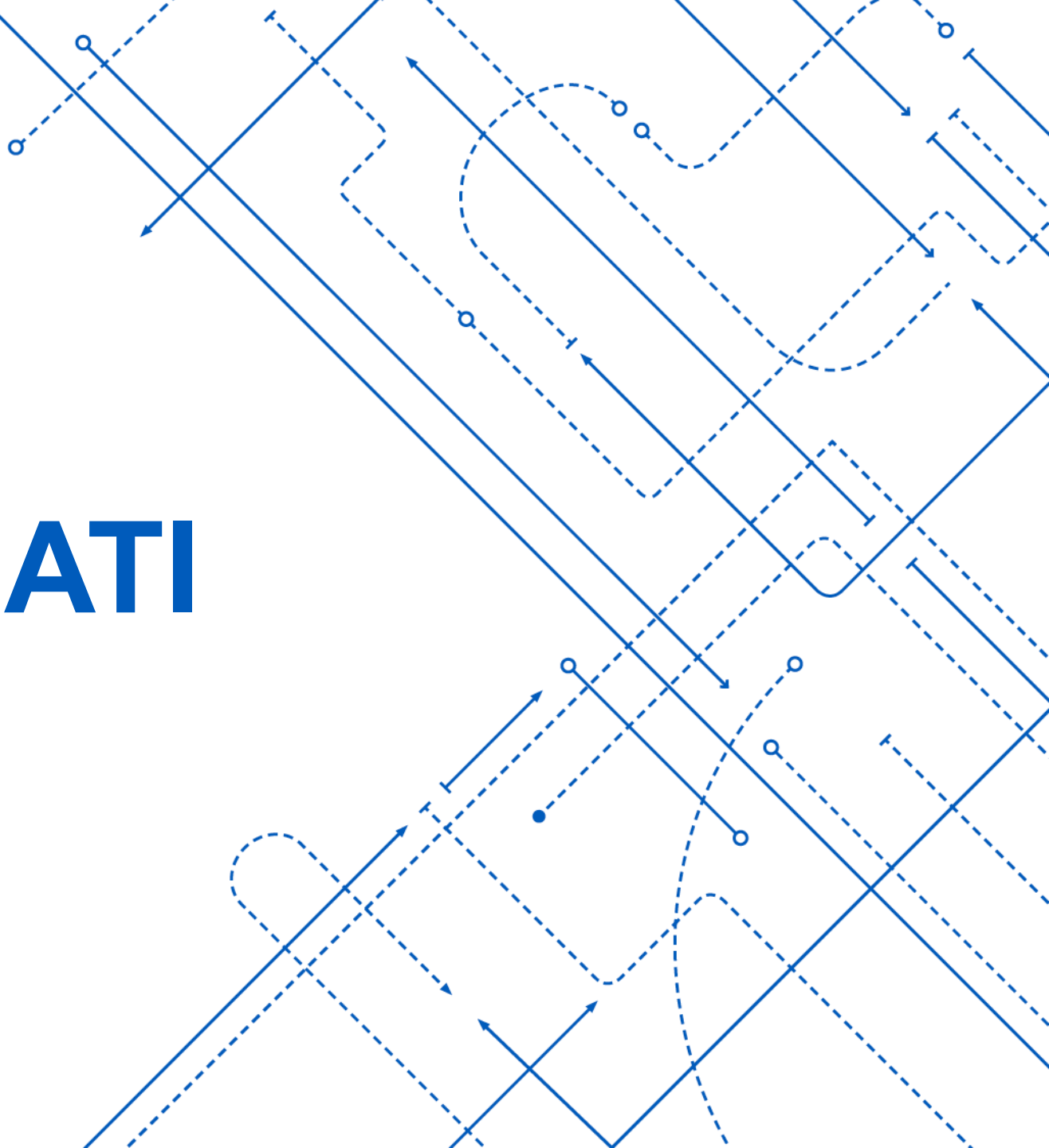
STATISTICA E ANALISI DEI DATI

Capitolo 6 – Introduzione al clustering

Dott. Stefano Cirillo
Dott. Luigi Di Biasi

a.a. 2024-2025



Dal mondo ideale … al mondo reale

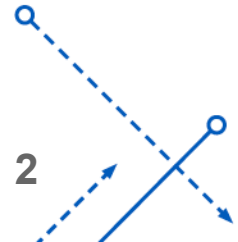
I metodi statistici visti fino ad ora permettono di studiare ***"come si comportano"*** uno o più fenomeni:

- tramite l'osservazione diretta di grafici;
- **misurando** e **calcolando** diverse quantità numeriche.

← Mondo ideale!

Nella realtà, si hanno a disposizione un **grande numero di osservazioni** che sono praticamente intrattabili.

C'è ne siamo già accorti, discutendo un po' dei dataset!



Statistica e analisi dei dati

```r
X <- c(24 ,26 ,30 ,25 ,29 ,27 ,20 ,29 ,27 ,28 ,18 ,21 ,26 ,30 ,28);
Y <- c(27 ,26 ,29 ,26 ,30 ,27 ,22 ,29 ,27 ,28 ,20 ,20 ,27 ,30 ,30);
df <-data.frame (X,Y);
#plot(df$X,df$Y);
median(df$X)
median(df$X)
sd(df$X)
median(df$Y)
median(df$Y)
sd(df$Y)

cov(df$X,df$Y);
cor(df$X,df$Y);
```

Mondo ideale!

Mondo reale!

	geography_desc	year_desc	category	commodity_desc	amount
25688	Zhejiang	1970	Grain production	total grain	11235
25689	Zhejiang	1970	Grain sown area	total grain	3203.2667
25690	Zhejiang	1969	Grain production	total grain	10480
25691	Zhejiang	1969	Grain sown area	total grain	3120,6
25692	Zhejiang	1968	Grain production	total grain	9745
25693	Zhejiang	1968	Grain sown area	total grain	3039.2667
25694	Zhejiang	1967	Grain production	total grain	9365

Last Updated	7 years ago
Owner	Agriculture
Created	7 years ago
Size	2.86 MB

Displaying 6 columns, 25,703 rows in table

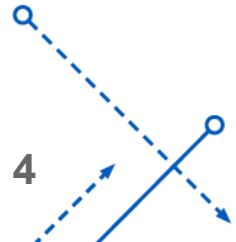
provincial_data

Grado di “naturale associazione”

Nella realtà, si hanno a disposizione un **grande numero di osservazioni** che sono praticamente intrattabili.

Potremmo però pensare di **classificarle in gruppi** per poter **considerare ogni gruppo** come **singola unità**.

In questo modo **«potremmo riuscire»** ad eseguire le nostre analisi in modo più agevole.

