METODI STATISTICI PER L'ANALISI DI DATI AMBIENTALI

Esempi di applicazione a studi di ecologia del paesaggio

planland ©®
studio tecnico daniel franco

Riferimenti

- Fabbris L., 1997. Statistica multivariata, analisi esplorativa dei dati. McGraw – Hill.
- Legendre P., legendre L., 2004.Numerical Ecology. Elsevier.
- www.dsa.unipr.it/soliani/soliani.html

Perché l'analisi statistica

" A un pensiero che isola e separa si dovrebbe sostituire un pensiero che distingue ed unisce. A un pensiero disgiuntivo e riduttivo occorrerebbe sostituire un pensiero del complesso, nel senso originario del termine complexus: ciò che è tessuto insieme."

Edgar Morin

Perché l'analisi statistica

- Complessità dei fenomeni osservati
- Multidimensionalità dei fenomeni reali

Eliminare:

1. L'informazione inutile
2. L'informazione ridondante
3. Rumore

Perché l'analisi statistica

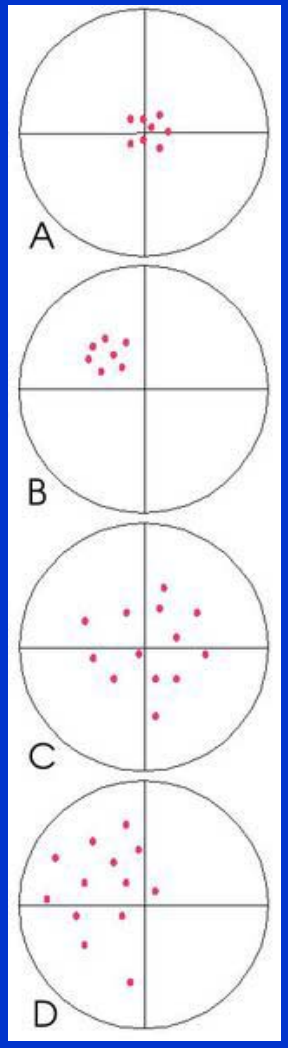
Consente di esplorare i dati sperimentali, al fine di

- descrivere in forma sintetica le informazioni
- evidenziare le strutture di relazione implicite (*pattern*) che li percorrono
- ricavare modelli interpretativi/predittivi della realtà.

Contenuti

- Terminologia
- Concetti di base
- Metodi statistici
  - Confronto set di dati sperimentali
  - Analisi esplorativa di dati multidimensionali
- Interpretazione dei risultati

Terminologia



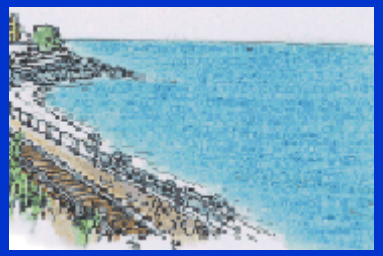
Il dato statistico:

E' l'elemento di base di ogni ricerca ed è quindi della massima importanza che sia "affidabile".

Concetti di

- Precisione: (vicinanza di misure ripetute al medesimo valore, esprimibile in termini di deviazione standard)
- Accuratezza: (vicinanza di un valore misurato al suo valore reale)

Terminologia



- *Popolazione statistica:* insieme delle unità oggetto di indagine, di cui si considera un
- *Campione*: frazione della popolazione statistica effettivamente osservato.
- *Variabili:* caratteristiche che si osservano in una popolazione statistica. Possono essere
  - Qualitative o categoriche: espresse da numeri interi e discreti
  - Quantitative: espresse su scala continua

Terminologia

- *Scala* di misura delle variabili:
  - **Nominale** o classificatoria: appartenenza a categorie qualitative; solo relazioni di identità-diversità**.** Caso particolare: variabili dicotomiche. Conteggio, frequenza, confronti tra frequenze.
  - **Ordinale** o per ranghi: scala monotonica, appartenenza a classi ordinate; non è possibile quantificare la distanza tra le classi (colore, giudizi,età).
  - ad **intervalli**: scala monotonica, aggiunge la possibilità di quantificare la distanza tra classi (temperatura, calendario).
  - di **rapporti**: scala con origine reale (0 significa quantità nulla). Hanno significato anche i rapporti tra le misure.

Contenuti

- Terminologia
- Concetti di base
- Metodi statistici
  - Confronto set di dati sperimentali
  - Analisi esplorativa di dati multidimensionali
- Interpretazione dei risultati

Distribuzione di frequenza
Media
Varianza
Deviazione standard
Significatività statistica

Concetti di base

- L'operazione più immediata su una serie di dati è la distribuzione ordinata di tutti i valori *(seriazione)*
- Successivamente la serie può essere raggruppata in *classi*
- Si ottiene una *distribuzione di frequenza o di intensità*

0	0	3	2	1	2	3
2	0	3	3	1	1	3
3	1	3	2	3	1	0

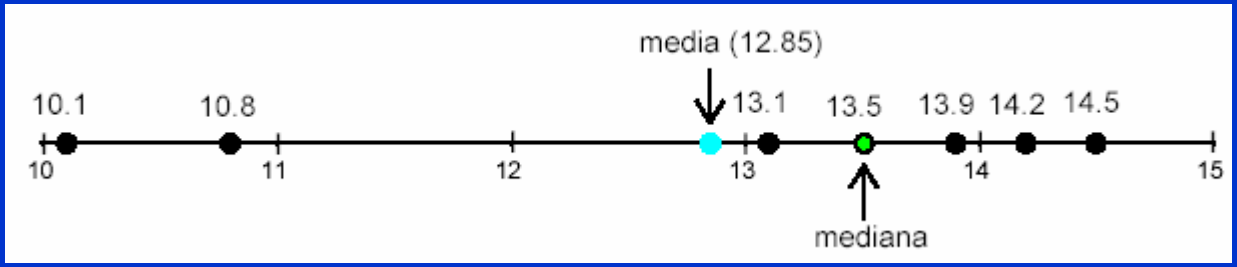
Classe	0	1	2	3
Freq. Assoluta	4	5	4	8
Freq. Relativa	0.19	0.24	0.19	0.38
Freq. cumulata	0.19	0.43	0.62	1.00

Concetti di base

- ***Tendenza centrale:*** valore attorno a cui sono raggruppati i dati

Ordine	Valore
1	10,1
2	10,8
3	13,1
4	13,9
5	14,2
6	14,5

- **Media aritmetica**: somma dei dati / numero delle osservazioni
- **Mediana**: valore che occupa la posizione centrale in un insieme ordinato di dati (N+1/2 per N dispari, media tra N/2 e il successivo se N pari). Rappresenta meglio il valore "tipico" di insieme di dati non simmetrici; non è influenzata dai valori estremi.
- **Moda**: valore più frequente di una distribuzione. Poco utilizzata perché meno stabile e oggettiva.



Concetti di base

- ***Varianza e deviazione standard***
  - **Devianza o scarto** (*deviate)* → X_i — **Xm**
    di quanto un elemento differisce dalla media
  - **Somma del quadrato degli scarti** (*sum of squared deviates)* **SS** = $(X_i$ — **Xm**$)^2$
  - **Varianza** (*variance)* σ^2→ media degli SS

    $\sigma^2 = \Sigma\ (X_i$ — **Xm**$)^2/N$
  - **Deviazione standard o scarto quadratico medio** (*standard deviation)* σ → radice della media degli SS

    σ = sqrt[$\Sigma\ (X_i$ — **Xm**$)^2/N$] = sqrt[varianza]

Concetti di base

- ***Ipotesi nulla***
  - ***Nel test statistico si verifica in termini probabilistici la validità di un'ipotesi statistica, detta appunto ipotesi nulla, di solito indicata con H_0***
  - ***Nel test di ipotesi è data dall'uguaglianza tra due quantità***

Concetti di base

- ***Significatività statistica: il p-level***
- Significatività statistica di un risultato = stima della misura del grado di "verità", rappresentatività della popolazione statistica.
- *p-level* = probabilità di errore connessa nell'accettare i risultati osservati come validi, ovvero rappresentativi della realtà.
  - In molte aree di ricerca il limite $p<.05$ (errore 5%) è considerato un livello accettabile di errore.
- Maggiore il valore del *p-level*, minore la probabilità che la relazione osservata tra le variabili del campione sia un indicatore rilevante della relazione tra le rispettive variabili nella popolazione statistica.

Contenuti

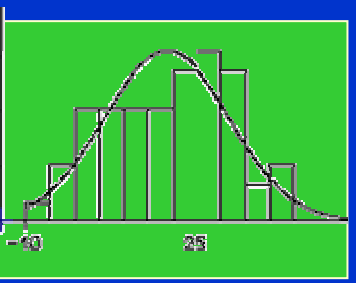
- Terminologia
- Concetti di base
- Metodi statistici
  - Confronto set di dati sperimentali
  - Analisi esplorativa di dati multidimensionali
- Interpretazione dei risultati

Metodi statistici

- I metodi statistici si distinguono in relazione
- alle caratteristiche dei dati – se permettono o meno il ricorso alla distribuzione normale:
  - Statistica parametrica
  - Statistica non parametrica
- al numero di variabili:
  - statistica univariata - bivariata,
  - statistica multivariata.

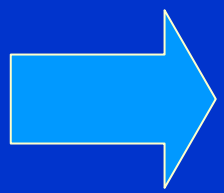
Metodi parametrici e non

- **Metodi parametrici**
- Prevedono l'esistenza di assunti :
  - 1. Indipendenza dei campioni: generati per estrazione casuale da una popolazione in cui ogni soggetto abbia la stessa probabilità di essere incluso in un gruppo qualsiasi. I fattori aleatori risultano casualmente distribuiti e non generano distorsioni o errori sistematici.
  - 2. Normalità della distribuzione: la non-normalità della distribuzione delle medie è indice di un'estrazione non casuale.
  - 3. Omogeneità delle varianze: se formati per estrazione casuale dalla medesima popolazione, i vari gruppi devono avere varianze eguali.
- Quando almeno uno dei presupposti non è rispettato, neppure dopo appropriata trasformazione dei dati, la validità dei risultati è in dubbio.



Metodi parametrici e non

- **Metodi non parametrici**
  - Non impongono vincoli ai dati: si utilizzano quando la forma della distribuzione dei dati è ignota, oppure non rispetta gli assunti precedenti.
  - Nella ricerca sperimentale tale condizione è frequente, spesso perché si dispone di un numero di dati insufficiente per dimostrare la normalità della distribuzione



- La soluzione adottata sempre più frequentemente consiste nell'utilizzare due metodi analoghi appartenenti alle due categorie: così è possibile verificare quanto l'allontanamento dagli assunti parametrici pesa sul risultato.

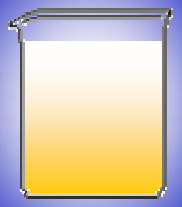
Contenuti

- Terminologia
- Concetti di base
- Metodi statistici
  - Confronto set di dati sperimentali
  - Analisi esplorativa di dati multidimensionali
- Interpretazione dei risultati

Contenuti – metodi statistici

- Confronto set di dati sperimentali
  - Test per 2 campioni indipendenti
    - Parametrico: t – test
    - Equivalenti non parametrici
  - Test per 2 campioni dipendenti
    - Parametrico: t –test
    - Equivalenti non parametrici
  - Analisi della varianza: test per più campioni
    - Parametrico: ANOVA
    - Equivalenti non parametrici

Confronto set di dati sperimentali



- ***t - test per campioni indipendenti***
- Metodo comunemente utilizzato per valutare la significatività delle differenze tra le medie di due gruppi di osservazioni indipendenti:
  - Media del campione e media attesa (o della popolazione);
  - Singolo dato e media del campione;
  - Medie di due campioni indipendenti.
- Metodo parametrico: i campioni devono avere distribuzione normale e la scala di misura deve essere equispaziata.

