie alla memoria* ***: mettendo insieme una serie di casi singoli si riesce ad arrivare ad una prima forma di generalizzazione («Che cosa»)***

POWERFUL DATA VISUALIZATION



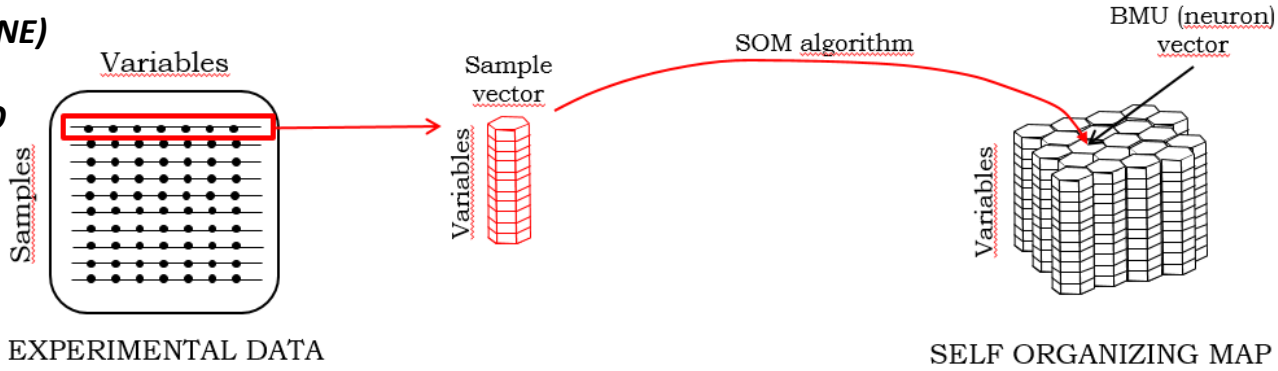
SIMPLIFICATION and INFORMATION EXTRACTION

- Kohonen, T. (2001). Self-organizing maps, third ed. Berlin: Springer.
- Himberg, J., Ahola, J., Alhoniemi, E., Vesanto, J., Simula, O., 2001. The Self-Organizing Map as a Tool in ***Knowledge*** Engineering DOI:10.1142/9789812811691_0002

DA «SENSAZIONE» VETTORE DI DATI RILEVATI DAI SENSORI A PATTERN MEMORIZZATO (NEURONE)

DA «TABULA RASA» A UN MODELLO DEI DATI SENSORIALI, UNA MAPPA DELLA REALTA' OGGETTO DI STUDIO

How the SOM algorithm works



1. A sample is presented to the initialized SOM (the map dimension ratio depends on the first two eigenvectors of the data - Vesanto 2000);
2. The algorithm identifies (in terms of distance) the Best Matching Unit (i.e. neuron) for the sample;
3. The **Best Matching Unit (BMU)** adjusts itself (decreasing the distance) and its neighbors according to the sample presented, thus it "learns" from the sample;
4. After all the samples are presented to the SOM 1 epoch is finished;
5. The process can be iterated for a wisely selected number of epochs avoiding overfitting.

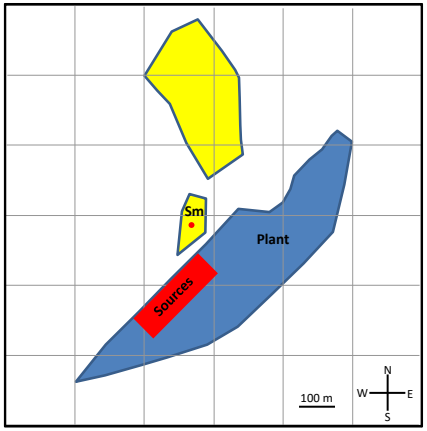
11 Visualizzazione bidimensionale non lineare - su più piani

I DIVERSI PIANI DELLA MAPPA RAPPRESENTANO CIASCUNO LA VARIABILITA' DI UN SENSORE NELLE SITUAZIONI DESCRITTE DALLA MAPPA (RAPPRESENTATIVITA'VERIFICABILE RISPETTO A NUOVI DATI SENSORIALI)

POSIZIONI VICINE NELLA MAPPA INDICANO GENERALMENTE RELATIVA SIMILITUDINE A SEGUITO DEL TRAINING CHE GENERA VETTORI CARATTERISTICI/NEURONI DELLA MAPPA MODIFICANDO BMU E SUO INTORNO, **MA** NON VI E' METRICA/DISTANZA LINEARE SULLA MAPPA (possono esserci aree omogenee e discontinuità, sulla base dell'eterogeneità dei dati sensoriali/vettori sperimentali impiegati nel training)

CASE STUDY Bio-waste composting facility

(August - December 2017)



Map of the investigated area
(**Sm**=sampling site;
blue area=plant;
red area= plant odor sources;
yellow areas=dwellings).