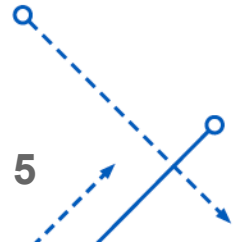


Introduzione al Clustering

Il clustering è un metodo che ci permette di **raggruppare un grande numero di osservazioni in categorie o gruppi caratterizzati (idealmente) da un'alta «naturale associazione»**.

Invece di analizzare ogni singola osservazione, possiamo trattare ogni gruppo come un'unità.

Nella pratica, cerchiamo di trovare somiglianze tra i dati per creare questi gruppi.



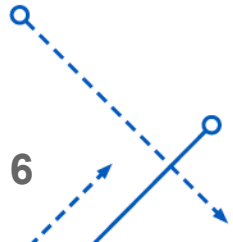
Introduzione dettagliata al Clustering

L'analisi dei cluster è una metodologia che permette di:

- raggruppare in sottoinsiemi (cluster) entità (unità) appartenenti ad un insieme più ampio;
- ottenere **raggruppamenti** in base alla **somiglianza** in modo che:
  - **gli elementi di uno stesso gruppo siano tra *«loro il più possibile simili»***
  - **gli elementi appartenenti a gruppi distinti siano *«tra loro il più possibile diversi».***

Nota: L'insieme originario delle entità su cui si attua l'analisi per ricavare i cluster **non è sottoposto ad alcuna restrizione**.

Può contenere variabili, individui, osservazioni, dati, misure ← **Che struttura dati di R vi ricorda?**

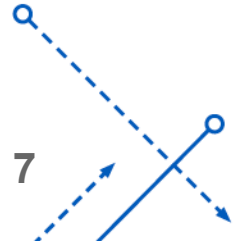


Scopo del clustering

Il clustering ha lo scopo di distribuire le osservazioni in gruppi, in modo tale che:

- il grado di ***naturale associazione*** sia alto tra i membri dello stesso gruppo …
- … e basso tra i membri di gruppi diversi.

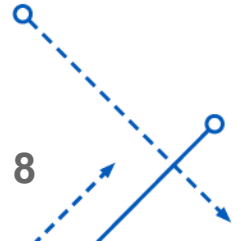
L'obiettivo è di ottenere quindi **un'alta omogeneità all'interno dei gruppi** e **un'alta eterogeneità tra** gruppi distinti.



Definizione del problema del Clustering

- Il problema dell'analisi dei cluster consiste nel determinare **m sottoinsiemi**, detti cluster, di individui in da un insieme di entità dimensione **n**, con **m** intero minore di **n**, tali che ogni **entità appartenga soltanto ad un unico sottoinsieme.**

- Gli individui che sono **assegnati allo stesso cluster** sono detti **simili** mentre gli individui che sono **assegnati a differenti cluster** sono detti **dissimili**.



Esempi di applicazione

- Informatica e l'ingegneria (per creare delle classificazioni o formare delle categorie);
- Scienze naturali (per affrontare problemi di tassonomia, per descrivere l'ecologia di comunità naturali);
- **La medicina (come ausilio nella diagnosi dei quadri clinici, nelle previsioni sulle malattie di individui o popolazioni e a scopo di diagnosi) – digital health!;**
- L'economia (per classificare regioni e identificare aree omogenee sulla base di particolari indici);
- L'archeologia, **la linguistica**, **le scienze sociali**.

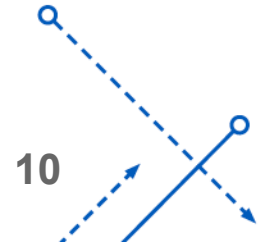
Nota: Esistono innumerevoli altri possibili casi d'uso, in blu riportiamo però quelli che si avvicinano ai dataset proposti.

Clustering e classificazione

Supponiamo di avere un database con informazioni su migliaia di clienti, come età, spesa annuale e frequenza degli acquisti.

Potremmo riuscire a clusterizzare i clienti nelle seguenti categorie:

- Clienti occasionali che comprano sporadicamente e spendono poco;
- Clienti fedeli che acquistano regolarmente e spendono di più;
- Clienti a rischio che prima erano clienti fedeli ma ora non comprano più.



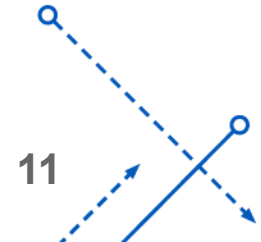
Clustering e Tassonomia

Pensate agli scienziati che raccolgono **dati su diverse specie di piante o animali**, come caratteristiche fisiche, habitat, e comportamenti.

Supponiamo di dover studiare una foresta.

Da ricercatori potremmo utilizzare il clustering per raggruppare le specie di piante in base a fattori come «Tipo di habitat» e «Tipologia di Foglie».

Ad esempio, piante che crescono in aree umide possono essere raggruppate insieme oppure piante con foglie simili (ad esempio, sempreverdi vs. decidue) possono formare un gruppo.



Tipologie di Clustering

- Possiamo dividere il clustering in due macro-tipologie: Supervisionato e non supervisionato.

- **Non Supervisionato:** L'algoritmo cerca di **identificare strutture nei dati** in modo autonomo e di raggruppare i dati in cluster in **base alla similarità**. Non si conoscono in anticipo i gruppi. (K-means, DBSCAN, Agglomerative Clustering). In genere utilizzato per esplorazione dei dati, analisi delle tendenze, riduzione della dimensionalità.

- **Supervisionato:** Creare modelli che possano prevedere le etichette per nuovi dati basati sulla similarità con i dati di addestramento. (K-Nearest Neighbors (KNN), Support Vector Machines (SVM)). In genere usato per Previsione, classificazione, riconoscimento di pattern.

Un po’ di AI (le reti SOM)

Le mappe **auto-organizzanti (SOM)** sono molto efficaci nella creazione di classificazioni.

Inoltre, le classificazioni mantengono informazioni topologiche su quali classi sono più simili tra loro.

Le mappe auto-organizzanti possono essere create con qualsiasi livello di dettaglio desiderato. Sono particolarmente adatte per il clustering di dati in molte dimensioni e con spazi delle caratteristiche complessi e connessi.

Chiaramente, tutto quello che fanno le reti SOM siamo in grado di farlo anche con tecniche di clustering non «black-box». Però, se la macchina può farlo al posto nostro, noi risparmiamo tempo.