Confronto set di dati sperimentali

- ***t - test per campioni indipendenti* - Esempi di applicazione**
- Media del campione e media attesa (o di popolazione);
  - Confronto tra i risultati dell'esperimento e i dati di letteratura
- Singolo dato e media del campione;
  - Es: confronto accrescimento pianta dopo sversamento su terreno di sostanza inquinante e valori accrescimento medi precedentemente stimati;
- Medie di due campioni indipendenti (caso tipico).
  - 2 gruppi soggetti a diversa condizione sperimentale = verificare quanto differiscono a seguito del trattamento.
  - Confronto tra l'esito di un esperimento e un "bianco".

Confronto set di dati sperimentali

- ***t - test per campioni indipendenti***
- A due code: si chiede se Ma ≠ Mb;
  A una coda: si chiede se Ma > Mb, escludendo già che sia minore (es. sversamento tossico vs. vegetazione).
- Il test si basa sulla differenza tra le medie:
  - Se i due gruppi non differiscono, la differenza non sarà significativamente diversa da zero.
  - Se si rileva una differenza, si deve provare che sia data da una vera diversità di comportamento.
- $t = (Xm_A - Xm_B)/\sigma_M$
  differenza tra medie/dev.std. media pop
  - Tavole: valori critici di t a seconda del livello di precisione richiesto (*p*-level).

Confronto set di dati sperimentali

non parametrico

- ***Equivalenti non parametrici:***
- **Wald-Wolfowitz runs test**
  - Testa l'ipotesi che i due campioni siano tratti da due popolazioni diverse per qualche carattere (non solo la media, anche la forma generale della popolazione).
- **Mann Whitney U Test**
  - È l'alternativa più sensibile al *t*-test, ed ha la stessa interpretazione.
- **Kolmogorov-Smirnov Two-Sample Test**
  - Simile nell'interpretazione al primo: tiene conto delle differenze globali tra popolazioni.

Confronto set di dati sperimentali

- ***t - test per campioni dipendenti***
  - Per la significatività delle differenze tra le medie di due gruppi di campioni dipendenti o correlati.
- La relazione tra due variabili dipende in larga parte dalla variabilità interna ai due gruppi:
  - Se la differenza tra le medie è maggiore di tale variabilità, allora è significativa.
  - La variabilità interna non serve ai fini del test perché è estranea a ciò che si vuole provare.
- Nel caso di gruppi di campioni correlati, la variabilità interna è maggiore della differenza fra le medie;
- Il test per campioni dipendenti serve a ignorare la variabilità intrinseca delle due popolazioni e a concentrarsi su quella piccola differenza che si vuole indagare.

Confronto set di dati sperimentali

- ***t - test per campioni dipendenti***
- Esempi di applicazione
- Due gruppi di osservazioni ricavate da uno stesso campione in due condizioni differenti
- Osservazione di una variabile
  - prima e dopo un trattamento;
  - in due stagioni diverse;
  - A due diverse condizioni di temperatura, pH...

Confronto set di dati sperimentali

- ***Equivalente non parametrico:***
- ***Sign test***
  - È l'equivalente del t-test per campioni dipendenti, ed ha la stessa interpretazione.
- ***Wilcoxon Matched Pairs Test***
  - Richiede che siano su scala ordinale anche le differenze tra variabili;
  - Se applicato nelle stesse condizioni del t-test, è più sensibile.

non parametrico

Contenuti – metodi statistici

- Confronto set di dati sperimentali
  - Test per 2 campioni indipendenti
    - Parametrico: t – test
    - Equivalenti non parametrici
  - Test per 2 campioni dipendenti
    - Parametrico: t –test
    - Equivalenti non parametrici
  - Analisi della varianza: test per più campioni
    - Parametrico: ANOVA
    - Equivalenti non parametrici

Confronto set di dati sperimentali

- ***ANOVA a una via per campioni indipendenti.***
  - Analisi parametrica
  - Evidenzia la significatività della differenza fra le medie di **tre o più** gruppi di campioni
  - Testa le differenze tra medie analizzando le varianze.
- L'**Analisi della Varianza** partiziona la varianza totale per confrontare la variabilità *tra* gruppi con la variabilità *dentro* i gruppi.
- Al cuore dell'analisi della varianza c'è il fatto che la varianza può essere partizionata.
  - la varianza è calcolata come somma dei quadrati della dev. std dalla media totale.

Confronto set di dati sperimentali

Confronto la variabilità totale TRA i gruppi ed ENTRO i gruppi.

Se la variabilità fra i gruppi è legata esclusivamente al caso la variabilità totale TRA i gruppi ed ENTRO i gruppi saranno simili.

La variabilità è misurata dalla varianza (QM), si definisce perciò:

QM(tra) = SS(tra)/GL(tra) TRA gruppi

QM(entro) = SS(entro)/GL(entro) ENTRO gruppi

Dove SS è la devianza o somma degli scarti $(Xi — \mathbf{Xm})^2$ e GL sono i gradi di libertà.

Perciò:

$$F=QM(tra)/QM(entro)$$

esprime la significatività delle differenza esistente tra i gruppi per un dato livello di probabilità.

Confronto set di dati sperimentali

	Gruppo A	Gruppo B
Obs 1	2	4
Obs 2	3	5
Obs 3	4	6
Mean	3	5
(SS)	2	2
Overall Mean	4	
Total SS	10	

Esempio:

- Le medie dei due gruppi sono diverse
- La somma dei quadrati della dev. std. è 2; sommandole tra loro ottengo:

SS (gruppo A)+SS (gruppo B)=4

Questa è la SS **ENTRO** i gruppi

- La SS totale è 10 (≠ 4).
- SS **TRA** gruppi sarà:

SS (totale)-SS (entro)=6

Confronto set di dati sperimentali

	Gruppo A	Gruppo B
Obs 1	2	4
Obs 2	3	5
Obs 3	4	6
Mean	3	5
(SS)	2	2
Overall Mean	4	
Total SS	10	

Esempio:

QM (tra)=SS (tra)/GL(tra)=6

QM (entro)=SS (entro)/GL(entro)=1

Concludendo:

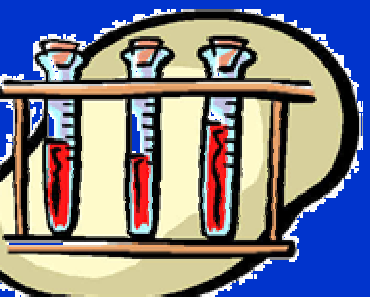
F=6/1=6

SS (tra) è anche detta MSEffect mentre SS (entro) è anche detta MSError

Nel caso nullo il loro rapporto tende a uno.

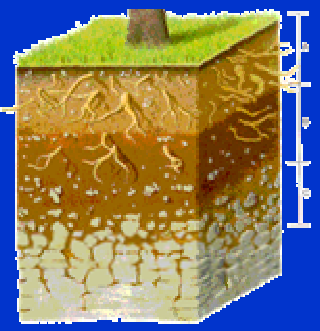
Confronto set di dati sperimentali

- Ad una via: si considera una sola variabile su più gruppi di campioni indipendenti, ognuno sottoposto ad un diverso livello della variabile (es: C somministrata).
- A due vie: viene analizzato l'effetto di due variabili indipendenti *contemporaneamente*.
  - si vede anche se le due variabili interagiscono nel loro effetto.
- A una via per campioni correlati.
  - campione sottoposto successivamente a più condizioni sperimentali.
  - come nel test t, l'utilità è sempre quella di eliminare la variabilità estranea all'esperimento.



Confronto set di dati sperimentali

- ANOVA: esempi di applicazione



  - Confronto della capacità di adsorbimento del fosforo di tre orizzonti del suolo.
  - Confronto dell'efficienza di abbattimento dei nutrienti tra più zone umide artificiali trattanti lo stesso refluo.
  - Confronto dei tassi potenziali di nitrificazione per bacino idrografico e stagione (ANOVA a due vie)
  - Confronto dei dati di biomassa relativi a diverse zone umide in relazione all'anno di campionamento, alla zona umida di appartenenza, alla posizione del campione all'interno della wetland.

Confronto set di dati sperimentali

In presenza di più gruppi descritti da più variabili

→ MANOVA

Analisi multidimensionale della varianza

Dalle variabili dipendenti devono essere preventivamente estratte delle nuove variabili dipendenti (artificiali) che derivano dalla combinazione lineare delle variabili originarie.

non parametrico

Confronto set di dati sperimentali

- ***Equivalente non parametrico:***
- *Kruskal-Wallis ANOVA by Ranks test*
  - Analisi della varianza utilizzabile nel caso in cui le condizioni di base non siano realizzate;
  - Si applica a campioni indipendenti.
- *Friedman ANOVA*
  - Analisi non parametrica della varianza per campioni dipendenti

Contenuti

- Terminologia
- Concetti di base
- Metodi statistici
  - Confronto set di dati sperimentali
  - Analisi esplorativa di dati multidimensionali
- Interpretazione dei risultati

Contenuti – metodi statistici

Analisi esplorativa di dati multidimensionali

- Caratteristiche
- Metodi fattoriali di tipo esplorativo
  - Analisi fattoriale
  - Analisi delle componenti principali
  - Analisi fattoriale delle corrispondenze
  - Multidimensional scaling (MDS)
- Analisi di correlazione
  - Analisi di correlazione semplice
  - Analisi di correlazione canonica
- Metodi di ordinamento e di classificazione
  - Cluster analysis
  - Analisi discriminante
- Analisi di regressione multipla
- Modelli di regressione non lineari
  - Generalized linear – non linear models (GLZ)
  - Modelli di regressione addittivi
  - Generalized Additive Models (GAM)

Analisi esplorativa di dati multidimensionali

- Le analisi multidimensionali sono metodiche particolarmente utili nell'analisi di sistemi ecologici complessi.
- Dal momento che molte variabili sono necessarie per la spiegazione del comportamento di un ecosistema, i dataset ecologici sono necessariamente multidimensionali, o multivariati.

- Analisi statistica **multivariata:** rilevate più variabili, si decide di analizzarle variabili a coppie o per gruppi, in modo da evidenziarne le relazioni.
- Analisi **multidimensionale**: metodiche di analisi appropriate nel caso in cui si assuma l'esistenza di più dimensioni sottostanti all'insieme di dati analizzato.