1 group of dwellings at 100 m

1 group of dwellings at 300 m

Sampling point at 100 m, near dwellings

Electronic nose – 33 sensors by Sensigent (Baldwin Park CA – U.S.A.)

Automatic sampling system (OdorPrep, by Lab Service Analytica S.r.L., Italy) for air samples collection

Three sources sampled in bags by a **manually activated "lung principle" sampler**

Odor concentration analysis (EN 13725:2003) by a dynamic olfactometer (WOLF **??**by ArcoSolutions s.r.l., Italy)

Remote activation in correspondence of citizens' complaints

S. Licen, S. Cozzutto, M. Angelucci, P. Barbieri «Self Organizing Map algorithm as a tool for analysis, visualization and interpretation of electronic nose high dimensional raw data" NOSE 2018 Milan, 9-12 September 2018

DATA AVAILABLE

Four months survey → more than **140'000** minutes recorded → x **33** sensors → more than **4.5 millions** of single sensor values

10 air samples collected at the receptor → Odor concentration in OUe m^{-3}

3 sources sampled and submitted to the electronic nose → E-nose records

HOW TO PUT IN A COMPREHENSIVE FRAME ALL THE DATA AVAILABLE?

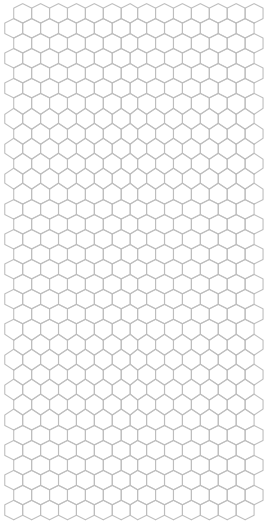
WE STARTED FROM THE ELABORATION OF E-NOSE AMBIENT AIR DATA BY SELF ORGANIZING MAP ALGORITHM

SELF ORGANIZING MAP BUILDING

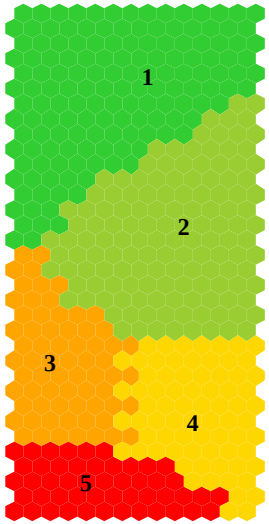
Four months survey → more than **140'000** minutes recorded → x **33** sensors → more than **4.5 millions** of single sensor values

SOM map composed by **476** neurons (34x14) representing "**air types**" perceived at the receptor

5 Clusters = **"air types" classification**



K-means clustering →



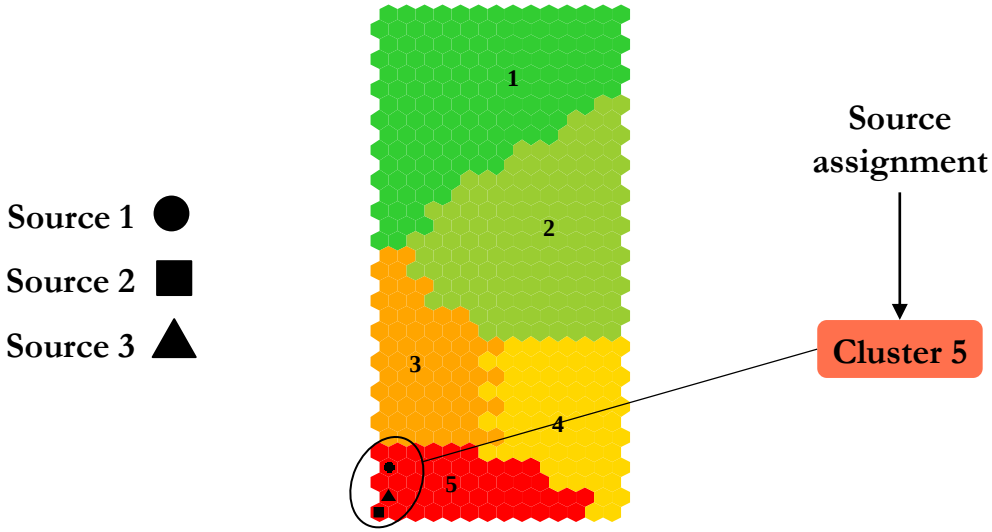
The neurons are depicted by hexagons stuck together in a 2D map

