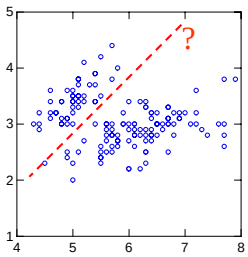
Classificazione
(aka Cluster Analysis)

- Classificazione non gerarchica
  - es. *k*-means
- Classificazione gerarchica divisiva
- Classificazione gerarchica agglomerativa
  - Legame: singolo, completo, medio, …
  - Coefficiente di correlazione cofenetica
- Classificazione con vincoli
- Metodi innovativi (machine learning)
  - es. Self-organizing maps

Classificazione



(= Cluster Analysis)

Obiettivo:
massimizzare l'omogeneità dei gruppi (o classi o clusters)
(cioè: gli oggetti simili devono essere nello stesso cluster)

Problema generale:
cercare le discontinuità nello spazio dei dati da classificare

Discontinuità

Reali o "naturali":

es. tassonomia

Arbitrarie:

es. ecologia delle comunità

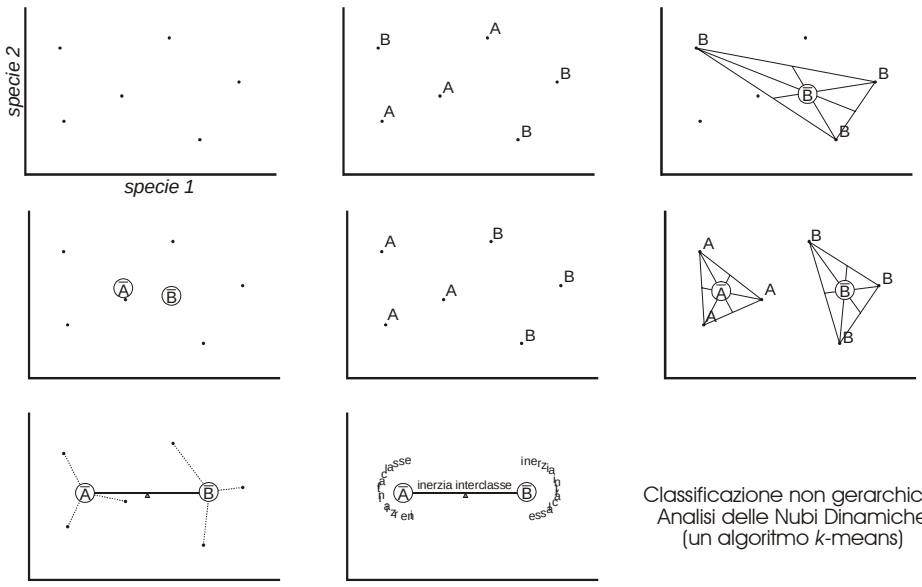
Clustering non-gerarchico: *k*-means

- Si utilizza direttamente la matrice dei dati calcolando distanze euclidee
- Si massimizza la varianza inter-classe per un numero dato *a priori* di classi
- In pratica, equivale a cercare le classi che massimizzano il rapporto F di una MANOVA a una via

k-means

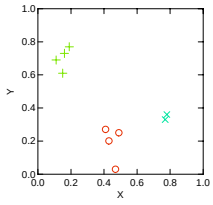
Procedura iterativa:

1. Scegli un numero di classi
2. Assegna gli oggetti alle classi
   (a caso o in base ad un'altra classificazione)
3. Sposta gli oggetti nelle classi il cui centroide è più vicino (la varianza intra-classe diminuisce)
4. Ripeti lo step 3 finchè non c'è più nessun cambiamento nella composizione delle classi



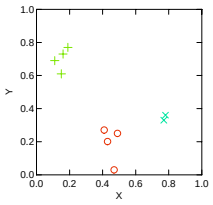
Classificazione non gerarchica:
Analisi delle Nubi Dinamiche
(un algoritmo *k*-means)

k-means con 3 classi



Variabile	InterSQ	gdl	IntraSQ	gdl	rapporto F
X	0.536	2	0.007	7	256.163
Y	0.541	2	0.050	7	37.566
** TOTALE **	1.078	4	0.058	14	

k-means con 3 classi



Classe 1 di 3 (contiene 4 casi)

Membri			Statistiche				
Caso	Distanza	Variabile	Min	Media	Max	Dev.St.
Caso 1	0.02	X	0.41	0.45	0.49	0.04
Caso 2	0.11	Y	0.03	0.19	0.27	0.11
Caso 3	0.06					
Caso 4	0.05					

Classe 2 di 3 (contiene 4 casi)

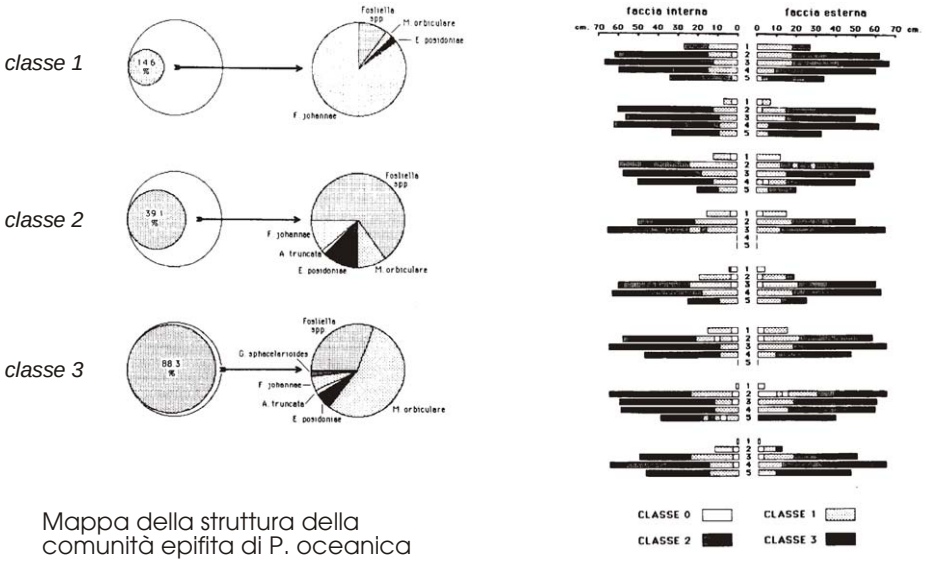
Membri			Statistiche			
Caso	Distanza	Variabile	Min	Media	Max	Dev.St.
Caso 7	0.06	X	0.11	0.15	0.19	0.03
Caso 8	0.03	Y	0.61	0.70	0.77	0.07
Caso 9	0.02					
Caso 10	0.06					

Classe 3 di 3 (contiene 2 casi)

Membri			Statistiche			
Caso	Distanza	Variabile	Min	Media	Max	Dev.St.
Caso 5	0.01	X	0.77	0.77	0.78	0.01
Caso 6	0.01	Y	0.33	0.35	0.36	0.02

Svantaggi della classificazione *k*-means

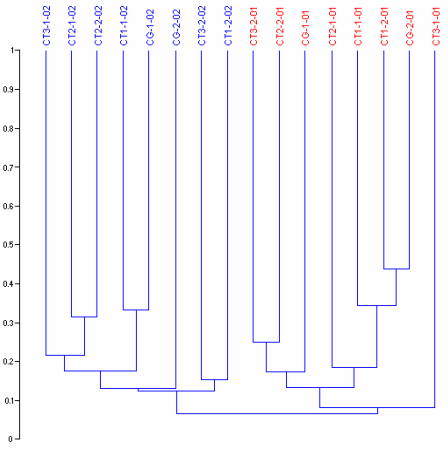
- Ottimizza le classi, ma non in modo globale (iterando si può ottenere una partizione più efficace)
- Bisogna scegliere il numero di classi *a priori*
  - Ma quante classi devono essere scelte?