Analisi esplorativa di dati multidimensionali

- *riduzione dei dati*:
  - descrizione dei dati rilevati attraverso eliminazione delle ridondanze
- *discriminazione e classificazione*:
  - determinazione delle regole per la separazione ottima degli insiemi di unità in esame - applicazione di regole prefissate per separare le unità.
- *raggruppamento (**clustering**) e ricerca tipologica*:
  - definizione delle classi di unità più somiglianti.
- *ricerca di determinanti*:
  - spiegare il comportamento di una o più variabili *criterio*, o *dipendenti*, in funzione di un insieme di variabili esplicative, *predittive*, o *indipendenti*.
- *costruzione di modelli*:
  - sistemi interrelati di variabili che spiegano un determinato fenomeno.
- *evidenziazione di strutture latenti*:
  - Le metodiche ipotizzano che le variabili osservate siano approssimazioni misurabili di variabili non manifeste

Analisi esplorativa di dati multidimensionali

- **SIMMETRIA**
  - Metodi simmetrici considerano le variabili sullo stesso piano causale, le relazioni sono considerate bidirezionali. Non evidenzia nessi causali.
  - Metodi asimmetrici intendono evidenziare relazioni di dipendenza tra più sistemi di variabili. Le variabili osservate sono divise tra variabili dipendenti e variabili predittive.
- **METRICA**
  - Analisi metrica: realizzata con dati quantitativi
  - Analisi non metrica: si basa su algoritmi che possono essere applicati qualunque sia la scala di misura delle variabili.
- **LINEARITA'**
  - per le analisi metriche si assume, in genere, che la relazione che lega le variabili sia esprimibile come una funzione lineare.

Contenuti – metodi statistici

Analisi esplorativa di dati multidimensionali

- Caratteristiche
- Metodi fattoriali di tipo esplorativo
  - Analisi fattoriale
  - Analisi delle componenti principali
  - Analisi fattoriale delle corrispondenze
  - Multidimensional scaling (MDS)
- Analisi di correlazione
  - Analisi di correlazione semplice
  - Analisi di correlazione canonica
- Metodi di ordinamento e di classificazione
  - Cluster analysis
  - Analisi discriminante
- Analisi di regressione multipla
- Modelli di regressione non lineari
  - Generalized linear – non linear models (GLZ)
  - Modelli di regressione addittivi
  - Generalized Additive Models (GAM)

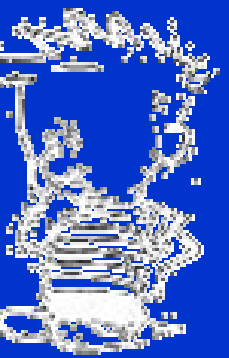
Metodi fattoriali di tipo esplorativo

- Ricavare dall'analisi di un fenomeno gli elementi conoscitivi fondamentali per l'individuazione di nuovi modelli o la verifica di ipotesi.
- Permettono di
  - ridurre la multidimensionalità delle matrici di dati attraverso l'estrazione di nuove variabili tra loro incorrelate (componenti principali o fattori);
  - rappresentare e sintetizzare il fenomeno mediante dimensioni nuove (assi fattoriali), ovvero differenti punti di vista per l'interpretazione.
- METODI
  - Analisi fattoriale
  - Analisi delle componenti principali (PCA)
  - Analisi fattoriale delle corrispondenze
  - Scaling Multidimensionale

Metodi fattoriali di tipo esplorativo

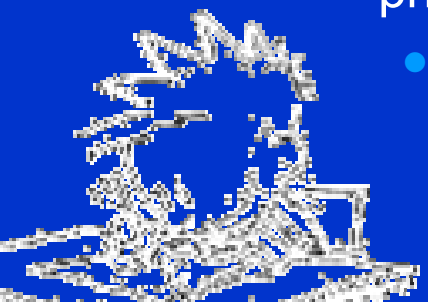
Analisi Fattoriale

- È' il primo (e progenitore) metodo di analisi multivariata
- Si applica a un insieme di variabili *quantitative* tra cui si assumono relazioni *simmetriche*.
- Scopo: spiegare la molteplicità delle relazioni esistenti tra variabili attraverso un numero ridotto di fattori
- I fattori sono ripuliti sia dalla variabilità ridondante, dovuta alla presenza di variabili correlate, sia dalla variabilità spuria, introdotta da variabili marginali rispetto ai fenomeni considerati.
- Principi: parsimonia nella rappresentazione matematica (poche dimensioni significative); robustezza essenziale dell'analisi, immediata percettibilità delle soluzioni grafiche ottenibili.
- ☺ finalità più ampie, assunti più specifici, problemi empirici più particolari della PCA
- ☹ indeterminatezza o non unicità della soluzione fattoriale (insiemi contraddittori di punteggi fattoriali sono ugualmente plausibili).



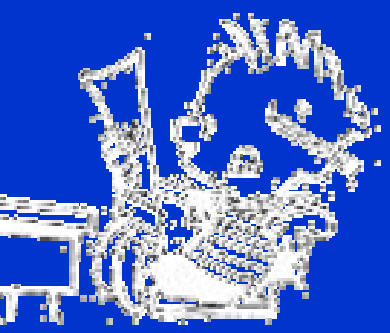
Metodi fattoriali di tipo esplorativo
Analisi delle componenti principali

- Scopo: sostituire le variabili con un numero inferiore di fattori indipendenti, eliminando la variabilità dovuta alla correlazione delle variabili.
  - Analisi *simmetrica*, su variabili *quantitative* organizzate in matrici di *correlazione*.
- Componenti principali = combinazioni lineari delle variabili la cui varianza complessiva uguagli quella osservata.
  - Tali componenti sono tra loro ortogonali, quindi indipendenti.
  - In genere, la maggior parte della varianza spiegata si concentra sulle prime, poche componenti principali.
- Autovalore: varianza della corrispondente componente principale.
  - Estrazione in ordine decrescente: spiegano progressivamente una quota sempre minore d'informazione.



Metodi fattoriali di tipo esplorativo
Analisi delle componenti principali

- Non esiste un criterio universale per determinare il n° di fattori da estrarre:
  - compromesso tra necessità di semplificare la struttura dei dati - volontà di spiegare la quota maggiore di varianza.
- I criteri più utilizzati sono:
  - *Numero di fattori prefissato*: in relazione all'ampiezza del campo di osservazione, o all'esigenze dell'analisi;
  - *Varianza estratta dai fattori*: estrazione dei fattori fino al raggiungimento della soglia di varianza prefissata;
  - *Criterio di Kaiser*: estrazione dei fattori il cui autovalore è maggiore o uguale a 1 (spiegano una quota di variabilità almeno pari a quella delle variabili di partenza)
  - *Rappresentazione grafica autovalori* in funzione del n° fattori: Se la spezzata mostra due tendenze si considerano rilevanti solo i valori che, più in alto del flesso, si staccano visibilmente dagli altri. È considerato il criterio di selezione più severo.



Metodi fattoriali di tipo esplorativo
Analisi delle componenti principali

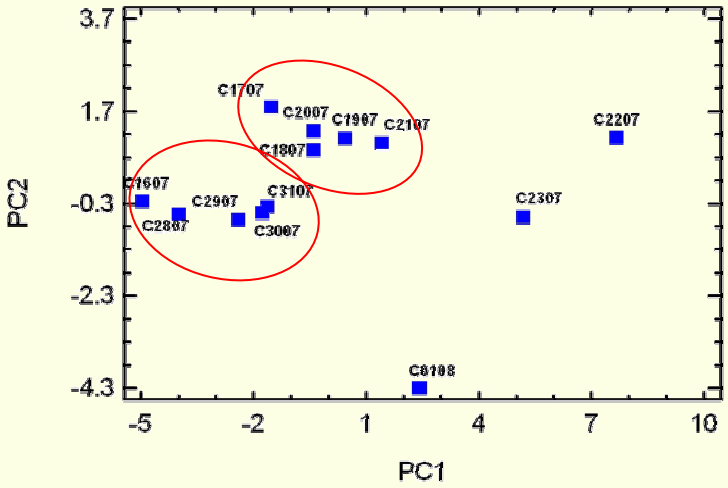
- ☹ come nel caso dell'analisi fattoriale, l'arbitrarietà dei criteri che guidano il risultato dell'analisi, soprattutto la scelta di porre fine all'estrazione.
- ☺ pone meno problemi definitori di un'analisi fattoriale e, nelle applicazioni pratiche, dà risultati simili
- L'attenzione da porre è all'utilizzo di variabili quantitative, possibilmente in numero minore dei casi osservati.

Metodi fattoriali di tipo esplorativo
Analisi delle componenti principali

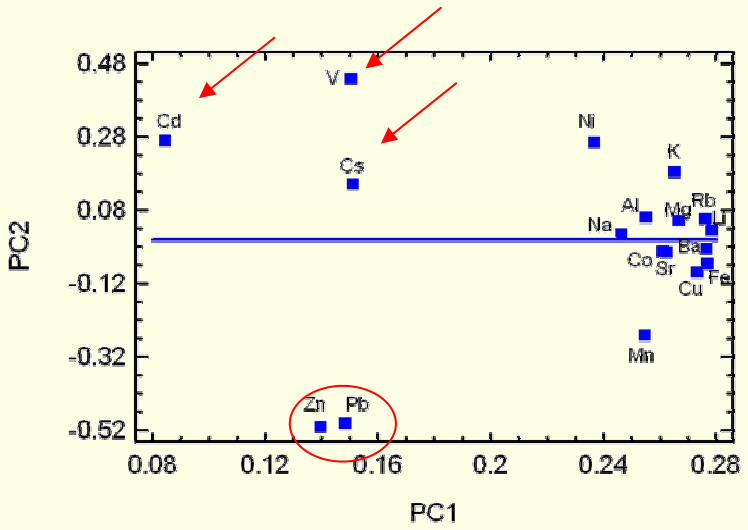
- Esempio di applicazione – Studio della qualità aria nella laguna di Venezia
  - Oggetti: campioni di particolato atmosferico (PM 2,5)
  - Variabili: concentrazioni di diversi tipi di microinquinanti inorganici nell'aria
  - Numero di componenti principali: definito dalla percentuale di varianza spiegata (2 componenti spiegano l'85% della varianza totale)
- Risultati:
  - Scatter plot: distribuzione degli oggetti (campioni) nello spazio delle componenti principali
  - Component weights: consente di analizzare il ruolo di ciascuna variabile (elementi) nello spazio delle componenti principali e quindi di stabilire l'esistenza di correlazioni e la loro importanza relativa

Metodi fattoriali di tipo esplorativo
Analisi delle componenti principali

Scatter plot



Component weights



Metodi fattoriali di tipo esplorativo
Analisi delle componenti principali

- Analisi dello scater plot

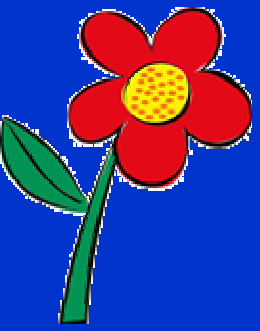
  Dall'analisi dello scatter plot si possono individuare due gruppi, evidenziati da un cerchio rosso, che corrispondo a periodi di campionamento diversi. I campioni C2207, C2307 e C0108 si presentano isolati.

- Analisi del component weights:
  - La prima componente principale tiene conto della somma degli elementi presenti in ogni campione
  - La seconda componente discrimina i due gruppi evidenziati in rosso, questi si differenziano soprattutto in base al diverso contenuto di V e Cd (peso elevato nella seconda componente principale)
  - Anche lo Zn e il Pb sono influenti su questa componente, ma dall'analisi dei dati originali non sembrano esserci delle differenze significative nei valori di questi due elementi tra il primo e il secondo periodo di campionamento. Hanno comunque la stessa dinamica ambientale.