Cluster characterization (I)

3 source profiles recorded by the electronic nose

The source profiles were **projected** onto the map, thus:

1. The source profile is presented to the SOM model;
2. The algorithm identifies (in terms of distance) the **Best Matching Unit** (i.e. **neuron**) for the sample;
3. A symbol is depicted onto the map in correspondence of the assigned neuron

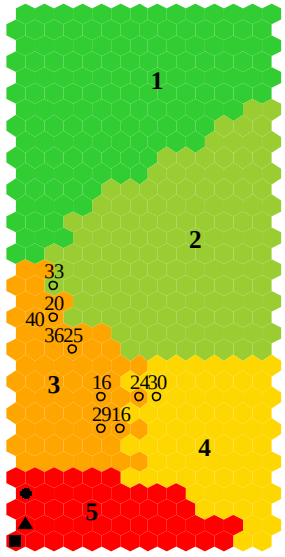


Cluster characterization (II)

10 air samplings followed by olfactometric analysis (EN 13725)

They **only common variable** between high frequency data and ambient air samplings is the **sampling date/time** (2-3 minutes for each sample), thus:

1. The air sampling with OdorPrep **date/time** was identified ;
2. The e-nose data vectors corresponding to the **same date/time** were identified;
3. The **map neurons** representing the abovementioned data vectors were identified;
4. The odor concentration values obtained by the olfactometric analysis were directly depicted onto the map over the abovementioned neurons (represented by hexagons)

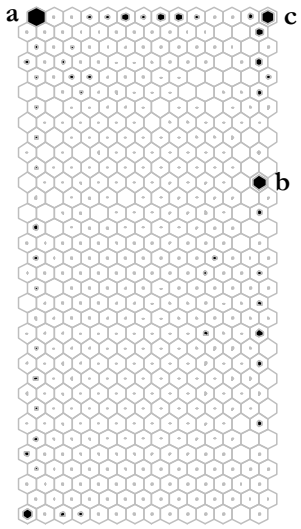


Odor concentration
(European Odour Units/m³)

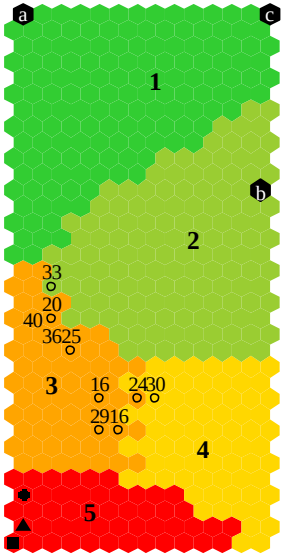
Cluster characterization (III): hits & cluster frequency (+ duration)

HITS = number of experimental vectors represented by each neuron

a-b-c: neurons representing the highest number of experimental vectors (hits):



The filling of the hexagons is proportional to the number of hits



Cluster assignment :

1 – not odorous;

2 – not odorous;

3 – modestly malodorous;

4 – not determined;

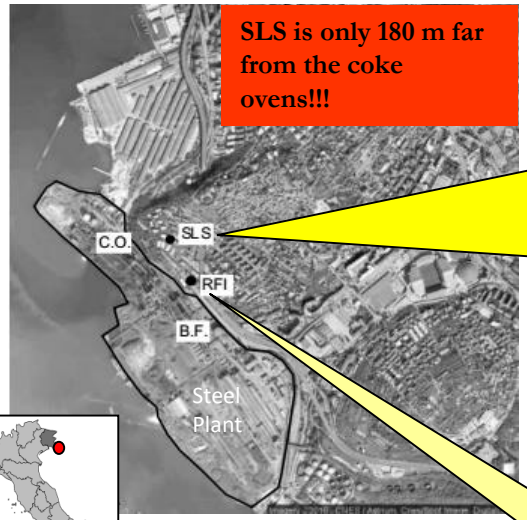
5 –malodorous.