Mappa della struttura della comunità epifita di P. oceanica

Classificazione gerarchica

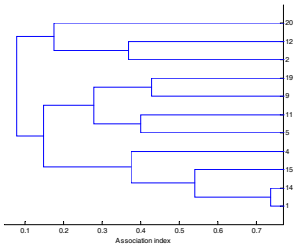
- Solitamente si rappresenta con un dendrogramma



Classificazione gerarchica

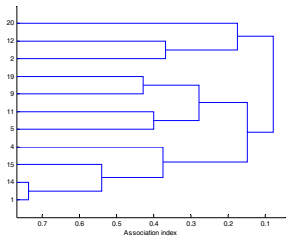
- Divisiva
  - Si cercano partizioni successive di un insieme di oggetti in due sotto-insiemi in modo da soddisfare un criterio predefinito
- Agglomerativa
  - Si aggregano gli oggetti in gruppi ed i gruppi fra loro secondo un criterio predefinito

Classificazione gerarchica divisiva



- Si inizia con tutti gli oggetti in una sola classe, poi:
  - si dividono mediante classificazioni k-means o altre tecniche divisive con partizioni in due classi;
  - si divide ogni volta la classe più eterogenea…
  - Oppure si dividono simultaneamente tutte le classi esistenti.
- Si tratta di tecniche efficienti, ma poco usate!

Classificazione gerarchica agglomerativa



- Si inizia con un oggetto per classe
- Le classi vengono fuse finchè ne rimane una soltanto
- E' la classe di metodi di gran lunga più ampiamente diffusa

Classificazione gerarchica agglomerativa

- Si lavora su una matrice di dissimilarità o distanza
- Procedura generale:
  1. Si calcola una matrice simmetrica di dissimilarità/distanza fra le classi
  2. Le classi fra cui la dissimilarità/distanza è minima vengono fuse fra loro
  3. Si calcola la dissimilarità/distanza fra la nuova classe ottenuta e tutte le altre

  N.B. I diversi algoritmi differiscono nel punto 3

Classificazione gerarchica agglomerativa

	A	B	C	D	E
A	0	.	.	.	.
B	0.35	0	.	.	.
C	0.45	0.67	0	.	.
D	0.11	0.45	0.57	0	.
E	0.22	0.56	0.78	0.19	0

Prima fusione: A *e* D

	AD	B	C	E
AD	0	.	.	.
B	?	0	.	.
C	?	0.67	0	.
E	?	0.56	0.78	0

Come calcolare le nuove distanze?

Classificazione gerarchica agglomerativa
Legame semplice

	A	B	C	D	E
A	0	.	.	.	.
B	0.35	0	.	.	.
C	0.45	0.67	0	.	.
D	0.11	0.45	0.57	0	.
E	0.22	0.56	0.78	0.19	0

	AD	B	C	E
AD	0	.	.	.
B	0.35	0	.	.
C	?	0.67	0	.
E	?	0.56	0.78	0

d(AD,B)=*Min*{**d**(A,B), **d**(D,B)}

Classificazione gerarchica agglomerativa
Legame completo

	A	B	C	D	E
A	0	.	.	.	.
B	0.35	0	.	.	.
C	0.45	0.67	0	.	.
D	0.11	0.45	0.57	0	.
E	0.22	0.56	0.78	0.19	0

	AD	B	C	E
AD	0	.	.	.
B	0.45	0	.	.
C	?	0.67	0	.
E	?	0.56	0.78	0

d(AD,B)=*Max*{**d**(A,B), **d**(D,B)}

Classificazione gerarchica agglomerativa
Legame intermedio

	A	B	C	D	E
A	0	.	.	.	.
B	0.35	0	.	.	.
C	0.45	0.67	0	.	.
D	0.11	0.45	0.57	0	.
E	0.22	0.56	0.78	0.19	0

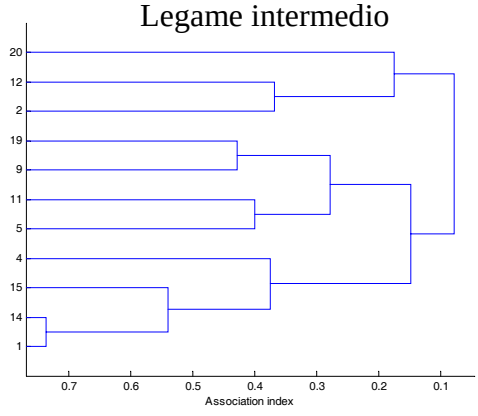
	AD	B	C	E
AD	0	.	.	.
B	0.40	0	.	.
C	?	0.67	0	.
E	?	0.56	0.78	0

d(AD,B)=*Mean*{**d**(A,B), **d**(D,B)}

Classificazione gerarchica agglomerativa
Metodo di Ward

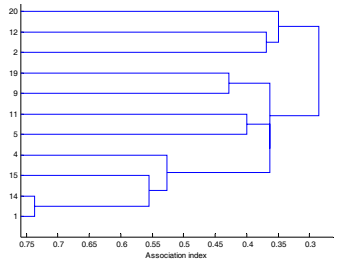
- Minimizza la varianza intra-classe
- Affine ad algortimi basati sul calcolo delle distanze dai centroidi di classe (come nel caso dell'algoritmo *k*-means)