Metodi fattoriali di tipo esplorativo

Analisi fattoriale delle corrispondenze

- Permette la suddivisione dell'insieme di unità in esame in due parti, una delle quali è usata per la ricerca della soluzione, l'altra per paragone. *Metodo asimmetrico.*
- Si applica a *tabelle di frequenze* derivanti dall'osservazione di due o più variabili *qualitative*.
- Scopo: individuare l'eventuale struttura di dipendenza che lega tra loro due insiemi, attraverso una rappresentazione grafica della loro distribuzione in uno spazio con un numero minore di dimensioni.
- Si evidenziano i fattori impliciti nei dati rilevati.
- In ecologia, è utilizzata prevalentemente per lo studio dei dati sulla presenza / assenza di specie in differenti punti di campionamento



Metodi fattoriali di tipo esplorativo
Analisi fattoriale delle corrispondenze

- ☺ È utile quando si vuole analizzare una tabella di ampie dimensioni derivata dalla scomposizione di un fenomeno secondo due caratteristiche (solitamente due variabili qualitative), per le quali ha interesse studiare la distribuzione congiunta.
- Diversamente dalla PCA, l'analisi delle corrispondenze prescinde da considerazioni sulla scala di misura delle variabili e da assunti funzionali tra le stesse.
- ☹ Essendo una tecnica esplorativa, non vi sono test statistici di significatività che siano ordinariamente applicati ai risultati conseguiti.

Metodi fattoriali di tipo esplorativo
Analisi fattoriale delle corrispondenze

- Esempio di applicazione
- Studio dell'organizzazione strutturale di tre paesaggi agricoli nel Veneto
  - Finalità: influenza dell'organizzazione strutturale sulla biodiversità
- L'analisi delle corrispondenze è stata utilizzata per individuare quali valori delle variabili utilizzate spiegavano maggiormente le diversità tra i tre paesaggi, a scale differenti
  - Variabili: valori di copertura delle siepi, delle aree urbanizzate, di *5* categorie colturali; valori di 3 indici di struttura.
  - Ognuna di queste è stata divisa in classi %, e ogni % ha costituito una variabile per l'analisi.

Contenuti – metodi statistici

Analisi esplorativa di dati multidimensionali

- Caratteristiche
- Metodi fattoriali di tipo esplorativo
  - Analisi fattoriale
  - Analisi delle componenti principali
  - Analisi fattoriale delle corrispondenze
  - Multidimensional scaling (MDS)
- Analisi di correlazione
  - Analisi di correlazione semplice
  - Analisi di correlazione canonica
- Metodi di ordinamento e di classificazione
  - Cluster analysis
  - Analisi discriminante
- Analisi di regressione multipla
- Modelli di regressione non lineari
  - Generalized linear – non linear models (GLZ)
  - Modelli di regressione addittivi
  - Generalized Additive Models (GAM)

Metodi fattoriali di tipo esplorativo

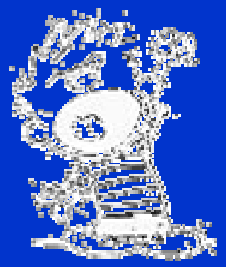
Scaling Multidimensionale (MDS)

non parametrico

- Metodo *non parametrico* di tipo esplorativo, considerato la principale alternativa all'analisi fattoriale.
- Insieme di procedure che, attraverso l'individuazione delle dimensioni significative sottostanti ai dati, consente di spiegare le differenze o le relazioni di somiglianza tra gli oggetti analizzati.
- Non richiede l'utilizzo di una matrice di correlazione, ma può essere applicato a variabili su scala *non metrica* organizzati in matrici di *prossimità*, *similarità* o *dissimilarità*.
  - L'unica assunzione sui dati è l'esistenza di relazioni monotone tra essi.
- Scopo: determinare le coordinate geometriche di un insieme di entità, ovvero individuare una configurazione degli oggetti in uno spazio costituito da un numero di dimensioni minore rispetto a quello di partenza.
- Stima della bontà dell'adattamento della configurazione stimata rispetto al *set* di dati: indice di stress
  - il valore oltre il quale l'indice è significativo non è determinabile statisticamente. Kruskal ha individuato cinque intervalli di valori dell'indice di stress, ad ognuno dei quali corrisponde una valutazione dell'adattamento della configurazione.

Metodi fattoriali di tipo esplorativo

*Scaling Multidimensionale (*MDS)

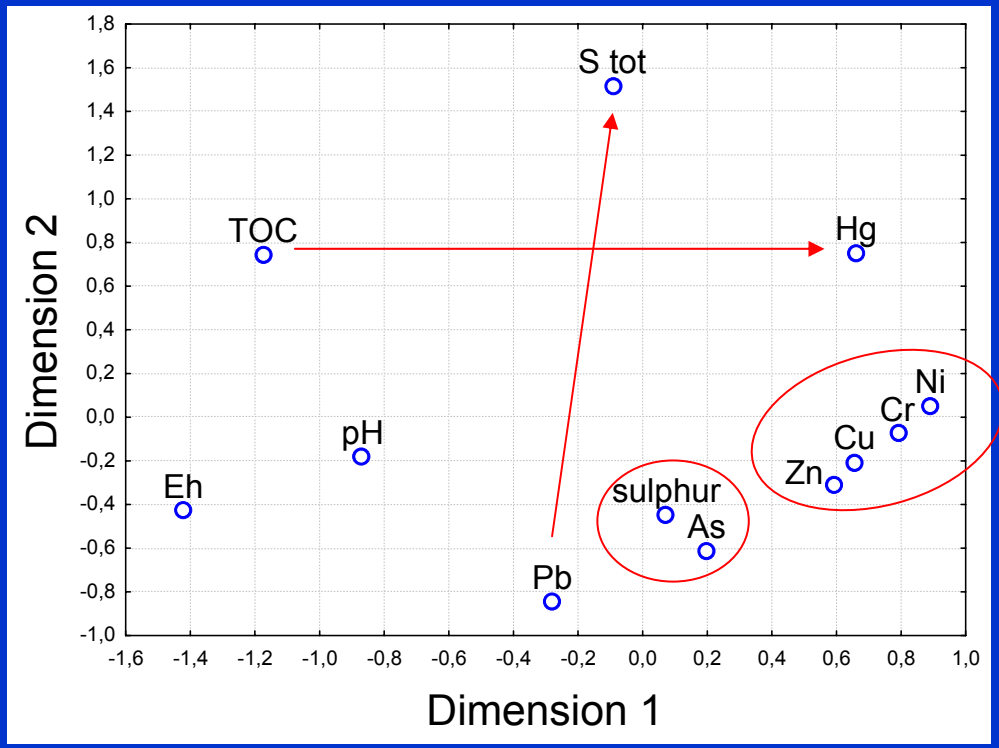


- ☺ non impone alcun vincolo al tipo di matrice da analizzare, non si richiede che i dati siano distribuiti come normali multivariate e nemmeno che le relazioni tra i dati siano lineari.
- ☹ Costringere i dati su due o tre dimensioni può tuttavia incidere sulla qualità della soluzione finale. E' necessario determinare fino a che punto è possibile ridurre le dimensioni senza che la qualità delle soluzioni abbia a soffrirne.

Metodi fattoriali di tipo esplorativo

Scaling Multidimensionale (MDS)

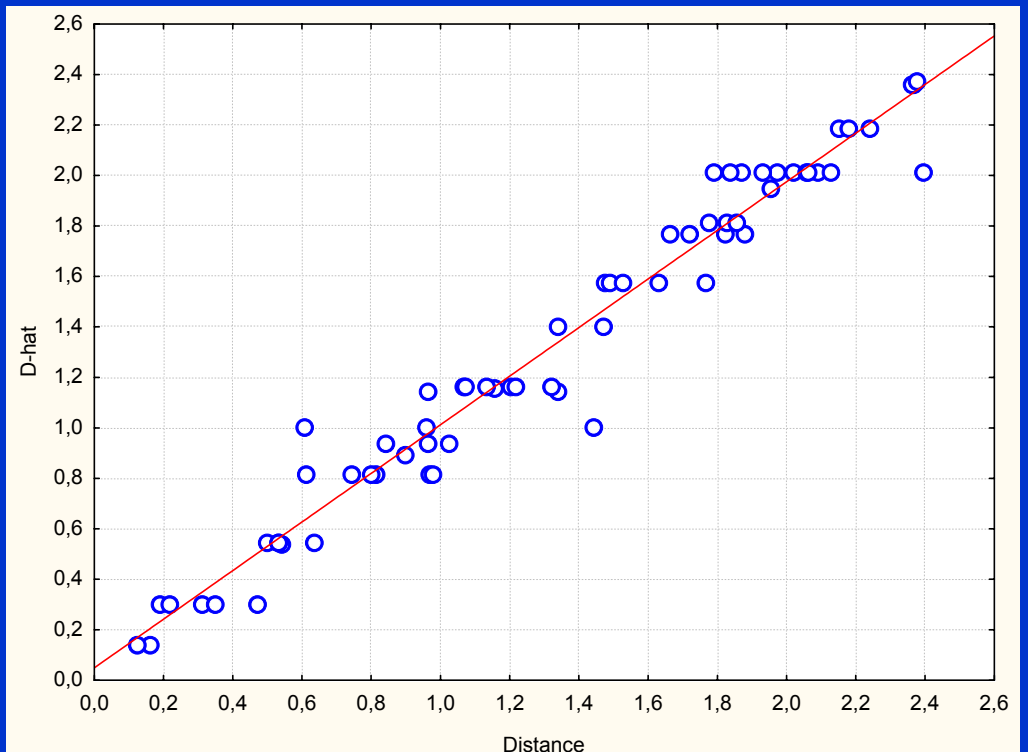
- Esempio di applicazione – Analisi delle dinamiche dei metalli nei sedimenti di un impianto di fitodepurazione



Metodi fattoriali di tipo esplorativo

Scaling Multidimensionale (MDS)

- Esempio di applicazione – Analisi del ciclo dei metalli nei sedimenti di un impianto di fitodepurazione



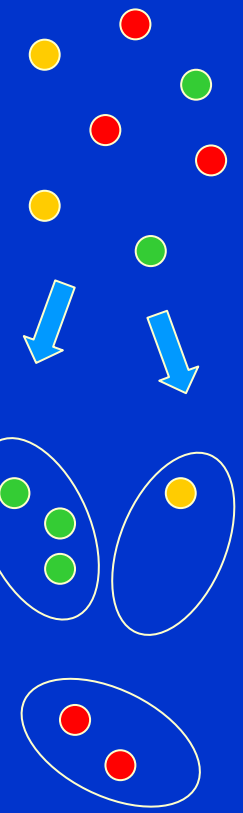
Contenuti – metodi statistici

Analisi esplorativa di dati multidimensionali

- Caratteristiche
- Metodi fattoriali di tipo esplorativo
  - Analisi fattoriale
  - Analisi delle componenti principali
  - Analisi fattoriale delle corrispondenze
  - Multidimensional scaling (MDS)
- Analisi di correlazione
  - Analisi di correlazione semplice
  - Analisi di correlazione canonica
- Metodi di ordinamento e di classificazione
  - Cluster analysis
  - Analisi discriminante
- Analisi di regressione multipla
- Modelli di regressione non lineari
  - Generalized linear – non linear models (GLZ)
  - Modelli di regressione addittivi
  - Generalized Additive Models (GAM)

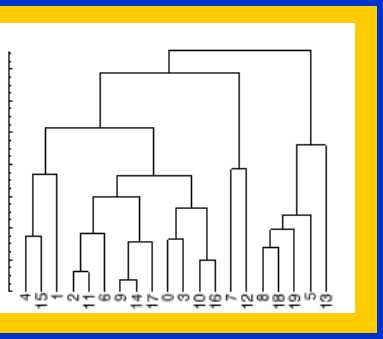
Metodi di ordinamento e classificazione

Cluster Analysis



- Insieme di tecniche volte a classificare l'insieme delle unità d'analisi in gruppi non definiti a priori
- Matrice dei dati: n osservazioni su p variabili,
  - in forma grezza o dopo opportuna trasformazione;
  - rappresentati su qualsiasi scala,
  - o resi indipendenti dall'unità di misura tramite standardizzazione.
- Scopo: assegnare le entità a poche categorie (o classi, o gruppi) non definite a priori, in modo da massimizzare l'omogeneità interna, ovvero massimizzare le differenze tra i gruppi.
- Non si fanno assunzioni in anticipo, ma si attende che le relazioni tra le unità siano evidenziate dall'analisi

Metodi di ordinamento e classificazione
Cluster Analysis



- Ambiti di applicazione:
  - ridurre i dati in forma grafica semplice e percepibile per evidenziare le informazioni rilevate
    - strumento di presentazione di analisi multivariate;
  - generare ipotesi di ricerca, con una classificazione che mostri le connessioni reali tra entità, e intuire in base a queste i patterns presenti nei dati.