Esempio di Tassonomia

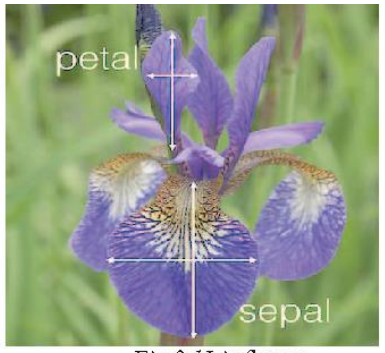
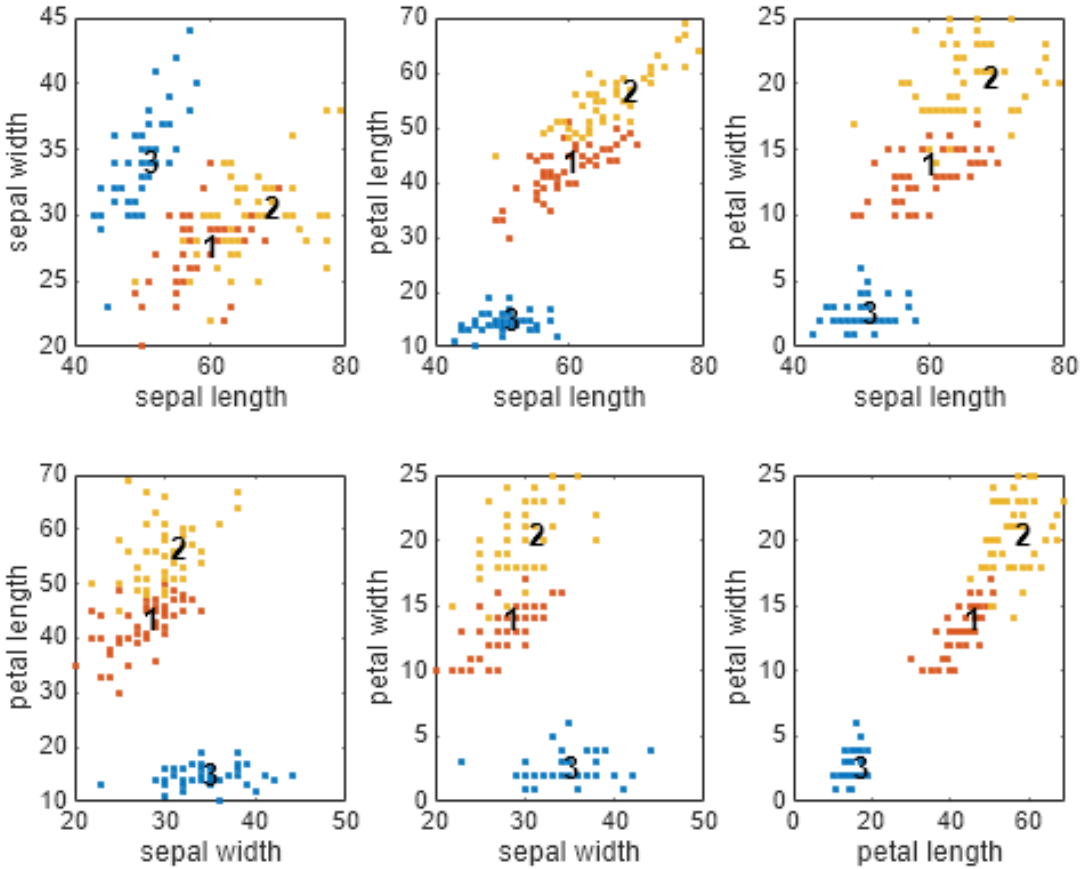


Fig.3.1Iris flower

Fig. 3.2 Three species of Iris flower

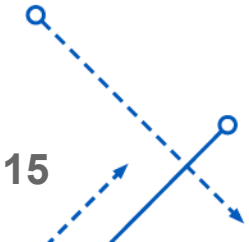


Clustering nella Medicina (Digital Health)

Un esempio di clustering in medicina è l'analisi dei dati dei pazienti per migliorare la diagnosi e la previsione delle malattie.

Ad esempio, utilizzando tecniche di clustering su un grande set di dati clinici, potremmo essere in grado di raggruppare i pazienti da diversi punti di vista:

- Pazienti con **sintomi simili** possono essere raggruppati per identificare pattern che possono suggerire una diagnosi comune.
- Si possono formare **gruppi di pazienti in base a caratteristiche (anagrafiche o di anamnesi)** come età, storia familiare e abitudini di vita, facilitando l'individuazione di fattori di rischio per determinate malattie.
- Andamento comune di parametri biometrici o clinici in risposta a stimoli esterni **(Il digital twin!)**



Clustering in Economia

Un esempio di clustering in economia è **l'analisi delle regioni per identificare aree omogenee** in base a diversi indici economici.

Ad esempio:

1. Le regioni possono essere classificate in base **al PIL pro capite** per identificare quelle più sviluppate rispetto a quelle meno sviluppate.
2. Raggruppando le regioni in base **ai tassi di disoccupazione,** è possibile identificare aree con problemi occupazionali simili.

Nota: nulla toglie che sia possibile utilizzare più indicatori in contemporanea, un po' come facciamo con la regressione multipla!

Clustering in Linguistica

Il clustering ha molte applicazioni in linguistica e nel Natural Language Processing (NLP).

Possiamo usarlo per **raggruppare documenti simili**, facilitando l'organizzazione di grandi raccolte di testi. Ad esempio, articoli di notizie o pubblicazioni accademiche possono essere **clusterizzati in base ai temi trattati.**

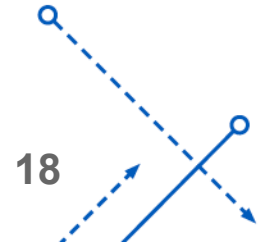
Il clustering può aiutare a identificare gruppi di frasi che esprimono **sentimenti simili (la sentiment analysis)**, consentendo una comprensione più profonda dello stato d'animo degli utenti.

Utilizzando tecniche di clustering, **è possibile identificare automaticamente argomenti ricorrenti** in un corpus di testi, facilitando la classificazione e l'individuazione di temi principali.

Clustering in Linguistica /2

Word Embeddings: Nel contesto di rappresentazioni vettoriali delle parole, il clustering può essere utilizzato per identificare gruppi di parole simili, permettendo di esplorare relazioni semantiche tra di esse.

Creazione di chatbots: Nel design di chatbot, il clustering può aiutare a raggruppare domande simili, migliorando la comprensione delle richieste degli utenti e facilitando risposte più accurate.