Legame intermedio

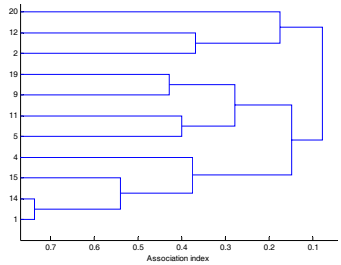


1	1.00									
2	0.00	1.00								
4	0.53	0.00	1.00							
5	0.18	0.05	0.00	1.00						
9	0.22	0.09	0.13	0.25	1.00					
11	0.36	0.00	0.17	0.40	0.33	1.00				
12	0.00	0.37	0.18	0.00	0.13	0.00	1.00			
14	0.74	0.00	0.30	0.20	0.23	0.17	0.00	1.00		
15	0.53	0.00	0.30	0.00	0.36	0.00	0.26	0.56	1.00	
19	0.00	0.00	0.17	0.21	0.43	0.32	0.29	0.09	0.09	1.00
20	0.04	0.00	0.17	0.00	0.14	0.10	0.35	0.00	0.18	0.25	1.00
	1	2	4	5	9	11	12	14	15	19	20

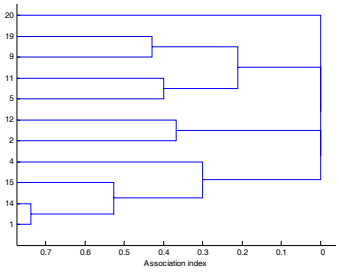
Legame singolo



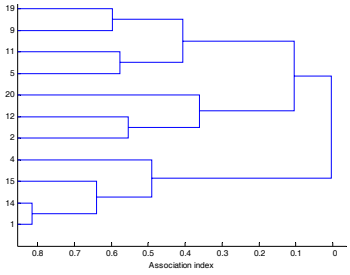
Legame intermedio



Legame completo

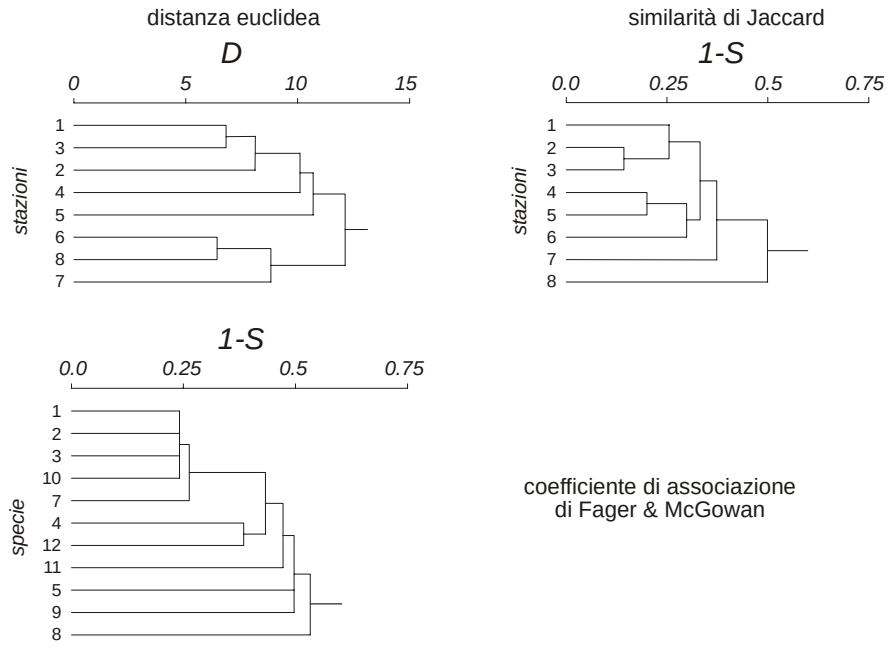


Algoritmo di Ward



Metodi di classificazione gerarchica agglomerativa

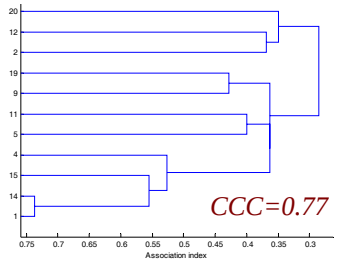
- Legame semplice
  - Produce “catene” di oggetti, “contrae” lo spazio intorno alle classi
  - Non appropriato se l’errore di campionamento è grande
  - Usato in tassonomia
  - Invariante rispetto alle trasformazioni dei dati
- Legame completo
  - Produce classi compatte e “dilata” lo spazio intorno a esse
  - Non appropriato se l’errore di campionamento è grande
  - Invariante rispetto alle trasformazioni dei dati
- Legame intermedio, centroide, di Ward
  - Maggiore probabilità di riconoscere partizioni “naturali”
  - Non invarianti rispetto alle trasformazioni dei dati



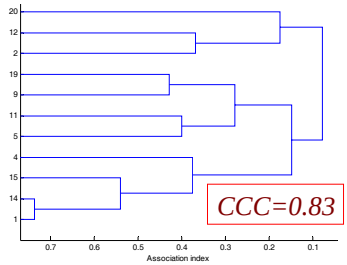
Coefficiente di correlazione cofenetica (CCC)

- Correlazione fra la matrice di dissimilarità originale e quella inferita in base alla classificazione
- CCC >~ 0.8 indica un buon accordo
- CCC <~ 0.8 indica che il dendrogramma non è una buona rappresentazione delle relazioni fra oggetti
- Si può usare il CCC per scegliere l'algoritmo ottimale

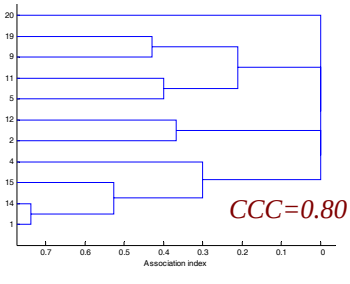
Legame singolo



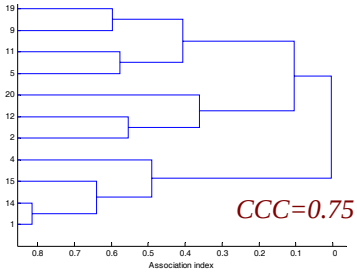
Legame intermedio



Legame completo



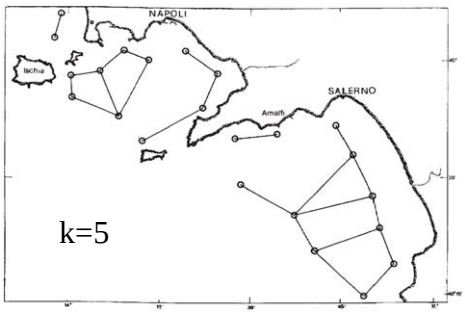
Algoritmo di Ward



Classificazione vincolata

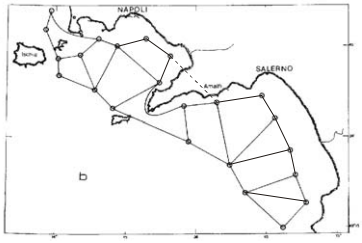
- Si definisce un vincolo di natura spaziale (es. contiguità fra oggetti) o di natura temporale (es. successione)
- Si definisce un metodo per verificare che il vincolo sia rispettato (es. rete di Gabriel per rappresentare la contiguità)
- Si applica il vincolo ad un qualsiasi algoritmo di classificazione
- Le classi ottenute possono essere più vicine alla nostra percezione della realtà
- Il fuoco si sposta dalla composizone delle classi alle discontinuità fra di esse