Cluster frequency
39 %
25 %
12 %
12 %
12 %

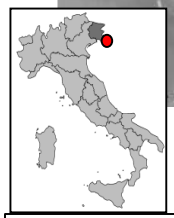
Cluster duration (%)

Cluster	0-1 h	1-2 h	2-4 h	4-8 h	8-12 h	12-24 h	24-48 h	48-60 h
1	71	8	4	4	6	5	1	1
2	82	9	4	3	1	1	0	0
3	86	5	4	4	1	0	0	0
4	93	3	1	2	1	0	0	0
5	84	4	4	4	4	0	0	0

A CASE STUDY AT RECEPTORS CLOSE TO A STEEL PLANT



C.O. = coke ovens; B.F. = blast furnace



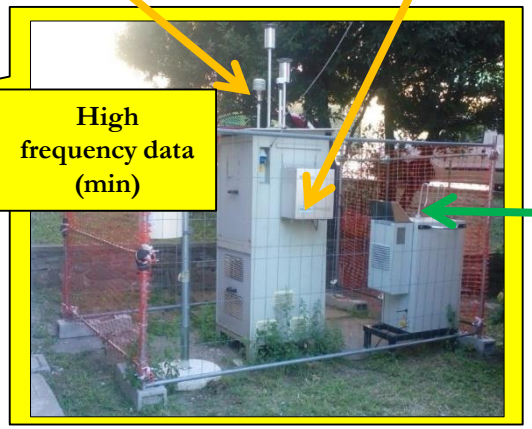
Italy,
Friuli Venezia-Giulia region,
Trieste city

Optical Particle Counter (MetOne Instruments)

Multisensor array MSEM-32 (Sensigent)

ARPA FVG

High frequency data (min)



Remotely activated **air sampler** Odorprep (Lab Service Analytica)

Analysis by dynamic olfactometry according to EN 13725 within 6h (ARCO Solutions srl)

Pollutant monitoring station (Regional Environmental Protection Agency ARPA-FVG)

Citizen complaints records (Local Police)

Meteorological data

S. Licen, S. Cozzutto, G. Barbieri, P. Barbieri «Assessing impacts from very variable sources of perceptible pollutants: smart tools for the integration of high frequency instrumental data with ancillary chemical, meteorological and subjective information - **XVII Congresso Nazionale di Chimica dell'Ambiente e dei Beni culturali Genova, 24-27 giugno 2018**

SELF ORGANIZING MAP BUILDING

Four months survey
(June-September 2015)

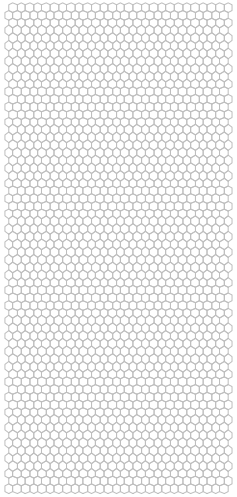
120'000 sample vectors
(1 vector per minute)

more than **3 millions**
of single data

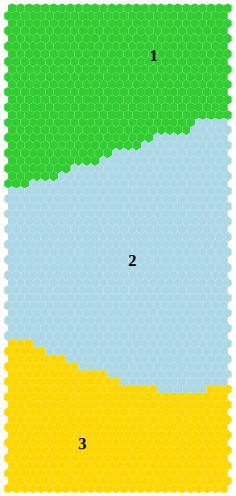
x **21** sensors (E-nose)
x **8** channels (OPC)

Self Organizing Map
composed by **1728** neurons (64x27)

(Dimension ratio depends on first 2 eigenvectors of the data, Vesanto 2000)