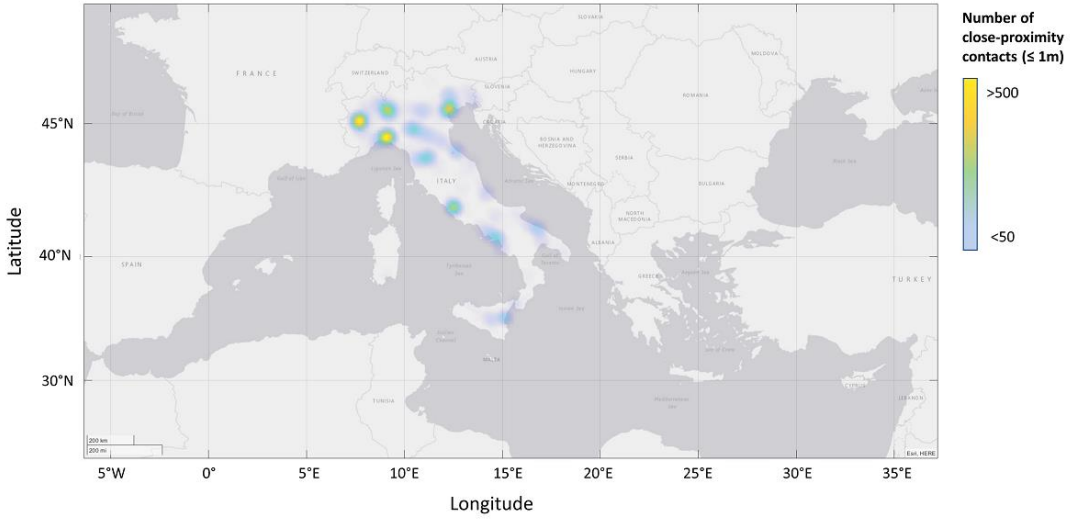
Esempio reale d’uso

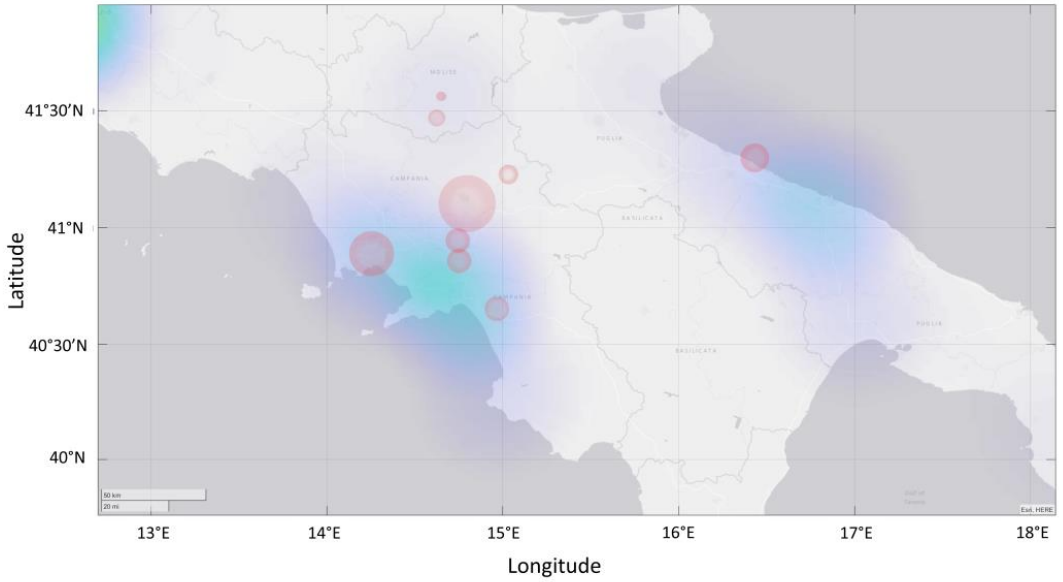
Con un semplice «word count» e applicando una semplice tecnica di clustering è possibile suddividere un insieme di libri per argomento e autore. (Svolgimento alla lavagna)