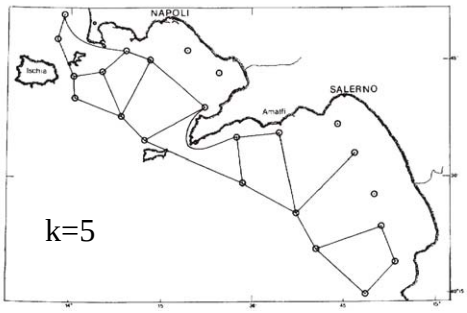
Distanza euclidea
(dati quantitativi)



k=5

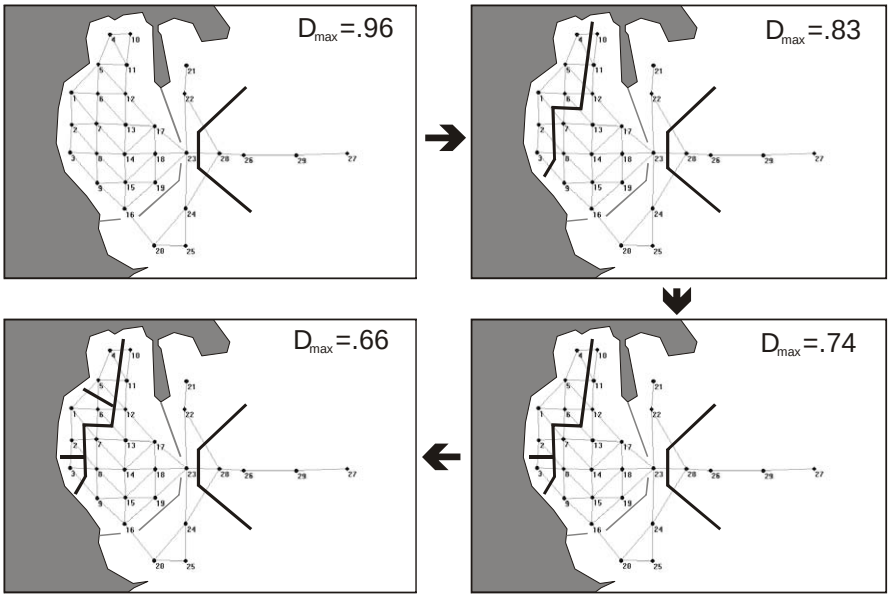
Rete di Gabriel





k=5

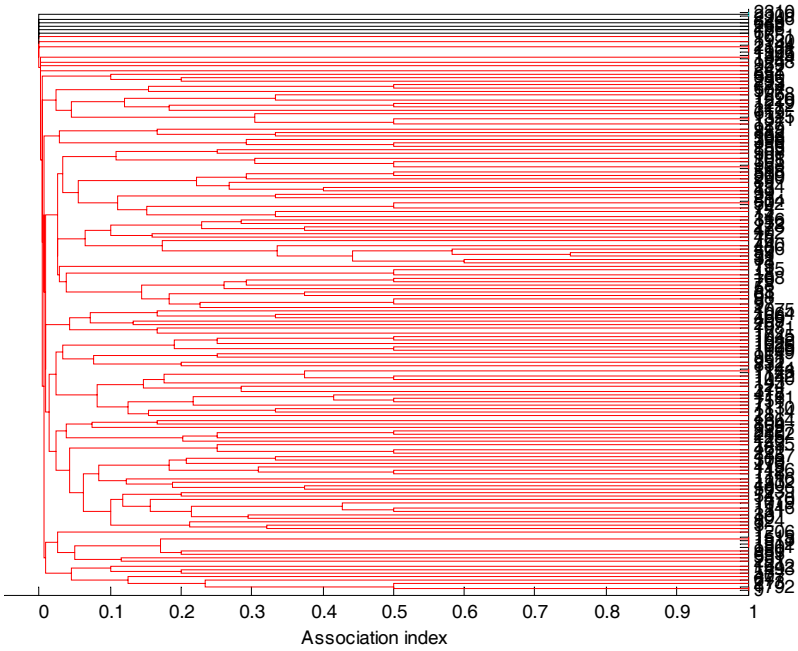
Dissimilarità di Jaccard
(dati binari)



dissimilarità di Jaccard

Problemi aperti

- Esistono realmente classi biologicamente o ecologicamente rilevanti?
- Il dendrogramma rappresenta la realtà biologica (web-of-life vs. tree-of-life)?
- Quante classi usare?
  - le regole per la selezione sono arbitrarie!
- Quale metodo usare?
  - la tecnica ottimale dipende dai dati!
- I dendrogrammi non sono efficienti nel visualizzare classificazioni complesse…