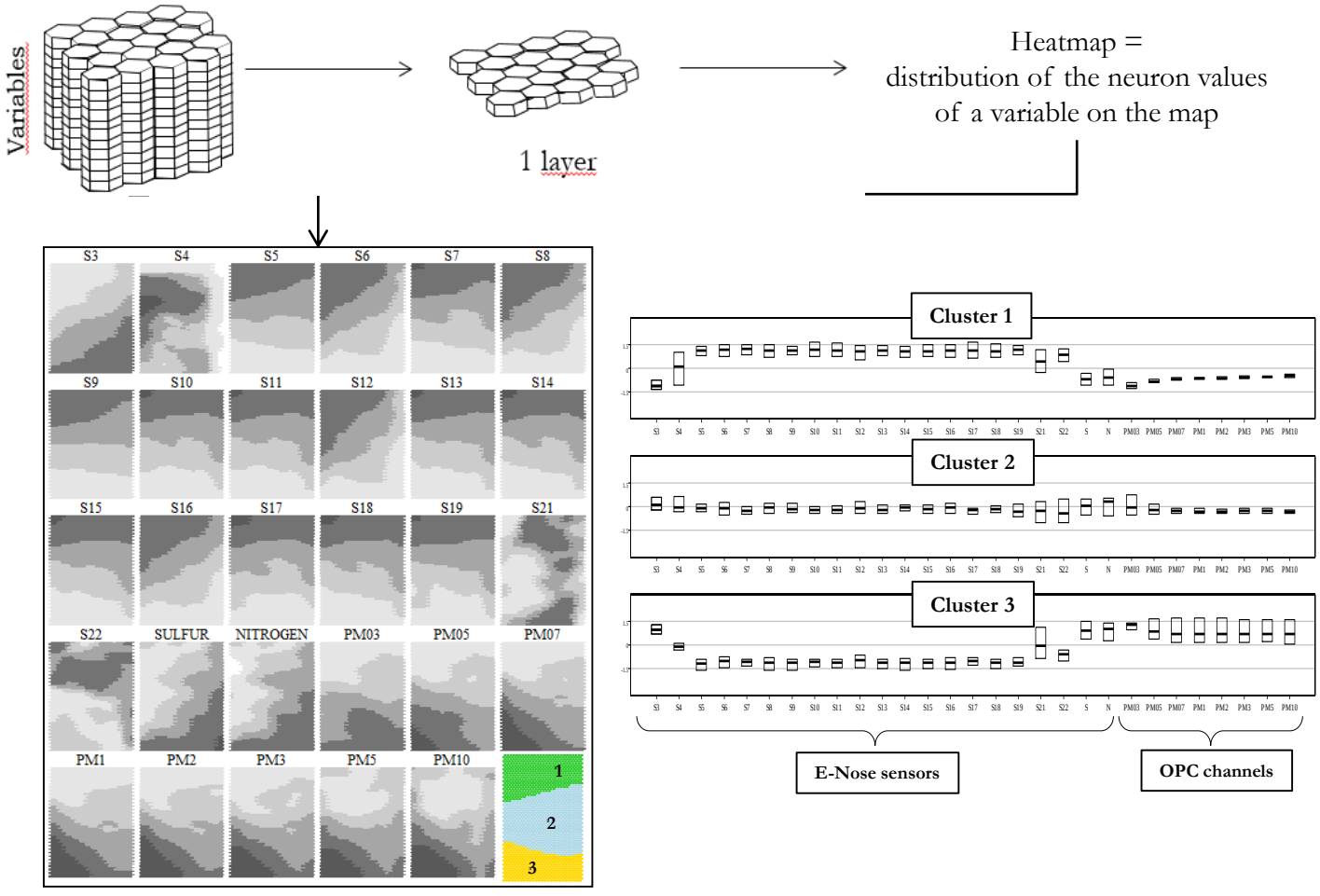


3 Clusters = "air types"



(MATLAB implemented with SOM toolbox + R implemented with openair package and in-house scripts)