- Non esiste un metodo univoco per condurre l'analisi.
  - selezionare una misura di prossimità, sulla base di cui confrontare gli elementi;
  - scegliere la tecnica di raggruppamento delle entità, ovvero il criterio di aggregazione.

Metodi di ordinamento e classificazione
Cluster Analysis

- **Criteri di classificazione**:
  - ottimizzano l'omogeneità interna delle classi e la loro reciproca eterogeneità, oppure
  - determinazione dei gruppi in base a una precisa struttura interna.
- A seconda del criterio, l'analisi si definisce:
  - *Gerarchica*: ripartizione in dendrogramma, ogni classe è contenuta nella successiva. Individua le partizioni senza indicarne a priori il numero.
    - Agglomerativa: aggregazione degli elementi in direzione ascendente, dalle singole entità ad un'unica classe.
    - Scissoria: Aggregazione degli elementi in direzione discendente.
  - *Non gerarchica*: produce gruppi non gerarchizzabili attraverso iterazioni successive, in cui ogni partizione sostituisce la precedente. Il numero di classi dev'essere definito a priori, ed è necessario partire da una soluzione provvisoria.

Metodi di ordinamento e classificazione
Cluster Analysis

Il **punto di partenza** di tutti gli algoritmi di clustering è un modello che prescinde completamente alla natura dei dati impiegati e dalle specifiche problematiche disciplinari. Si fa riferimento in generale ad una *matrice dei dati* contenente informazioni su N oggetti (casi o osservazioni; righe della matrice) specificate dai valori assegnati a V variabili (colonne della matrice).

→ Dalla matrice dei dati originaria si passa ad una **matrice di distanze o di similarità** tra casi.

Metodi di ordinamento e classificazione
Cluster Analysis

La matrice delle distanze è una matrice in cui sono riportate le “distanze” tra le unità oggetto di indagine.

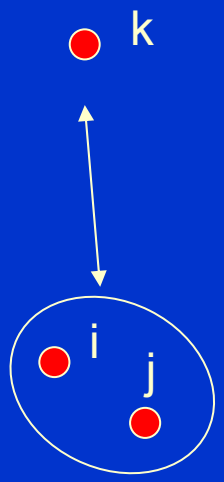
CARATTERISTICHE:
- Quadrata
- Simmetrica
- I valori sulla diagonale sono nulli

→ Esistono vari approcci per calcolare la distanza tipica tra entità:
Distanza euclidea (geometrica), delle media assoluta, etc..

Per le analisi non metriche si parla più propriamente di matrici di somiglianza/dissomiglianza

Metodi di ordinamento e classificazione
Cluster Analysis

- **Analisi Gerarchica Agglomerativa**:
  - Si procede nel seguente modo:
    - Calcolo della matrice di prossimità (somiglianza/dissomiglianza).
    - Estrazione della coppia di unità che sono più prossime (I gruppo)
    - All'interno del gruppo la distanza viene assunta essere nulla.
    - Si calcola la distanza tra il nuovo gruppo e le rimanenti entità. Esistono diversi tipi di algoritmi.
    - Si prosegue fino a che tutte le entità fanno parte di un unico gruppo.



Metodi di ordinamento e classificazione
Cluster Analysis

- **Analisi Gerarchica Scissoria:**
- Concettualmente opposto al processo di aggregazione: si parte da un unico gruppo e si arriva alla formazione di n gruppi costituiti da una sola unità.
- I criteri di divisione sono più generali perché permettono la formazione di un n° qualsiasi di sottogruppi.
  - Suddivisione binaria: un attributo dicotomico alla volta.
  - Suddivisione politetica: tutti gli attributi ad un tempo.

Metodi di ordinamento e classificazione
Cluster Analysis

- **Criteri per determinare il numero ottimo di gruppi**
  - rappresentazione grafica di soluzioni: dove è più evidente la discontinuità;
  - ispezione del dendrogramma: si seleziona all'altezza del massimo salto tra livelli di prossimità.
- **Criteri per interpretare una soluzione gerarchica:**
  - La lettura del dendogramma è effettuata a partire dall'ultima fusione tra gli elementi fino al punto in cui il numero di gruppi è considerato significativo.
  - Si identificano, all'indietro, le caratteristiche dei diversi gruppi.
  - Di ogni gruppo si calcolano le statistiche qualificanti.

Metodi di ordinamento e classificazione

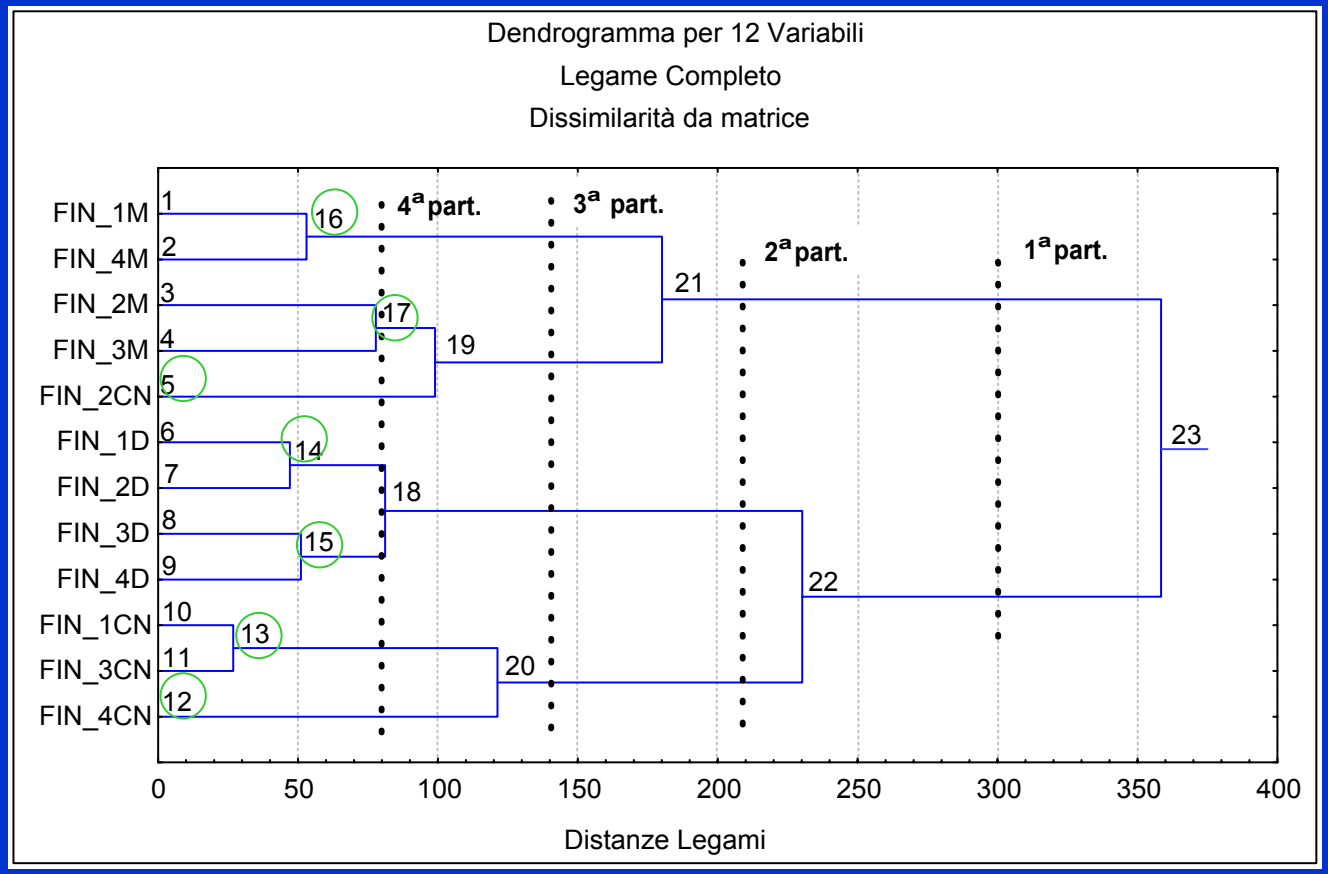
Cluster Analysis e Analisi fattoriale

- Forniscono informazioni diverse e complementari, svolgendo un ruolo di sussidiarietà.
  - l'analisi fattoriale è una tecnica appropriata per lo studio delle *relazioni tra le variabili*; assume che tali relazioni siano lineari.
  - l'analisi dei gruppi permette di *raggruppare le entità descritte* da tali variabili; la forma delle relazioni tra variabili è trascurabile.
- Dopo aver eseguito una cluster analysis si possono rendere evidenti le variabili più discriminanti tra le entità; dopo un'analisi dei fattori, le unità che più sono simili rispetto ai fattori trovati.

Metodi di ordinamento e classificazione
Cluster Analysis

- **Esempio di applicazione**
- Studio dell'organizzazione strutturale di tre paesaggi agricoli nel Veneto
- L'analisi di classificazione è stata utilizzata per identificare, tra i paesaggi analizzati, gruppi omogenei rispetto alle variabili studiate.
  - Relazioni di somiglianza
  - Modelli strutturali caratterizzanti

Metodi di ordinamento e classificazione
Cluster Analysis



Metodi di ordinamento e classificazione

Analisi discriminante

- A differenza dei metodi classificazione automatica, la suddivisione degli elementi avviene in base a classi *già definite,* sulla base di una tassonomia pre-esistente.
- Si applica a variabili *quantitative* o *qualitative*;
- Scopo: determinazione delle caratteristiche che permettono di distinguere tra loro due o più oggetti definiti da un insieme di variabili.