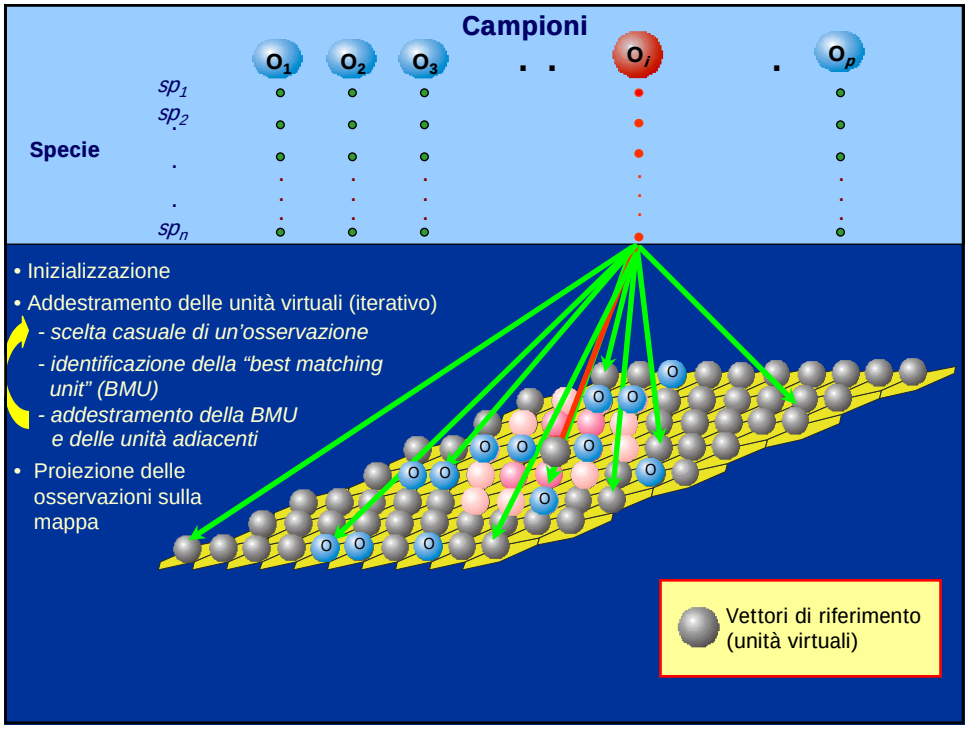


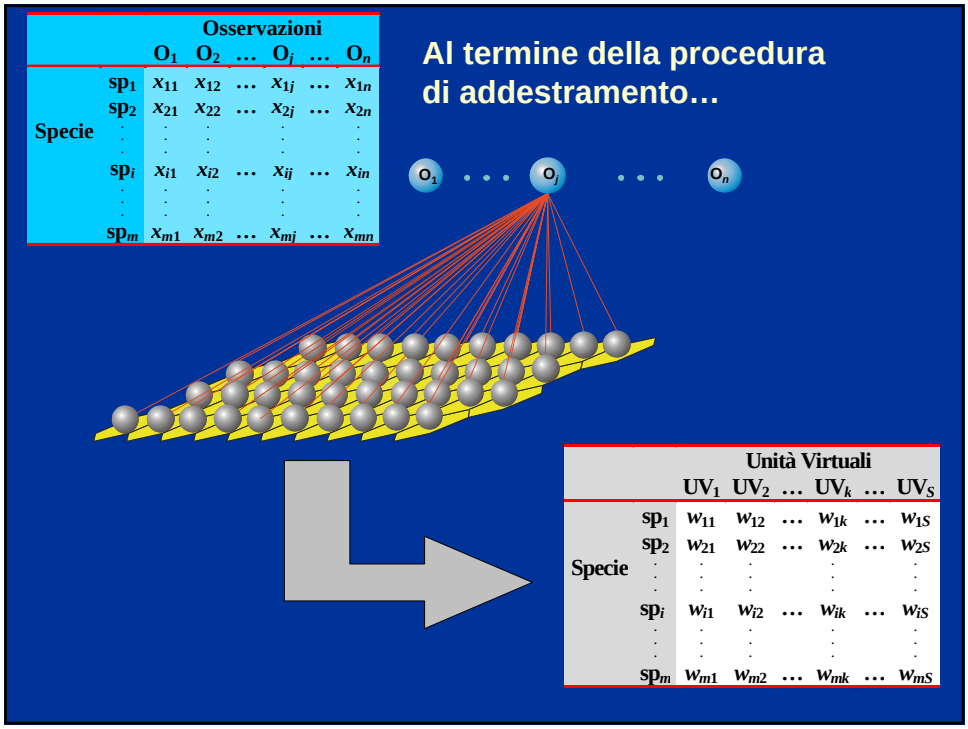
C’è qualcosa di meglio per casi complessi?

- Esistono metodi di nuova generazione, che non sono ancora molto diffusi in Ecologia
- La loro diffusione sta crescendo, ma bisogna che siano accettati dagli ecologi
- In molti casi, sfruttano le capacità di calcolo attualmente disponibili per tutti
- Si tratta soprattutto di tecniche mutuate dal campo del Machine Learning, dell’Intelligenza Artificiale e del Data Mining

Ordinamento e classificazione:

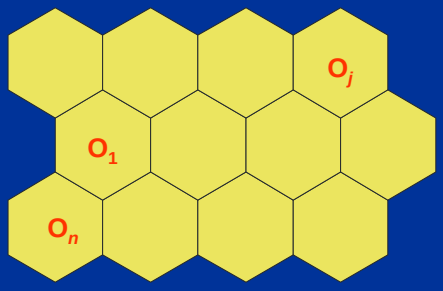
Self-Organizing Maps (SOMs) o Mappe di Kohonen


Campioni
O_1 O_2 O_3 . . O_i . O_p
Specie
sp_1
sp_2
sp_n
• Inizializzazione
• Addestramento delle unità virtuali (iterativo)
- scelta casuale di un'osservazione
- identificazione della "best matching unit" (BMU)
- addestramento della BMU e delle unità adiacenti
• Proiezione delle osservazioni sulla mappa
Vettori di riferimento (unità virtuali)


Osservazioni
O_1 O_2 … O_j … O_n
Specie
sp_1 x_{11} x_{12} … x_{1j} … x_{1n}
sp_2 x_{21} x_{22} … x_{2j} … x_{2n}
sp_i x_{i1} x_{i2} … x_{ij} … x_{in}
sp_m x_{m1} x_{m2} … x_{mj} … x_{mn}
Al termine della procedura di addestramento…
O_1 O_j O_n
Unità Virtuali
UV_1 UV_2 … UV_k … UV_S
Specie
sp_1 w_{11} w_{12} … w_{1k} … w_{1S}
sp_2 w_{21} w_{22} … w_{2k} … w_{2S}
sp_i w_{i1} w_{i2} … w_{ik} … w_{iS}
sp_m w_{m1} w_{m2} … w_{mk} … w_{mS}

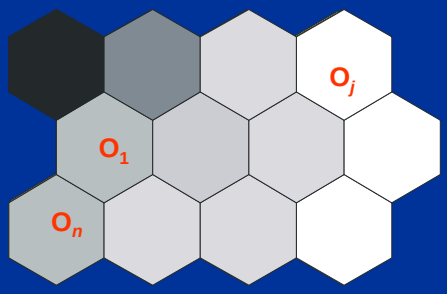
- Visualizzazione delle osservazioni

		Osservazioni					
		O_1	O_2	…	O_j	…	O_n
Specie	sp_1	x_{11}	x_{12}	…	x_{1j}	…	x_{1n}
	sp_2	x_{21}	x_{22}	…	x_{2j}	…	x_{2n}
	⋮	⋮	⋮		⋮		⋮
	sp_i	x_{i1}	x_{i2}	…	x_{ij}	…	x_{in}
	⋮	⋮	⋮		⋮		⋮
	sp_m	x_{m1}	x_{m2}	…	x_{mj}	…	x_{mn}



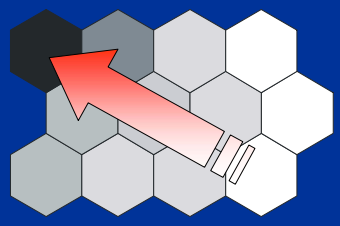
- Visualizzazione delle Unità Virtuali

		Unità Virtuali					
		UV_1	UV_2	…	UV_k	…	UV_S
Specie	sp_1	w_{11}	w_{12}	…	w_{1k}	…	w_{1S}
	sp_2	w_{21}	w_{22}	…	w_{2k}	…	w_{2S}
	⋮	⋮	⋮		⋮		⋮
	sp_i	w_{i1}	w_{i2}	…	w_{ik}	…	w_{iS}
	⋮	⋮	⋮		⋮		⋮
	sp_m	w_{m1}	w_{m2}	…	w_{mk}	…	w_{mS}