Statistica e analisi dei dati

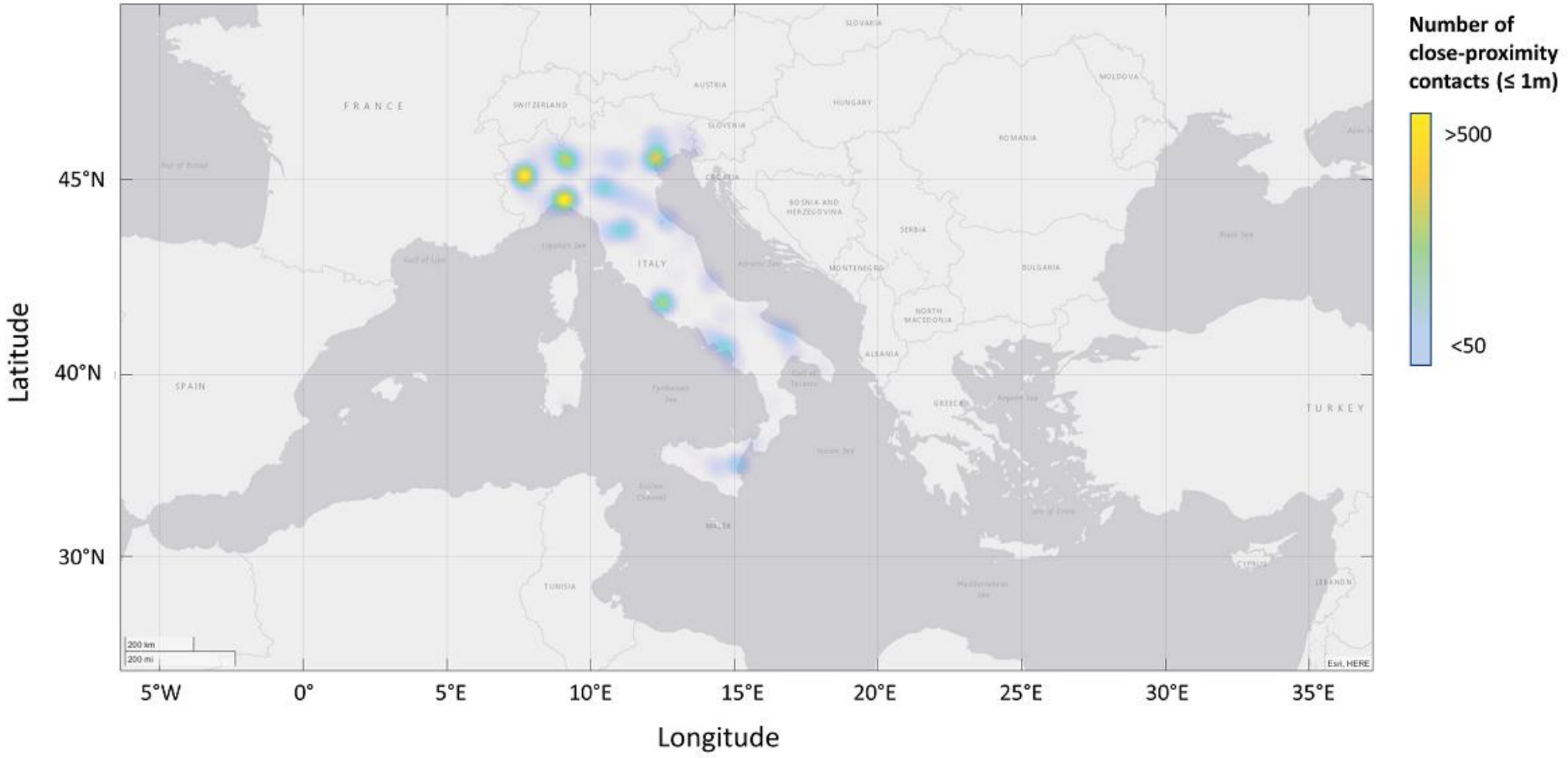
Clustering nelle scienze sociali

Un esempio di clustering nelle scienze sociali è l'analisi dei dati demografici per comprendere le dinamiche delle popolazioni. Un esempio eclatante lo abbiamo visto durante la pandemia del Sar-Cov-2.

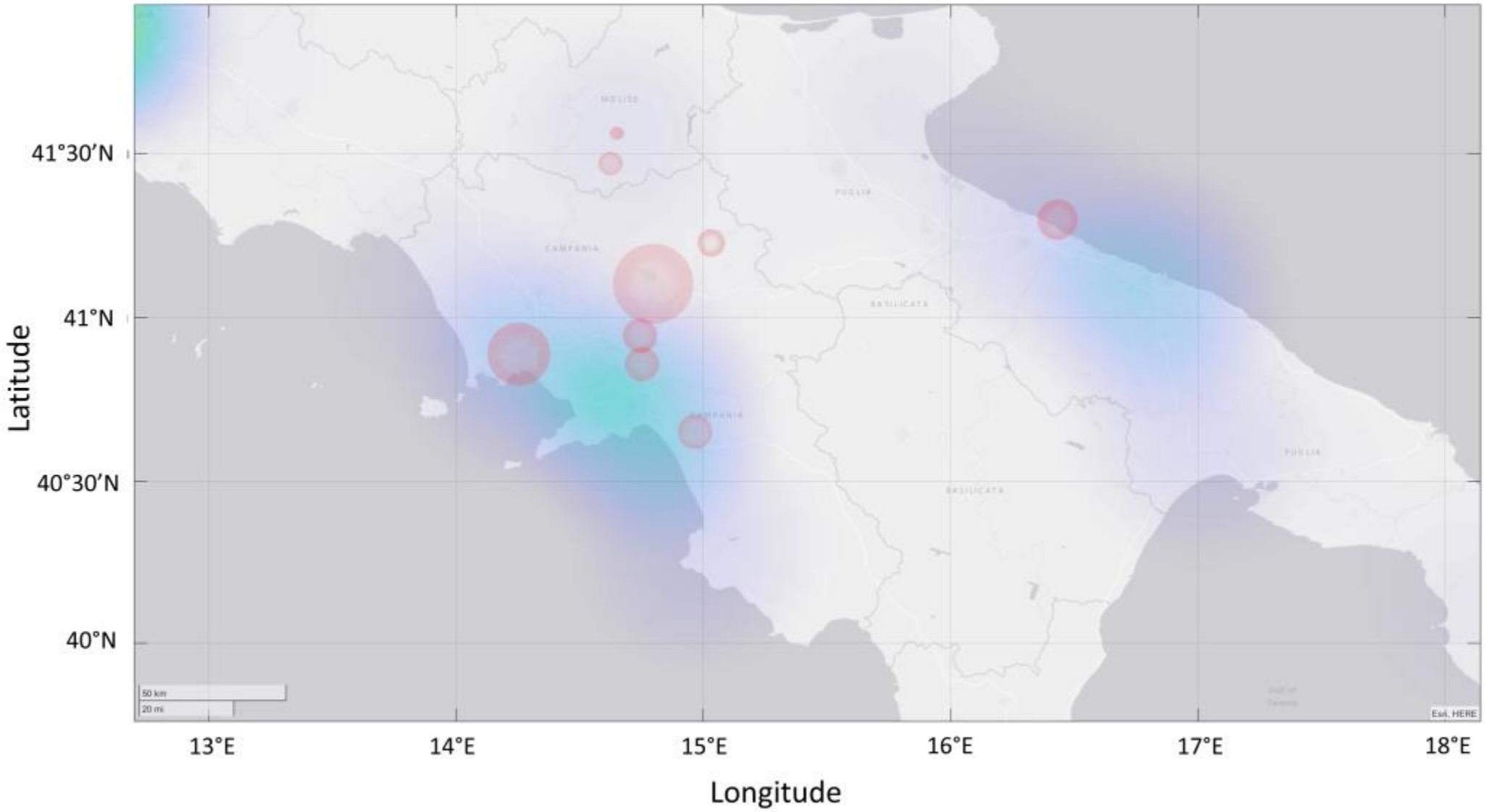




Statistica e analisi dei dati

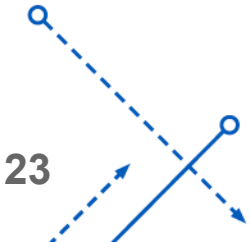


Statistica e analisi dei dati



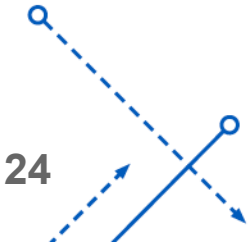
Obiettivi raggiungibili

- Individuazione di una reale tipologia;
- Previsioni basate su gruppi;
- Esplorazione dei dati;
- **Generazione di ipotesi di ricerca**;
- Verifica di ipotesi;
- Riduzione della complessità dei dati (PCA).



Ipotesi e verifica di ipotesi

- L'analisi dei cluster può anche essere usata per **generare ipotesi sulla natura dei dati**.
  - In questo caso occorre successivamente verificare le ipotesi fatte e ogni test deve essere effettuato su nuove osservazioni senza usare i dati da cui le ipotesi sono state generate.
- In alcuni campi di indagine scientifica si possono usare i metodi dell'analisi dei cluster per produrre gruppi che formano la base di uno schema di classificazione utile in studi successivi per scopi di previsioni di un qualche tipo.
  - Esempio ABCDE nella classificazione skin cancer.