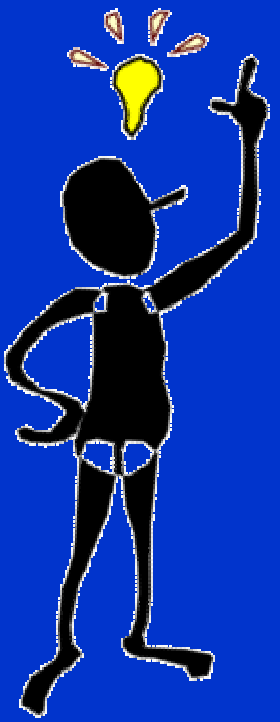
Contenuti – metodi statistici

Analisi esplorativa di dati multidimensionali

- Caratteristiche
- Metodi fattoriali di tipo esplorativo
  - Analisi fattoriale
  - Analisi delle componenti principali
  - Analisi fattoriale delle corrispondenze
  - Multidimensional scaling (MDS)
- Analisi di correlazione
  - Analisi di correlazione semplice
  - Analisi di correlazione canonica
- Metodi di ordinamento e di classificazione
  - Cluster analysis
  - Analisi discriminante
- Analisi di regressione multipla
- Modelli di regressione non lineari
  - Generalized linear – non linear models (GLZ)
  - Modelli di regressione addittivi
  - Generalized Additive Models (GAM)

Analisi di regressione multipla

- È un metodo di analisi *asimmetrico:*
  - verificare le relazioni tra una variabile dipendente (criterio) e *un set* di variabili indipendenti (predittori).
  - Nel modello multivariato (più di criterio) l'analisi è ripetuta per ogni criterio con le stesse variabili predittive.
  - determinare la funzione di regressione lineare che interpreta al meglio la variabilità dei criteri, pur utilizzando un numero ridotto di predittori.
- Assunzioni (metodo parametrico):
  - Linearità delle relazioni tra variabili;
    - in pratica, non è influenzata molto da piccole deviazioni da questo assunto;
    - se la deviazione è evidente, si può considerare la trasformazione delle variabili o la ricerca di componenti non lineari.
  - Normalità della distribuzione dei valori residuali (differenza tra valori previsti e valori osservati).



Analisi di regressione multipla

- Il metodo di analisi assume che ogni osservazione y sia esprimibile come una combinazione lineare delle **x** con coefficienti b e di una variabile non osservata ε, errore residuale della regressione.
- Le procedure stimano un'equazione lineare:

$$Y=a+b_1X_1+b_2X_2+...+b_pX_p$$

- *b:* coefficienti di regressione:
  - contributo indipendente di ogni variabile alla previsione;
  - il segno del coefficiente dà la direzione della relazione;

- Se i predittori tra loro hanno correlazione nulla, per sapere quali escludere si valuta la correlazione con la variabile criterio;
  - vale il cosiddetto "principio di parsimonia": ogni variabile che non contribuisce significativamente alla definizione del modello dev'essere esclusa.

Analisi di regressione multipla

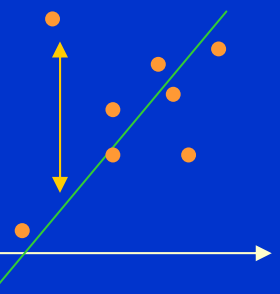
- I metodi di regressione multipla si possono classificare secondo tre criteri di analisi:
- 1. selezione progressiva (*forward selection*):
  - inserire una variabile alla volta, finché non è soddisfatto un criterio di arresto della procedura;
- 2. eliminazione a ritroso (*backward selection*):
  - rimozione di una variabile alla volta dall'equazione di regressione.
  - Ha il vantaggio di tener conto di predittori congiuntamente esplicativi ma individualmente non correlati con la variabile criterio;

Analisi di regressione multipla

- 3. regressione *stepwise* :
  - Prevede l'inserimento o la cancellazione, una alla volta, delle singole variabili, tenendo conto dell'apporto interpretativo della sequenza
  - una variabile è inclusa nell'equazione se dà il contributo più significativo all'interpretazione della variabile criterio, ma può essere rimossa nelle fasi successive.
- Si utilizza quando non ci sono conoscenze teoriche sufficienti per impostare una equazione di regressione (non si sa quali variabili adoperare).
- È la procedura più utilizzata, che soddisfa ogni necessità generica di selezione.

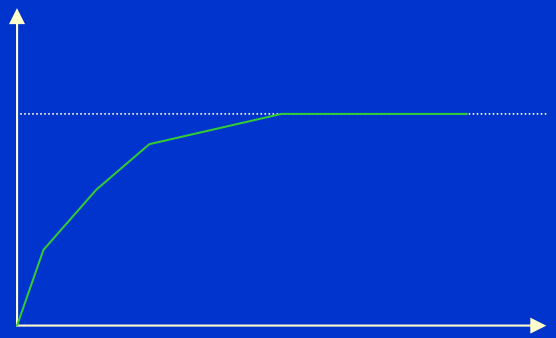
Analisi di regressione multipla

- **Criteri per valutare l'esito di una regressione**
- La deviazione di un punto dalla retta di regressione (o valore atteso) è detto **valore residuale**.
  - Minore è la variabilità dei valori residuali attorno alla retta, migliore è la previsione.
  - Nella maggior parte dei casi, il rapporto della variabilità residuale della variabile dipendente sulla variabilità originale cade tra 0 e 1.
  - 1 meno questo rapporto è detto R2 o coefficiente di determinazione.
- *R^2, o coefficiente di determinazione:*
  - frazione di variabilità originale spiegata dal modello di regressione (Es: $R^2 = 0.4$ significa che il modello ha spiegato il 40% della variabilità originale).
  - La capacità esplicativa di un predittore è valutata in relazione alla diminuzione di varianza (o aumento di R^2) che segue l'inserimento.
  - A questa capacità si rifanno i criteri per decidere la selezione dei predittori e il momento in cui arrestare il processo.



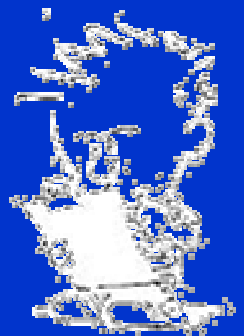
Analisi di regressione multipla

- *rappresentazione di R^2*:
- detto k il numero di passi del processo, si osserva che R^2 cresce rapidamente con le prime variabili, per poi attestarsi attorno al valore finale (curva con asintoto orizzontale).
- Si evidenzia, cioè, un brusco cambiamento di pendenza in corrispondenza del punto k*. Il numero di variabili ottimo è, allora, k-1; le altre variabili non determinano un aumento significativo della varianza.



Analisi di regressione multipla

- **Importanza di un predittore**:
- il *quadrato* del *coefficiente* di *correlazione* tra y e il predittore xj misura la capacità esplicativa globale del predittore xj.
  - È valido solo se le variabili predittive selezionate sono poco correlate.
- *Contributo netto*:
  - il *quadrato* del *coefficiente* di *regressione parziale* tra y e xj (b2)
  - frazione di devianza di y interpretata da xj al netto di tutti gli altri predittori
- *Contributo indipendente*:
  - prodotto del *coefficiente di correlazione* e del *coefficiente di regressione parziale* (b e r); sommato su tutti i predittori dà R2.
  - In presenza di coefficienti con segni opposti, da un valore negativo non interpretabile come contributo.

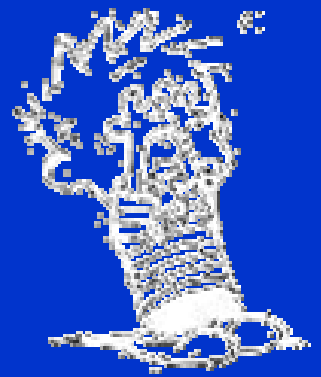


Analisi di regressione multipla

- **Criteri per decidere l'arresto del processo di selezione.**
- Non è individuabile un criterio assolutamente oggettivo e valido:
- n°massimo di predittori:
  - l'interpretazione è ardua se i predittori sono intercorrelati e il numero di variabili è alto.
  - la maggior parte degli autori raccomanda un numero di osservazioni pari ad almeno 10 o 20 volte il numero di variabili, altrimenti il risultato della regressione è instabile e improbabile da replicare
- Frazione di devianza complessivamente spiegata:
  - si può essere soddisfatti quando la frazione spiegata è significativa.
- Frazione di devianza spiegata dal predittore marginale:
  - il contributo esplicativo dell'ultimo predittore entrato si misura con l'incremento di R2 conseguente.
- Tolleranza:
  - l'originalità di un predittore è valutata in funzione della correlazione con i predittori presenti nell'equazione (escluso se è correlato).
  - Un valore selettivo di T può arrestare l'analisi alle prime battute.

Analisi di regressione multipla

- ☺ Il modello è:
  - additivo: il contributo interpretativo di un predittore è sommabile agli altri
  - asimmetrico: assume dipendenza tra variabile criterio e predittori
- ☹ limite concettuale delle regressioni:
  - si possono solo scoprire delle relazioni, ma non si può mai essere sicuri che non siano il risultato di meccanismi casuali.

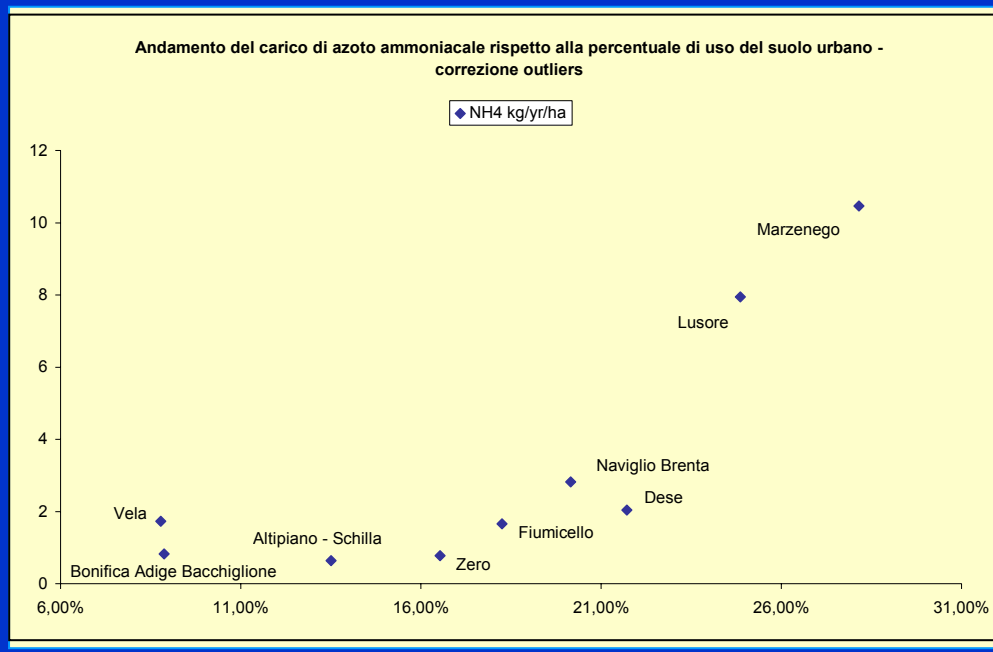


Analisi di regressione multipla e PCA

- **Trasformazione in componenti principali**:
- Se le variabili predittori fossero indipendenti, il contributo di ognuna di esse sarebbe la correlazione singola con la variabile criterio.
- si può passare attraverso una trasformazione dei predittori in componenti principali, e utilizzare quest'ultime nell'analisi di regressione.
- Applicato all'intero insieme di potenziali predittori, la PCA è una forma grezza di selezione delle variabili perché individua quelle che, contribuendo poco alla spiegazione della variabilità di y, possono essere escluse ai fini dell'equazione di regressione.