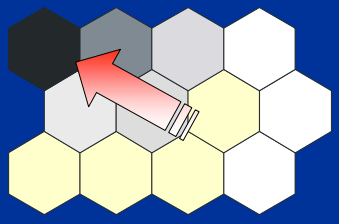


Ogni specie ha una diversa distribuzione…

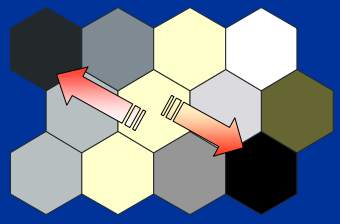
		Unità Virtuali					
		UV_1	UV_2	…	UV_k	…	UV_S
Specie	sp_1	w_{11}	w_{12}	…	w_{1k}	…	w_{1S}
	sp_2	w_{21}	w_{22}	…	w_{2k}	…	w_{2S}
	⋮	⋮	⋮		⋮		⋮
	sp_i	w_{i1}	w_{i2}	…	w_{ik}	…	w_{iS}
	⋮	⋮	⋮		⋮		⋮
	sp_m	w_{m1}	w_{m2}	…	w_{mk}	…	w_{mS}



Specie 1

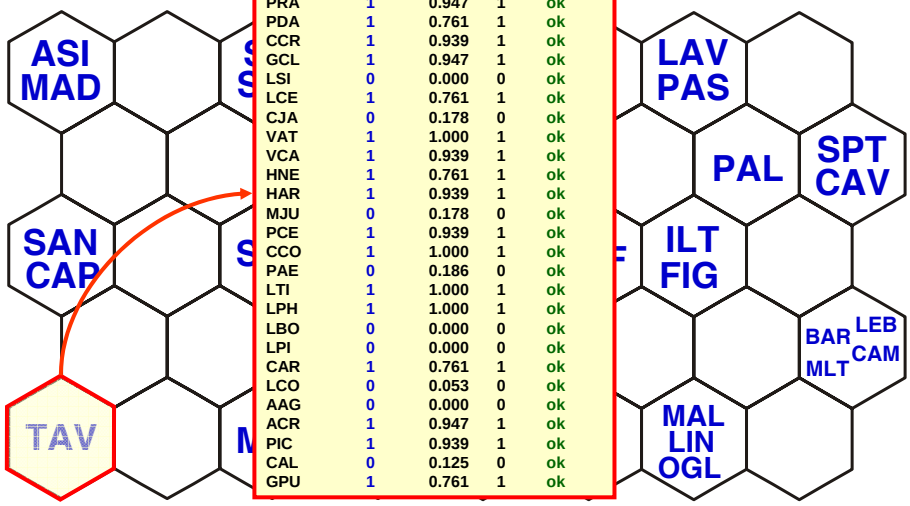


Specie 2



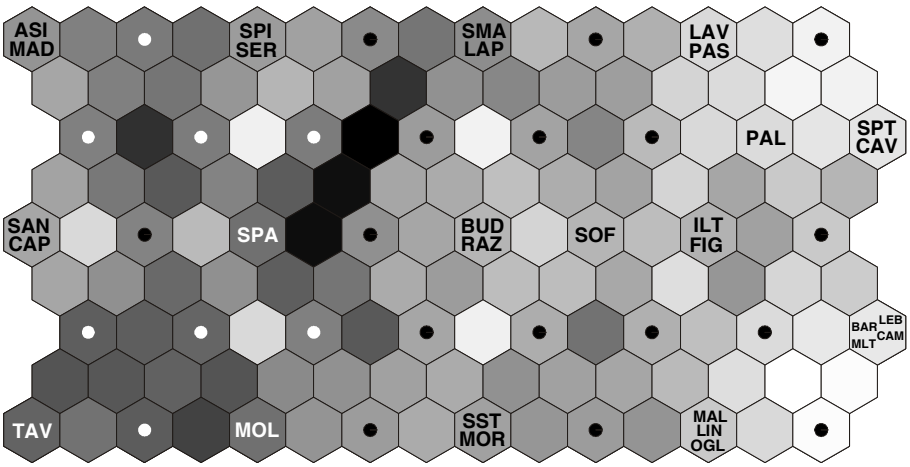
Specie *i*

Una SOM 8 x 5 bas… …ssenza delle farfalle diurne in un so… …ole circumsarde



Specie	TAV	UV		Accordo
PMA	0	0.000	0	ok
PBR	0	0.000	0	ok
PRA	1	0.947	1	ok
PDA	1	0.761	1	ok
CCR	1	0.939	1	ok
GCL	1	0.947	1	ok
LSI	0	0.000	0	ok
LCE	1	0.761	1	ok
CJA	0	0.178	0	ok
VAT	1	1.000	1	ok
VCA	1	0.939	1	ok
HNE	1	0.761	1	ok
HAR	1	0.939	1	ok
MJU	0	0.178	0	ok
PCE	1	0.939	1	ok
CCO	1	1.000	1	ok
PAE	0	0.186	0	ok
LTI	1	1.000	1	ok
LPH	1	1.000	1	ok
LBO	0	0.000	0	ok
LPI	0	0.000	0	ok
CAR	1	0.761	1	ok
LCO	0	0.053	0	ok
AAG	0	0.000	0	ok
ACR	1	0.947	1	ok
PIC	1	0.939	1	ok
CAL	0	0.125	0	ok
GPU	1	0.761	1	ok

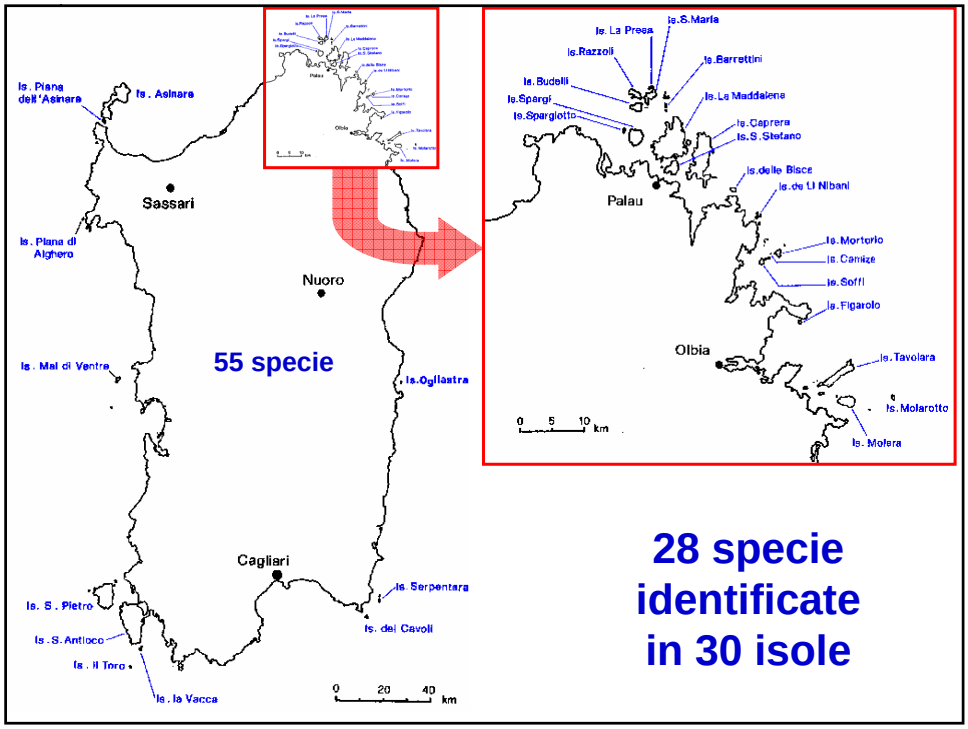
La matrice U consente di visualizzare le distanze fra gli elementi di una SOM.