Cluster characterization: heatmaps & cluster profiles

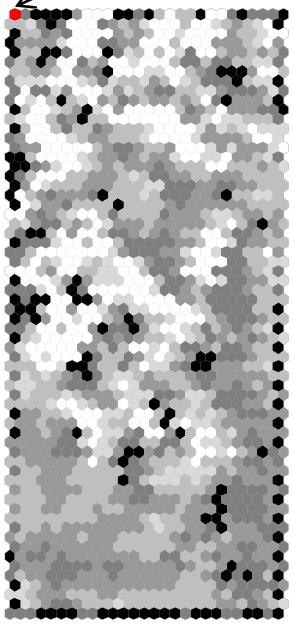


Cluster characterization: hits & cluster frequency

HITS = number of experimental vectors represented by each neuron

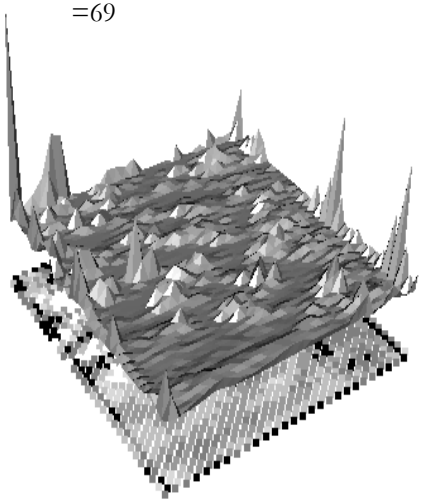
Neuron representing the maximum number of hits:

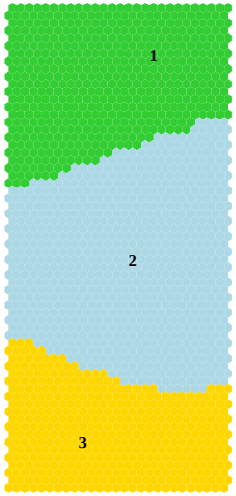
3743 hits



Grayscale from white to black
→ Growing number of hits

Mean n° of hits =69





31 %

40 %

29 %

Cluster characterization: scattered ancillary data

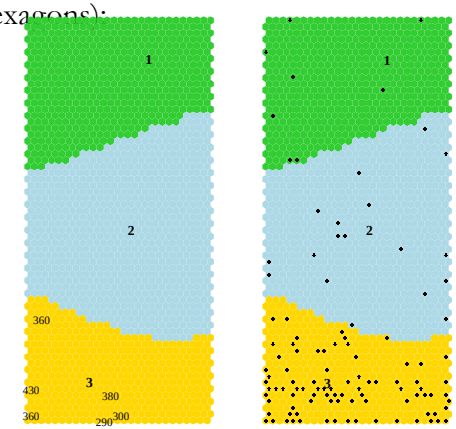
150 citizen complaints records classfied as «Odor» and/or «Dust»

6 air samplings followed by olfactometric analysis (EN 13725)

They **only common variable** between high frequency data and scattered ancillary data is the **sampling date/time**, thus:

1. The scattered ancillary data **date/time** was identified ;
2. The high frequency data vectors corresponding to the **same date/time** were identified;
3. The **map neurons** representing the abovementioned data vectors were identified;
4. The scattered ancillary data were directly depicted onto the map over the abovementioned neurons (represented by hexagons):

Odor concentration *(European Odour Units/m³)*



Citizen complaints

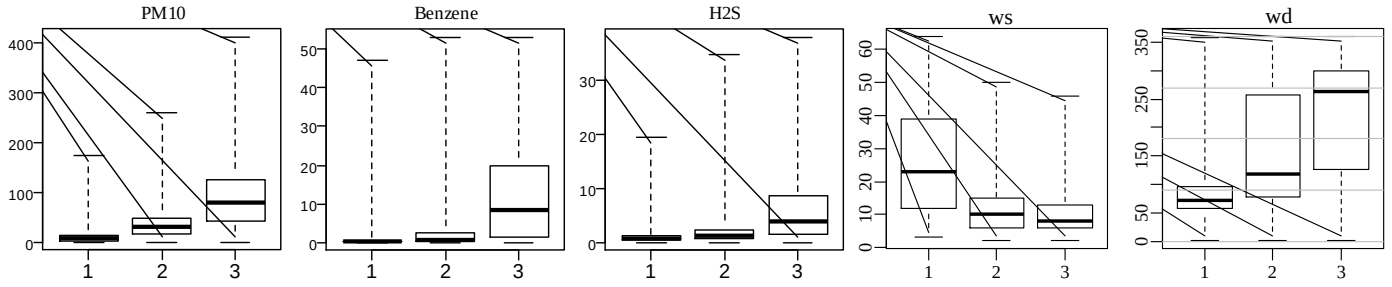
Cluster characterization: «medium» frequency ancillary data

2000 hourly vectors of pollutant data (benzene, CO, PM_{10}, H_2S, NO_2, NO_x)

2000 hourly vectors of wind data (speed and direction)

Pollutant and wind speed and direction data (collected 200 m apart – RFI) can be used as well, depicted in **boxplots** according to clusters:

1. Each cluster represented a number of e-nose experimental data collected at specific date/time values.
2. The abovementioned devices outputs were grouped according to the same date/time values (i.e. according to the clusters) and the groups obtained were represented in box-plots.