Analisi di regressione multipla

- Esempio di applicazione
- Studio qualità H_2O bacino scolante
- Modello di regressione per spiegare i carichi di ammoniaca:
  - % uso del suolo urbano
  - Indice di eterogeneità



Contenuti – metodi statistici

Analisi esplorativa di dati multidimensionali

- Caratteristiche
- Metodi fattoriali di tipo esplorativo
  - Analisi fattoriale
  - Analisi delle componenti principali
  - Analisi fattoriale delle corrispondenze
  - Multidimensional scaling (MDS)
- Analisi di correlazione
  - Analisi di correlazione semplice
  - Analisi di correlazione canonica
- Metodi di ordinamento e di classificazione
  - Cluster analysis
  - Analisi discriminante
- Analisi di regressione multipla
- Modelli di regressione non lineari
  - Generalized linear – non linear models (GLZ)
  - Modelli di regressione addittivi
  - Generalized Additive Models (GAM)

Modelli di regressione non lineari

- **Generalized Linear – Nonlinear Models (GLZ)**
  - utili quando la relazione tra la variabile dipendente e i predittori è chiaramente non lineare.
- I valori della variabile dipendente sono predetti da una *combinazione lineare di predittori*, legata alla variabile dipendente da una funzione non più lineare, ma la cui forma è decisa a priori:

$$Y=f(a+b_1X_1+b_2X_2+...+b_pX_p)$$

- Procedura
  - Scelta della forma della relazione che si vuole testare
  - Stima dei parametri della funzione, ottimizzando l'adattamento ai dati tramite iterazioni successive.
  - Regole per fermare l'iterazione:
    - soglia minima di variazione dell'R2 ad ogni passaggio,
    - numero massimo di iterazioni.
- ☺ non richiedono distribuzione normale della variabile dipendente
- non pongono vincoli sulla forma della relazione da descrivere.

Modelli di regressione non lineari

- **Modelli di regressione addittivi**:
- Si differenziano dai modelli standard poiché la relazione lineare ricercata non coinvolge direttamente i predittori, ma una loro funzione.

- Per ogni predittore si stima non il coefficiente di regressione ma la funzione, che entra nella regressione lineare tramite una procedura iterativa:

$$y = \sum f_i(x_i)$$

Modelli di regressione non lineari

- **Generalized Additive Models (GAM):**
- Combinazione dei due metodi: massimizza la qualità della previsione, stimando le funzioni *non parametriche* di ogni predittore collegate alla variabile dipendente da una funzione *g*:

$$y = g(\sum f_i(x_i))$$

- ☺ flessibilità: non pone nessun vincolo alle variabili e alle relazioni esistenti tra esse.
- ☹ Proprio per questo motivo, particolare attenzione va posta nella loro costruzione:
  - limitare il numero di variabili inserite nel modello,
  - valutare se il passaggio da un modello di regressione lineare ai modelli generalizzati migliora la robustezza del modello, per non aggiungere complessità inutile ai fini dell'interpretazione.

Contenuti – metodi statistici

Analisi esplorativa di dati multidimensionali

- Caratteristiche
- Metodi fattoriali di tipo esplorativo
  - Analisi fattoriale
  - Analisi delle componenti principali
  - Analisi fattoriale delle corrispondenze
  - Multidimensional scaling (MDS)
- Analisi di correlazione
  - Analisi di correlazione semplice
  - Analisi di correlazione canonica
- Metodi di ordinamento e di classificazione
  - Cluster analysis
  - Analisi discriminante
- Analisi di regressione multipla
- Modelli di regressione non lineari
  - Generalized linear – non linear models (GLZ)
  - Modelli di regressione addittivi
  - Generalized Additive Models (GAM)

Analisi di Correlazione

Correlazione lineare semplice

- **Pearson r correlation**
- Correlazione: misura della relazione tra due o più variabili.
- Scala di misura almeno a intervalli.
- Il coefficiente di correlazione r rappresenta la relazione lineare tra due variabili (retta di regressione o dei minimi quadrati).
- Elevato al quadrato (r^2) rappresenta la proporzione di varianza in comune tra le due variabili, ovvero, la forza della relazione.

Analisi di Correlazione

Analisi della correlazione canonica

- Il metodo è considerato una generalizzazione della **regressione multipla**.
- Variabili: quantitative, standardizzate
- Scopo: Ricerca le relazioni tra due set di variabili
  - permette di ridurre a un numero limitato di parametri l'analisi delle interdipendenze tra due insiemi di variabili;
  - Valuta il tipo interdipendenza esistente tra gli insiemi di variabili se è ritenuta significativa
- Esempio di applicazione
  - Matrice di composizione in specie & Matrice di descrittori ambientali
  - Chimica delle acque superficiali & caratteristiche geomorfologiche del bacino

Analisi di Correlazione
Analisi della correlazione canonica

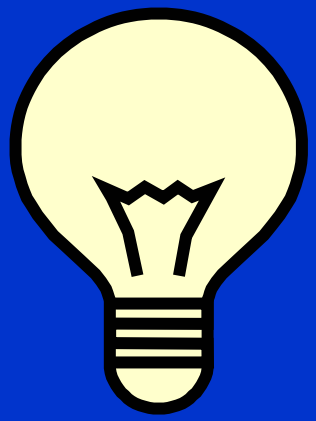
- Per collegare l'informazione di due set di variabili, si ricercano una serie di trasformate dette *radici canoniche*:
  - somme pesate dei loro elementi: a1 y1+…ap yp=b1 x1+…bq xq.
  - Si impone la condizione che le due somme pesate abbiano il valore massimo di correlazione
- Numero di radici estraibili = numero minimo di variabili in ciascuno dei due insiemi.
- La proporzione di varianza spiegata dalla correlazione tra le radici canoniche è descritta dagli autovalori.
  - Ogni radice successiva spiegherà quote sempre più piccole di variabilità.
- I pesi canonici permettono di comprendere in che modo ogni variabile di ciascun insieme contribuisce in maniera unica alla definizione della rispettiva radice canonica e di conseguenza alla relazione tra gli insiemi.

Analisi di Correlazione
Analisi della correlazione

- Esempio di applicazione: relazione tra i parametri in ingresso nelle acque superficiali, sono evidenziate quelle significative ($p < 0,5$).

Acqua in ingresso	pH	Ec	SS	$P_{tot.}$	P_{sol}	$N\text{-}NH_4$	$N\text{-}NO_3$	$N\text{-}NO_2$	N_{org}
pH	1,00	-0,11	-0,10	-0,03	-0,05	*0,27*	*0,44*	0,17	-0,04
Ec	-0,11	1,00	0,12	0,16	0,13	0,24	0,01	0,08	*0,48*
SS	-0,10	0,12	1,00	0,18	0,06	0,11	0,11	0,24	0,15
$P_{tot.}$	-0,03	0,16	0,18	1,00	*0,53*	*0,33*	*0,27*	*0,45*	0,18
$P_{sol.}$	-0,05	0,13	0,06	*0,53*	1,00	0,19	0,21	0,14	-0,03
$N\text{-}NH_4$	*0,27*	0,24	0,11	*0,33*	0,19	1,00	*0,57*	*0,38*	-0,06
$N\text{-}NO_3$	*0,44*	0,01	0,11	*0,27*	0,21	*0,57*	1,00	*0,42*	*-0,29*
$N\text{-}NO_2$	0,17	0,08	0,24	*0,45*	0,14	*0,38*	*0,42*	1,00	0,17
N_{org}	-0,04	*0,48*	0,15	0,18	-0,03	-0,06	*-0,29*	0,17	1,00

Analisi di Correlazione o Analisi di regressione?



Analisi di correlazione:

Per misurare l'intensità dell'associazione tra due variabili quantitative che variano congiuntamente senza che tra essa esista una relazione diretta di causa-effetto.

Analisi di regressione:

Per ricavare un modello statistico dai dati campionari che predica i valori di una variabile Y detta dipendente, individuata come effetto a partire dai valori dell'altra variabile X, detta indipendente, individuata come causa.

Contenuti

- Terminologia
- Concetti di base
- Metodi statistici
  - Confronto set di dati sperimentali
  - Analisi esplorativa di dati multidimensionali
- Interpretazione dei risultati

Interpretazione dei risultati

- I **risultati** ottenuti con metodi di analisi statistica **non sono assoluti**, ma devono essere valutati in funzione delle conoscenze pregresse relative all'argomento (e.g. Brotons *et al.*, 2003; Luoto *et al.*, 2002; Mac Nally *et al.*, 2000;Tufford *et al.*, 1998)
- Approccio "*ecologically focused*" (Luoto *et al.*, 2002):
  - Nella costruzione di modelli interpretativi di fenomeni ambientali è necessario considerare, al di la della robustezza e della significatività statistica, il significato ecologico delle relazioni individuate
  - Le relazioni tra i predittori e le variabili dipendenti indagate devono essere ecologicamente e plausibilmente interpretabili.
- L'analisi statistica dei dati è uno **strumento** di analisi: non è possibile scindere le analisi dalla loro interpretazione.

L'approccio *ecologically focused*
Costruzione di modelli

- L'applicazione di questo approccio può avvenire sia nella fase di costruzione dei modelli sia in quella di valutazione dei risultati ottenuti attraverso l'analisi statistica.
- Nella costruzione di modelli interpretativi *ecologically focused*, particolare attenzione viene posta nella plausibilità ecologica delle relazioni tra le variabili di risposta e i predittori.
- I modelli vengono costruiti inserendo forzatamente alcuni predittori di cui è nota la capacità di influenzare i fenomeni in esame (ad esempio dall'esame di studi precedenti).

L'approccio *ecologically focused*
Esempio di applicazione_1

- Hamer et al., 2006
  - Lo studio vuole esaminare il grado con cui alcuni fattori determinano il pattern osservato di ricchezza di specie di uccelli nell'est del Wyoming.
  - La scelta delle variabili predittive è stata effettuata in modo da riflettere il processo gerarchico di selezione degli habitat da parte degli individui e i suoi effetti sulla distribuzione della diversità: si è ipotizzato che gli uccelli selezionino l'habitat valutando in primo luogo l'estensione, in secondo luogo la struttura, in terzo luogo le caratteristiche della matrice e infine la disponibilità di prede, in un modo tale che i requisiti a scala maggiore agiscano da filtro o da costrizione per quelli a scala più piccola.
  - Le variabili indipendenti sono state quindi inserite progressivamente nella regressione: ad ogni step sono stati individuati una serie di modelli adatti a spiegare la relazione tra le variabili, generalmente analizzando esempi di letteratura e/o sulla base del valore di R^2.

L'approccio *ecologically focused*
Esempio di applicazione_2

- Chapa-Vargas e Robinson, 2006
  - Lo studio si pone come obiettivo di esaminare gli effetti di diverse tipologie ed estensioni di habitat, della struttura della vegetazione, delle distanze dai bordi e delle tipologie di questi, sul successo in termini di sopravvivenza dei nidi e di parassitismo sulla covata per una specie di uccello canterino in Illinois.
  - Sono state individuate una serie di ipotesi su quali fattori influenzino le due variabili dipendenti selezionate (ossia la sopravvivenza della nidiata e il parassitismo su di essa).
  - L'inclusione di questi fattori nei modelli di previsione si è basata sull'esame di lavori precedenti e sull'analisi esplorativa (correlazione parziale tra variabili indipendenti e variabili dipendenti e tra ipotesi e variabili indipendenti).

L'approccio *ecologically focused*

- Abbiamo visto che è possibile operare secondo l'approccio *ecologically focused* nella fase di selezione delle variabili, cioè **PRIMA** di effettuare l'analisi statistica con cui possiamo individuare le relazioni esistenti tra predittori e variabili dipendenti.
- Tuttavia è possibile usare questo approccio anche **DOPO** aver effettuato le analisi statistiche, ossia nella fase in cui siamo chiamati ad operare una scelta tra i diversi modelli in grado di descrivere il processo in esame.