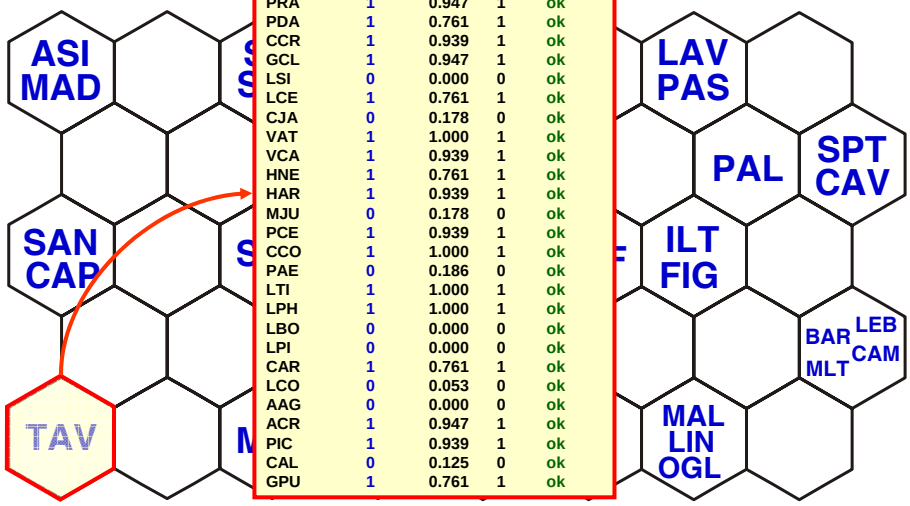
Un'applicazione delle **Self-Organizing Maps** (SOMs)

Dati estratti da:

Marina Cobolli, Marco Lucarelli e Valerio Sbordoni (1996).
Le farfalle diurne delle piccole isole circumsarde.
***Biogeographia*, 18: 569-582**

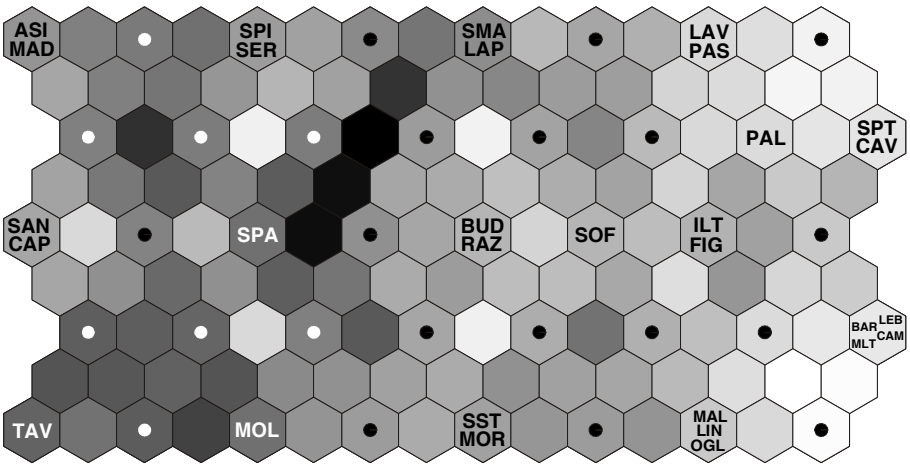


Una SOM 8 x 5 bas… …ssenza delle farfalle diurne in un so… …ole circumsarde

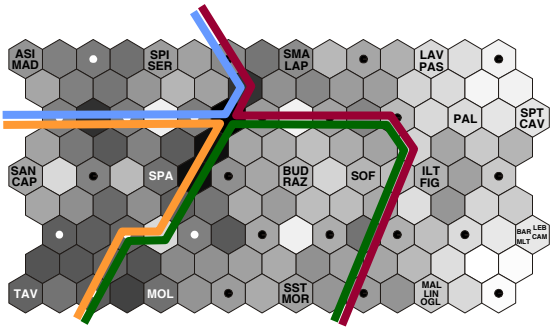


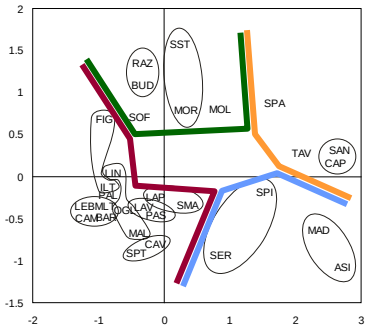
Specie	TAV	UV		Accordo
PMA	0	0.000	0	ok
PBR	0	0.000	0	ok
PRA	1	0.947	1	ok
PDA	1	0.761	1	ok
CCR	1	0.939	1	ok
GCL	1	0.947	1	ok
LSI	0	0.000	0	ok
LCE	1	0.761	1	ok
CJA	0	0.178	0	ok
VAT	1	1.000	1	ok
VCA	1	0.939	1	ok
HNE	1	0.761	1	ok
HAR	1	0.939	1	ok
MJU	0	0.178	0	ok
PCE	1	0.939	1	ok
CCO	1	1.000	1	ok
PAE	0	0.186	0	ok
LTI	1	1.000	1	ok
LPH	1	1.000	1	ok
LBO	0	0.000	0	ok
LPI	0	0.000	0	ok
CAR	1	0.761	1	ok
LCO	0	0.053	0	ok
AAG	0	0.000	0	ok
ACR	1	0.947	1	ok
PIC	1	0.939	1	ok
CAL	0	0.125	0	ok
GPU	1	0.761	1	ok

La matrice U consente di visualizzare le distanze fra gli elementi di una SOM.