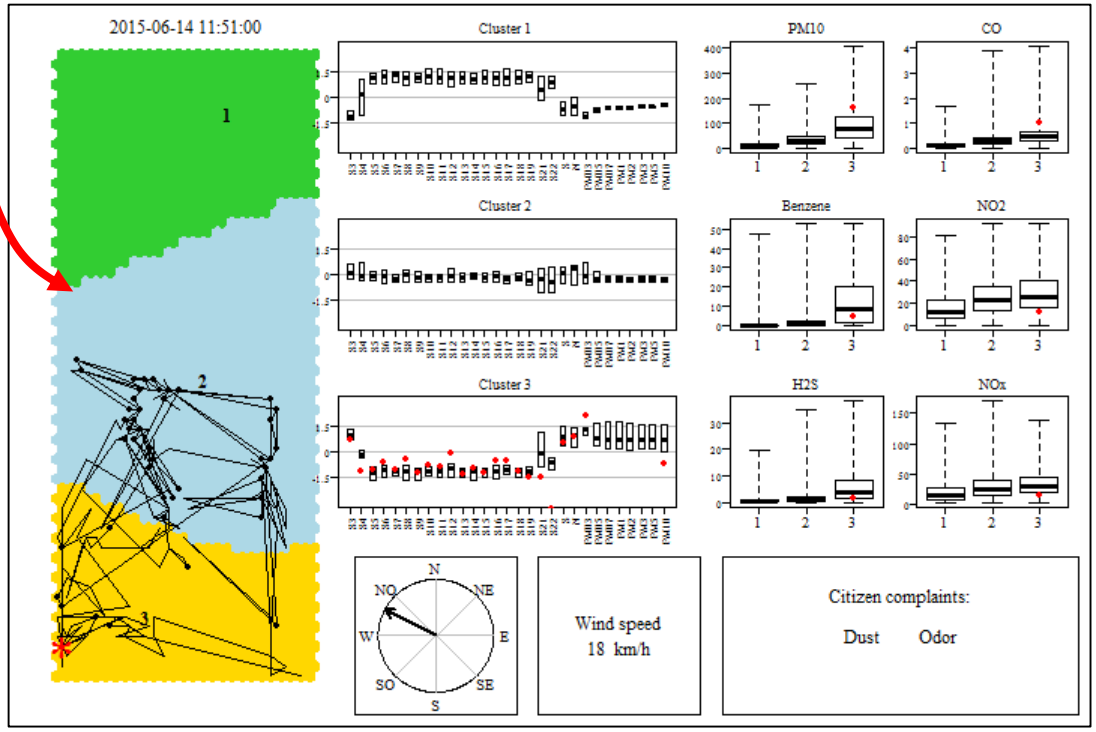


The combined use of all the available information led to the following cluster assignment:

Cluster assignment : 1- low impact; 2 – medium impact; 3 – high impact

BACK TO DETAIL

Follow the trajectory!!!



13. Interpretazione («il perché») dei raggruppamenti di vettori caratteristici/neuroni si ottiene correlando il modello (la mappa) dei dati prodotti dal sistema multisensore con altre matrici di dati che condividono con i dati del sistema multisensore una modalità di raccolta: il tempo a cui avviene l'evento.
Al tempo t_x in cui il dato prodotto dal sistema multisensore corrispondeva a vettore caratteristico k sulla mappa, si sono verificate:

- segnalazione di molestia olfattiva della cittadinanza,

- concentrazione di X ug/m3 di marker di sorgente (es. B/T o H2S) da il monitoraggio continuo «tradizionale»

- concentrazione di odore di Y uoE/m3, rilevata in laboratorio su campione prelevato in situ da campionatore attivato da remoto al tempo

$t_x + \Delta_1 t$ ($\Delta_1 t$ piccolo)

- il vento spirava da direzione Z,

- evento critico in impianto*!!! *a $(t_x - \Delta_2 t)$

è DATA FUSION (DATA STACKING)

si identificano attributi caratteristici della tipologie d'aria identificate dal sistema multisensore

https://doi.org/10.1016/S0003-2670(01)00991-6

https://www.researchgate.net/publication/261181646_Clustering_in_Analytical_Chemistry