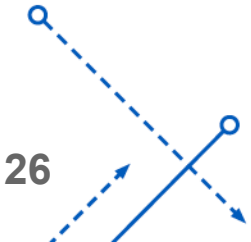
Esempio ABCDE

- Il metodo dermatologico ABCDE è un approccio utile per classificare le lesioni cutanee e identificare i segni del cancro della pelle.
- Iniziamo raggruppando le lesioni cutanee in base a caratteristiche comuni, utilizzando il metodo ABCDE: A (Asimmetria) B (Bordi) C (Colore) D (Diametro) … la E per semplicità la ignoriamo in questo esempio.
- Dall'analisi dei cluster, potremmo formulare ipotesi sulla relazione tra queste caratteristiche e il rischio di cancro della pelle.
    - Ad esempio:
        - RQ1 "Le lesioni asimmetriche (A) tendono a essere associate a un rischio più elevato di melanoma rispetto a quelle simmetriche."
        - "Lesioni con bordi irregolari (B) mostrano una maggiore incidenza di diagnosi di cancro della pelle rispetto a quelle con bordi regolari."

Responsabilità dell'analista (vostra!) /1

- E importante osservare **che la scelta delle variabili** (delle caratteristiche da osservare) è strettamente condizionata **allo scopo dell'indagine** e **presuppone** l'esistenza, seppure ad uno stato iniziale, di un modello logico.

- **La scelta delle variabili effettuata senza alcun ricorso a criteri matematici** o statistici e riflette il **giudizio dell'analista (VOSTRO)** sull'importanza delle proprietà utili a descrivere il fenomeno in funzione del quale deve essere svolta l'analisi dei cluster.



Responsabilità dell'analista (vostra)! /2

Esempio 1

Si desidera effettuare l'analisi dei cluster sui comuni di una regione in funzione dell'ambiente socio–economico, si dovranno prendere in considerazione i parametri rilevati in ciascun comune che:

- descrivono la situazione sanitaria,
- la situazione demografica,
- l'occupazione;

Responsabilità dell'analista (vostra)! /2

Esempio 2

Si desidera effettuare l'analisi dei cluster sui comuni di una regione in funzione del comportamento elettorale.

Si dovranno prendere in considerazione i parametri rilevati che descrivono le percentuali di voti conseguite dai diversi partiti politici in ciascun comune.

Responsabilità dell'analista (vostra) ! /3

Uno dei problemi che si presenta nell'analisi dei cluster riguarda **la standardizzazione o meno** delle variabili poiché attribuire un peso diverso a ciascuna caratteristica potrebbe condurre a risultati differenti circa la classificazione a seconda delle tecniche di clustering utilizzate.

In molti metodi di clustering si raccomanda la standardizzazione di ogni variabile (caratteristica) usando la media campionaria e la deviazione standard **campionaria entrambe derivate dall'insieme completo di individui della popolazione.**

Responsabilità dell'analista (vostra) ! /4

Esempio 1 (scelta delle variabili)
Si desidera effettuare l'analisi dei **cluster sui comuni di una regione** in funzione dell'ambiente socio–economico:

- **è preferibile utilizzare misure in forma standardizzata** poiché potrebbe risultare difficile stabilire se parametri relativi all'occupazione debbano avere un peso maggiore o minore di parametri relativi alla situazione demografica;

Esempio 2 (scelta delle variabili)

- Si desidera effettuare l'analisi dei cluster **sui comuni di una regione in funzione del comportamento elettorale:**
  - è preferibile utilizzare **pesi proporzionali alle percentuali medie dei suffragi ottenuti da ciascun partito politico**, ossia delle misure non standardizzate.

Ridurre la complessità (idea di base) /1

- Nella realtà, si hanno a disposizione un **grande numero di osservazioni** che sono praticamente intrattabili.
- Relativamente al fenomeno osservato, il numero delle caratteristiche misurate per ogni individuo può essere grande e **non sempre è necessario utilizzarle tutte!**

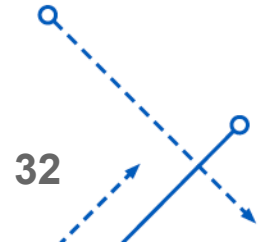
IMPORTANTISSIMO

- Includere variabili con un **piccolo potere discriminante** potrebbe **appiattire le differenze tra gruppi**.
- Includere variabili con un **elevatissimo potere discriminante** può rendere **inutile** l'inclusione di altre variabili strettamente collegate al fenomeno analizzato dal punto di vista logico.

Basso potere discriminante

- Supponiamo di voler segmentare i clienti di un negozio di abbigliamento in base a variabili come età, reddito, stile di vita e preferenze di acquisto.

- Se **includiamo variabili con un basso potere discriminante**, come il colore preferito o la marca di abbigliamento preferita, **potremmo non notare differenze significative tra i gruppi**. Queste variabili potrebbero non contribuire a definire chiaramente le caratteristiche dei clienti, appiattendo le differenze e portando a segmenti poco distintivi.



Alto potere discriminante

- Se **includiamo variabili con un elevato potere discriminante**, come il reddito annuale o il comportamento di acquisto (frequenza di acquisto, importo medio speso), **potremmo ottenere gruppi molto distinti**.

- Tuttavia, l'inclusione di queste variabili potrebbe rendere meno rilevante l'inclusione di altre variabili, come le preferenze di stile di vita, che potrebbero avere una connessione logica ma non aggiungono valore significativo all'analisi.

Ridurre la complessità (idea di base) /2

- Poichè in molte tecniche di clustering il **tempo di calcolo cresce drammaticamente con il crescere del numero delle variabili**, prima di utilizzare tale analisi è utile ridurre il numero di variabili a quelle più direttamente collegate al fenomeno in esame.
- Un metodo che permette di effettuare tale riduzione delle variabili originarie è **l'analisi delle componenti principali (PCA)**.
  - Le tecniche di clustering possono essere applicate alle prime q componenti principali (q < p) che possono essere considerate come nuove variabili di input.