L'approccio *ecologically focused*
Selezione del modello interpretativo

- Una volta individuati i modelli che sono in grado di spiegare i dati, spesso ci si trova di fronte alla necessità di scegliere quale sia il più "adatto".
- Considerati il numero limitato di dati e gli errori ad essi associati, non esiste un modo per sapere con sicurezza quale sia il modello "migliore".
- Esistono tuttavia dei metodi in grado di fornire delle risposte in termini di probabilità.
- L'applicazione di questi metodi non deve prescindere dall'esame della significatività ecologica delle relazioni individuate.

L'approccio *ecologically focused* deve accompagnare l'applicazione di tali strumenti.

Selezione del modello interpretativo

- Tra gli strumenti in grado di operare un confronto tra modelli è stato selezionato

  l' **Akaike Information Criterion** (A.I.C.)

- Perché?
  - è innovativo

    L'approccio non si basa sul tradizionale paradigma statistico dell'*hypothesis testing*, ma sulla teoria dell'informazione.
  - è intuitivo

    I calcoli si eseguono senza difficoltà e i risultati sono di facile interpretazione.
  - è diffuso

    Sono in continuo aumento gli studi in cui viene utilizzato (e.g. Westphal *et al*.,2003; Wiersma *et al*., 2004; Chapa-Vargas & Robinson, 2006; Hamer *et al*., 2006; Koper & Schmiegelow, 2006)

L'Akaike Information Criterion

- Il criterio informativo di Akaike (A.I.C.) è uno strumento che consente di confrontare diversi modelli fra loro e di determinare quale tra essi è più verosimilmente corretto, quantificando questa verosimiglianza.
- Le basi teoretiche del metodo risiedono principalmente nella teoria dell'informazione e non sono di facile comprensione (per approfondimenti si veda di Burnham & Anderson "*Model selection and multimodel inference: a practical information-theoretic approach*" Springer, NY, 1998).
- A.I.C è uno strumento che permette di selezionare il più robusto in maniera "oggettiva" ed indipendente dalla semplice robustezza statistica (es. R^2 delle regressioni).
- A differenza di altri metodi non calcola un *p-level*, non giunge a conclusioni riguardanti la significatività statistica e non rigetta nessun modello.

L'Akaike Information Criterion

- L'utilizzo di questo strumento parte dall'assunto che l'applicazione di alcuni modelli su un set di dati può essere riassunta da un criterio informativo, appunto quello sviluppato da Akaike:

$$A.I.C. = N \cdot \ln\left(\frac{SS}{N}\right) + 2K$$

dove N è la numerosità del campione, K è il numero di parametri e SS è la somma dei quadrati delle distanze verticali dei punti dalla retta.

- Il metodo non si basa sul singolo valore assunto dal criterio per un determinato modello, ma sul confronto dei valori di A.I.C. corrispondenti a modelli diversi tra loro per numero di variabili.

L'Akaike Information Criterion
Confronto di due modelli

- Si applichi il criterio per confrontare due modelli diversi.
- Detto A il modello più semplice e B quello più complesso (quello con un maggior numero di parametri), si ha:

$$\Delta A.I.C. = A.I.C._B - A.I.C._A$$

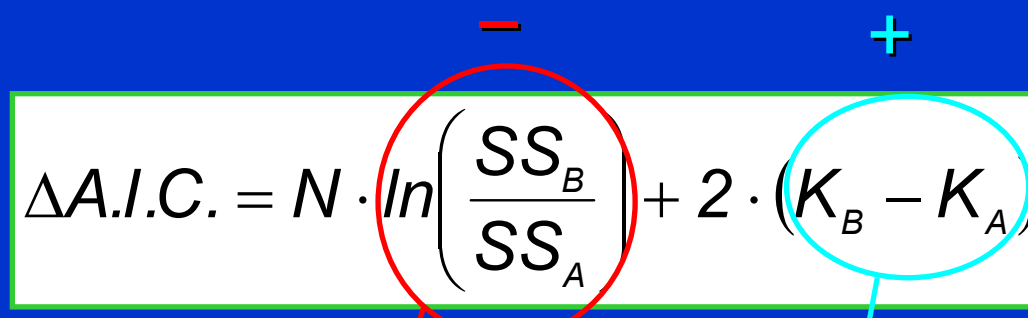
da cui, riprendendo l'equazione precedente:

$$\Delta A.I.C. = N \cdot ln\left(\frac{SS_B}{SS_A}\right) + 2 \cdot (K_B - K_A)$$

- Il valore di A.I.C. bilancia il cambiamento nella bontà della regressione (valutato attraverso la somma dei quadrati) con il cambiamento nel numero dei parametri.

L'Akaike Information Criterion
Confronto di due modelli

- Essendo il modello A quello più semplice, la sua applicazione ai dati sarà quasi sempre peggiore, quindi SS_A sarà maggiore di SS_B; poiché il logaritmo di una frazione è sempre negativo, ne risulta che il primo termine dell'equazione sarà negativo.
- Il modello B ha più parametri, quindi K_B è maggiore di K_A: questo fa sì che il secondo termine dell'equazione sia positivo.



$$SS_A > SS_B$$

$$\frac{SS_B}{SS_A} < 1$$

$$\ln\left(\frac{SS_B}{SS_A}\right) < 0$$

$$K_B > K_A$$

$$(K_B - K_A) > 0$$

L'Akaike Information Criterion
Confronto di due modelli

- Se il risultato complessivo è negativo, significa che la differenza nelle SS è maggiore di quella che ci si aspetta basandosi sulla differenza nel numero di parametri: dunque si può concludere che il modello B (quello più complesso) è più probabile. Se la differenza nel valore di A.I.C. è positiva, allora il cambiamento nelle SS non è così ampio come ci si aspetta dal cambiamento nel numero dei parametri; ciò significa che c'è una maggior probabilità che i dati siano meglio interpretati dal modello A (quello più semplice).
- Riassumendo:

valore di ΔA.I.C.	scelta
ΔA.I.C.>0	modello A
ΔA.I.C.<0	modello B

L'Akaike Information Criterion

- Nel caso in cui ci si trovi a dover confrontare più di due modelli, una volta calcolati i singoli valori di A.I.C., si può procedere alla scelta del modello che presenta il valore più basso: questo modello è quello che presenta la maggior probabilità di essere corretto.
- I singoli valori assunti da A.I.C. per i diversi modelli possono essere positivi o negativi (il segno dipende dall'unità di misura con cui si è deciso di esprimere i dati), ma questo non ci dà alcuna informazione: la sola cosa che si deve considerare è la DIFFERENZA tra i valori.

L'Akaike Information Criterion
A.I.C. corretto

- Nel caso in cui N sia piccolo al confronto con il valore di K, è preferibile adottare un valore di A.I.C. "corretto" ($A.I.C._C$), che viene calcolato con l'equazione seguente:

$$A.I.C._c = A.I.C. + \frac{2K(K+1)}{N-K-1}$$

- Nelle condizioni sopra indicate, il valore di $A.I.C._C$ risulta più accurato di quello di A.I.C. e consente di effettuare il confronto in maniera più agevole; proprio per questo si consiglia di usare sempre il valore corretto.

[Si tenga presente che $A.I.C._C$ può essere calcolato solo se il numero di dati è almeno due volte il numero dei parametri.]

L'Akaike Information Criterion
Applicazione

- Recentemente molti autori hanno fatto ricorso a questo strumento per selezionare i modelli che con maggior probabilità sono in grado di interpretare correttamente i dati (Westphal *et al.* 2003, Wiersma *et al.* 2004, Chapa-Vargas e Robinson 2006, Hamer *et al.* 2006, Koper e Schmiegelow 2006); in tutti questi lavori è stato utilizzato l'Akaike Information Criterion corretto ($A.I.C._C$) e, una volta selezionato il modello "migliore" (quello con $A.I.C._C$ più basso), sono stati individuati come eventuali modelli competitivi quelli con $\Delta A.I.C._C<2$.

L'Akaike Information Criterion
Applicazione

- Alcuni autori, selezionati i modelli che presentano la maggior probabilità di essere corretti, determinano il valore di questa probabilità attraverso un parametro, w_i, che rappresenta la probabilità del modello i-esimo di essere il migliore in un set di modelli. Le probabilità calcolate vengono chiamate i "pesi" di Akaike e risultano particolarmente utili quando i valori di $A.I.C._C$ sono molto vicini e non si può scegliere con sicurezza un modello piuttosto che l'altro.

L'Akaike Information Criterion
Vantaggi

- consente di lavorare con ipotesi multiple poiché confronta l'intero set di modelli, non limitandosi ad un confronto a coppie; risulta perciò particolarmente utile negli studi ecologici, dove si ha la necessità di confrontare diverse ipotesi tra loro;

L'Akaike Information Criterion
Vantaggi

- permette non solo di selezionare il modello che presenta la maggiore probabilità di essere corretto, ma di quantificare questa probabilità;
- è in grado di operare confronti tra i modelli senza essere influenzato dall'intercorrelazione tra i predittori; questo rende possibile il suo utilizzo per determinare quale tra un insieme di modelli annidati (*nested*) è più adatto ad interpretare i dati.

[Due modelli sono detti annidati quando uno rappresenta il caso più semplice dell'altro; ad esempio un modello cinetico ad una fase è un caso più semplice dello stesso modello a più fasi.]

L'Akaike Information Criterion
Esempio di applicazione

- Studio della relazione tra organizzazione strutturale del paesaggio e qualità delle acque nel bacino scolante della laguna di Venezia.
- Sono stati selezionati alcuni modelli sulla base della significatività statistica (p-level<0.1);
- Per individuare i modelli "migliori" è stato applicato il criterio di Akaike attraverso un approccio che tenesse comunque in considerazione la significatività ecologica dei modelli selezionati.

SCALA	VARIABILI	AIC_C	ΔAIC xnum variabili	ΔAIC x scala	ΔAIC tra scale
Bacino	N	-22,15	4,70	4,70	5,32
	S	-26,85	**0,00**	**0,00**	**0,61**
	ZOO	-23,26	3,59	3,59	4,20
	SHA	-25,70	**1,15**	**1,15**	**1,76**
	AR	-15,50	11,36	11,36	11,97
	F1	-19,39	7,46	7,46	8,08
	N-AR	-23,67	3,19	3,19	3,80
100m	AR	-14,48	12,99	12,99	12,99
	N	-19,33	8,14	8,14	8,14
	S	-27,47	**0,00**	**0,00**	**0,00**
	ZOO	-17,98	9,49	9,49	9,49
	SHA	-22,00	5,47	5,47	5,47
	F1	-19,07	8,40	8,40	8,40
	SHA-F1	-23,44		4,03	4,03
	N-AR-SHA	-19,62		7,85	7,85
50m	N	-18,86	6,83	6,83	8,61
	S	-25,69	**0,00**	**0,00**	**1,78**
	F1	-19,31	6,38	6,38	8,16
	N-AR	-20,33	**0,54**	5,36	7,14
	N-IP	-20,31	**0,56**	5,38	7,16
	ZOO-F1	-20,87	**0,00**	4,82	6,60
	N-AR-ZOO	-24,93		**0,76**	2,54
MIN		-27,47			