Ridurre la complessità (idea di base) /3

- Trasformare le informazioni dall'insieme completo di **n** individui all'informazione circa **m** gruppi di individui (ovviamente **m** deve essere molto più piccolo di **n**).
- In questo modo è possibile fornire una più concisa e più comprensibile descrizione delle osservazioni considerate.
- In altre parole, occorre ricercare delle semplificazioni **del problema originario con la minima perdita di informazione.**
- **Risparmiare tempo di calcolo!**

Ridurre la complessità (idea di base) /4

- Ovviamente, l'analisi delle componenti principali permette di ridurre il numero delle variabili ma non risolve il problema della standardizzazione e della correlazione delle variabili.
- Inoltre, le variabili di input determinano la classificazione ed è **possibile determinare una diversa classificazione utilizzando le prime q componenti principali invece che tutte le caratteristiche iniziali.**
- Nel caso di dati ben strutturati (gruppi ben differenziati) le differenze tra le due classificazioni sono piccole, ma nel caso di gruppi non ben differenziati, possono essere riscontrate delle grandi differenze

Definizioni di base

- Sia $I = \{I_1, I_2, \ldots, I_n\}$ un insieme di **n** individui appartenenti ad una popolazione.
- Assumiamo che esista un insieme di caratteristiche (features) $C = \{C_1, C_2, \ldots, C_p\}$ che sono osservabili e sono possedute da ogni individuo in **I**.
- Denotiamo con il simbolo x_{ij} il valore della misura della caratteristica j–esima relativa all'individuo I_i e con $X_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ $(i = 1, 2, \ldots, n)$ il vettore di cardinalità $1 \times p$ di tali misure.

$$X = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix},$$

$$X_i = (x_{i1}, x_{i2}, \ldots, x_{ip}) \ (i = 1, 2, \ldots, n).$$

Verosomiglianza tra caratteristiche /1

- Possiamo costruire una matrice (o tabella, o data-frame) di tutti i dati osservati per ogni singolo individuo, per ogni singola caratteristica osservabile.

$$
\begin{array}{cccccc}
x_{11} & x_{12} & \dots & x_{1j} & \dots & x_{1p} \\
x_{21} & x_{22} & \dots & x_{2j} & \dots & x_{2p} \\
\vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
x_{i1} & x_{i2} & \dots & x_{ij} & \dots & x_{ip} \\
\vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
x_{n1} & x_{n2} & \dots & x_{nj} & \dots & x_{np}
\end{array}
$$

Verosomiglianza tra caratteristiche /1

Clustering /1

L'analisi dei cluster è una metodologia che permette di:

- raggruppare in sottoinsiemi (cluster) entità (unità) appartenenti ad un insieme più ampio;
- ottenere **raggruppamenti** in base alla **somiglianza** in modo che:
  - **gli elementi di uno stesso gruppo siano tra *«loro il più possibile simili»***
  - **gli elementi appartenenti a gruppi distinti siano *«tra loro il più possibile diversi».***

Verosomiglianza tra caratteristiche /2

- Con gli strumenti a nostra disposizione, possiamo già essere in grado di «tirar fuori» qualche informazione interessante relativamente alla tabella delle osservazioni.
- Siamo in grado infatti di procedere con l'analisi esplorativa della nostra tabella così come faremmo normalmente.

	x_{11}	x_{12}	$\dots$	x_{1j}	$\dots$	x_{1p}
	x_{21}	x_{22}	$\dots$	x_{2j}	$\dots$	x_{2p}
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
	x_{i1}	x_{i2}	$\dots$	x_{ij}	$\dots$	x_{ip}
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
	x_{n1}	x_{n2}	$\dots$	x_{nj}	$\dots$	x_{np}
mean	$\overline{x}_1$	$\overline{x}_2$	$\dots$	$\overline{x}_j$	$\dots$	$\overline{x}_p$
var	s_1^2	s_2^2	$\dots$	s_j^2	$\dots$	s_p^2

$$\overline{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij},$$

$$s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_{ij} - \overline{x}_j\right)^2$$

Distanza e Similarità /1

Per risolvere il problema di clustering:

- Ci servono **strumenti** per capire **«chi si appaia con chi»** e in quale cluster;
- Dobbiamo **definire i termini somiglianza o differenza** in modo quantitativo;
- Occorre precisare cosa significa la **somiglianza di due individui I_i e I_j assegnati allo stesso cluster**;
- Occorre precisare cosa significa la **differenza di due individui assegnati a differenti cluster**.

Distanza e Similarità /2

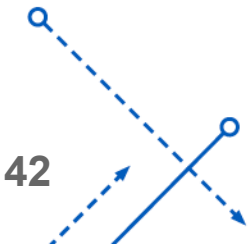
Dati due individui I_i e I_j $(i \neq j)$.

Cosa significa **somiglianza** di due individui $\mathbf{I_i}$ e $\mathbf{I_j}$ assegnati allo stesso cluster?

$coefficiente\ di\ similarità\ s_{ij} = s(X_i, X_j)$ ← **E' quanto più vicino al massimo possibile.** Spostare $\mathbf{X_i}$ o $\mathbf{X_j}$ in un altro cluster ridurrebbe questo valore!

Cosa significa **differenza** di due individui assegnati a differenti cluster?

$d_{ij} = d(X_i, X_j)$ ← **E' massima!** Spostare $\mathbf{X_i}$ o $\mathbf{X_j}$ in un altro cluster ridurrebbe questo valore e sarebbe possibile trovare un altro cluster nel quale questa distanza sia più grande.



Proprietà importanti di s_{ij} e d_{ij}

$$s_{ij} = s(X_i, X_j)$$

Il **coefficiente** di **similarità** assume valori **nell'intervallo** [0,1]

$$d_{ij} = d(X_i, X_j)$$

Le misure di distanza possono assumere **qualsiasi valore reale maggiore o uguale a zero**.

Importante: è possibile clusterizzare usando il coefficiente di similarità oppure la funzione di distanza. Entrambi sono indicatori che ci dicono «chi è simile a chi» e «chi dovrebbe appaiarsi con chi».

Clustering “naïve”

Un criterio per risolvere il problema di clustering potrebbe essere quello di assegnare due individui **Ii** e **Ij** ($i \neq j$):

- allo stesso cluster se il coefficiente di similarità tra i punti $\mathbf{X}_i$ e $\mathbf{X}_j$ è prossimo ad 1 **oppure** se la distanza tra i punti $\mathbf{X}_i$ e $\mathbf{X}_j$ è sufficientemente piccolo;

- a differenti cluster se il coefficiente di similarità tra i punti è prossimo ad 0 **oppure** se la distanza tra i punti è sufficientemente grande.

Ricordate i melanomi?

Come calcoliamo s_{ij} d_{ij}

Funzione di distanza /1

Le misure metriche di somiglianza sono soprattutto basate sulle funzioni distanza tra i vettori delle caratteristiche. Occorre quindi definire tale funzione.

Una funzione a valori reali d(Xi, Xj) `e detta funzione distanza **se e soltanto se essa soddisfa** le seguenti condizioni:

$$(i)\ d(X_i, X_j) = 0\ se\ e\ solo\ se\ X_i = X_j,\ con\ X_i\ e\ X_j\ in\ E_p;$$

La distanza tra un elemento e se stesso è zero.

$$(ii)\ d(X_i, X_j) \geq 0\ per\ ogni\ X_i\ e\ X_j\ in\ E_p;$$

La distanza è una funzione NON negativa

$$(iii)\ d(X_i, X_j) = d(X_j, X_i)\ per\ ogni\ X_i\ e\ X_j\ in\ E_p;$$

La distanza tra Xi e Xj è simmetrica (e la stessa tra Xj e Xi)

$$(iv)\ d(X_i, X_j) \leq d(X_i, X_k) + d(X_k, X_j)\ per\ ogni\ X_i,\ X_j\ e\ X_k\ in\ E_p.$$

Funzione di distanza /2

$$(iv)\ d(X_i, X_j) \leq d(X_i, X_k) + d(X_k, X_j)\ \textit{per ogni}\ X_i,\ X_j\ \textit{e}\ X_k\ \textit{in}\ E_p.$$

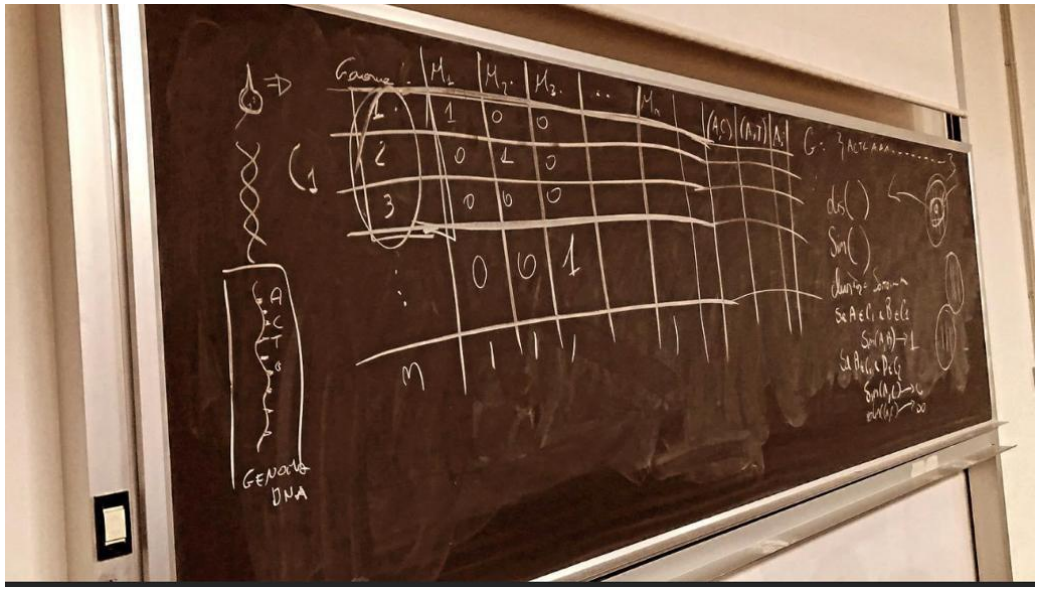
Diseguaglianza triangolare

La distanza tra Xi e Xj deve essere sempre minore o uguale della somma delle distanze di ognuno dei due vettori considerati da qualunque altro terzo vettore Xk.

In parole povere: se ci sta «un vettore in mezzo» tra Xi e Xj, allora quel vettore deve essere o nullo (0) oppure darà un contributo e aumenterà la distanza tra Xi e Xj

Distance Matrix (matrice delle distanze)

Negli esempi visti alla lavagna abbiamo avuto modo di apprezzare la forma di una **matrice di similarità.**



$$D = \begin{pmatrix} 0 & d_{12} & \dots & d_{1n} \\ d_{21} & 0 & \dots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \dots & 0 \end{pmatrix},$$

La matrice delle distanze contiene i valori di dij per le varie coppie anziché sij

Chiarimento!

$$X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix},$$

$$D = \begin{pmatrix} 0 & d_{12} & \dots & d_{1n} \\ d_{21} & 0 & \dots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \dots & 0 \end{pmatrix},$$

	x_{11}	x_{12}	$\dots$	x_{1j}	$\dots$	x_{1p}
	x_{21}	x_{22}	$\dots$	x_{2j}	$\dots$	x_{2p}
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
	x_{i1}	x_{i2}	$\dots$	x_{ij}	$\dots$	x_{ip}
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
	x_{n1}	x_{n2}	$\dots$	x_{nj}	$\dots$	x_{np}
mean	$\overline{x}_1$	$\overline{x}_2$	$\dots$	$\overline{x}_j$	$\dots$	$\overline{x}_p$
var	s_1^2	s_2^2	$\dots$	s_j^2	$\dots$	s_p^2

Distance Matrix (matrice delle distanze) /1

Le **distanze** tra tutte le possibili coppie di unità sono inserite in **una matrice simmetrica D** di cardinalità n × n, ossia

$$D = \begin{pmatrix} 0 & d_{12} & \dots & d_{1n} \\ d_{21} & 0 & \dots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \dots & 0 \end{pmatrix},$$

$$d_{ij} = d(X_i, X_j)$$

$(i)\ d(X_i, X_j) = 0\ se\ e\ solo\ se\ X_i = X_j,\ con\ X_i\ e\ X_j\ in\ E_p;$

I termini sulla diagonale principale sono tutti uguali a zero mentre

$(iii)\ d(X_i, X_j) = d(X_j, X_i)\ per\ ogni\ X_i\ e\ X_j\ in\ E_p;$

I termini simmetrici sono uguali a due a due

Distance Matrix (matrice delle distanze) /2

$$D = \begin{pmatrix} 0 & d_{12} & \dots & d_{1n} \\ d_{21} & 0 & \dots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \dots & 0 \end{pmatrix}, \quad d_{ij} = d(X_i, X_j)$$

OTTIMIZZAZIONE INFORMATICA!

E sufficiente considerare la matrice triangolare al di sopra o al di sotto della diagonale principale di D.
Risparmiamo RAM!

Distance Matrix (matrice delle distanze)

Non esiste una sola funzione distanza, ma esiste un'intera famiglia di funzioni che rispettano almeno le quattro proprietà precedenti. Abbiamo inoltre che:

(a) se d e d' sono due metriche anche $d + d'$ è una metrica;

(b) se d è una metrica e c un numero reale positivo allora anche $c\,d$ è una metrica;

(c) se d è una metrica e c un numero reale positivo allora anche $d' = d/(c + d)$ è una metrica.

ATTENZIONE: Il prodotto di due metriche (in particolare il quadrato di una metrica) **non necessariamente soddisfa la disuguaglianza triangolare** e quindi può non essere una misura di distanza.