Self-Organizing Map vs. Analisi delle Coordinate Principali





Gonepteryx cleopatra (L., 1767)

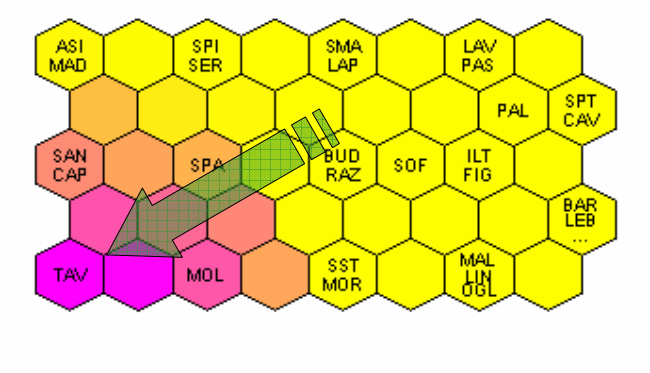
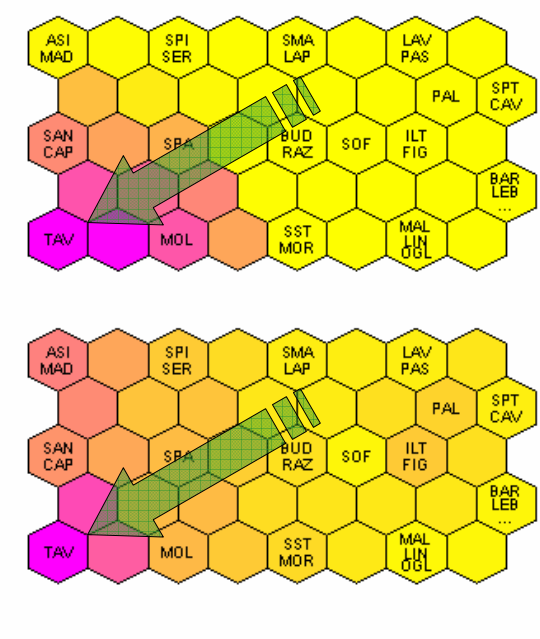


Foto: www.leps.it



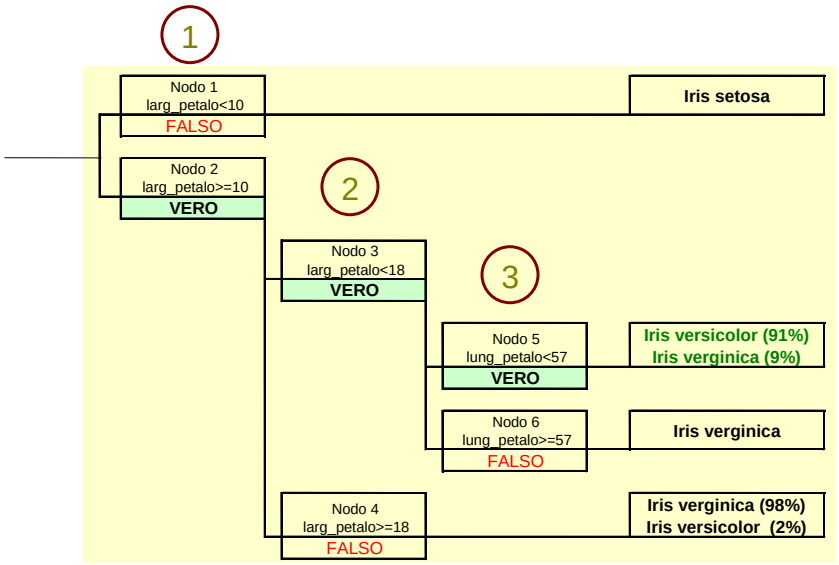
Gonepteryx cleopatra

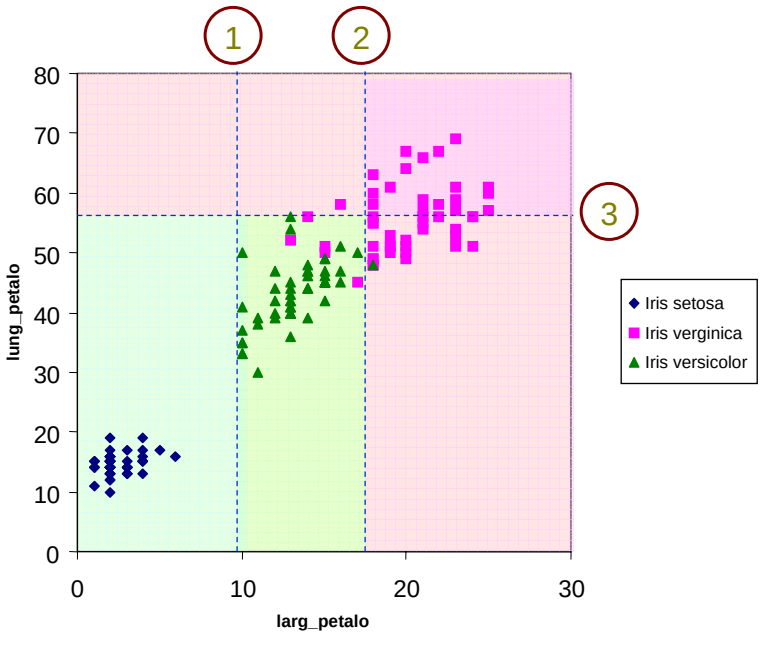
Altitudine

Foto: www.leps.it

Classification Trees

Il metodo: un esempio classico...





Modelli strutturali: classification tree

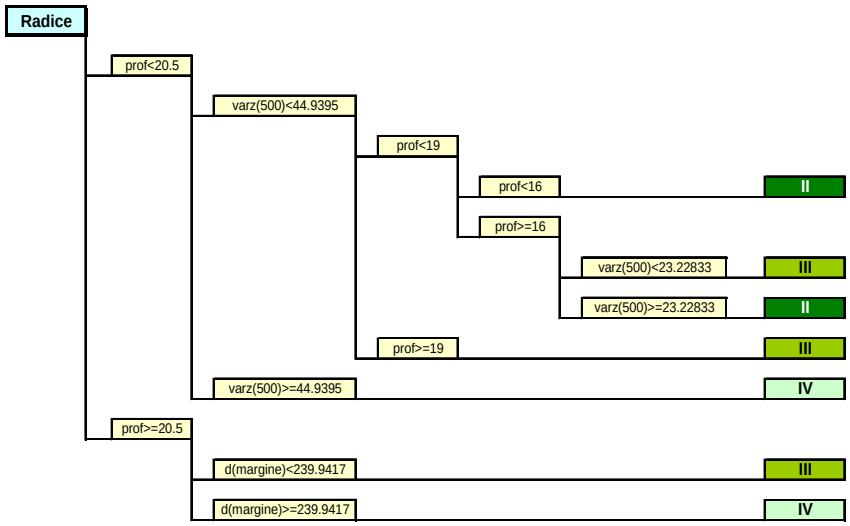
Variabili predittive:

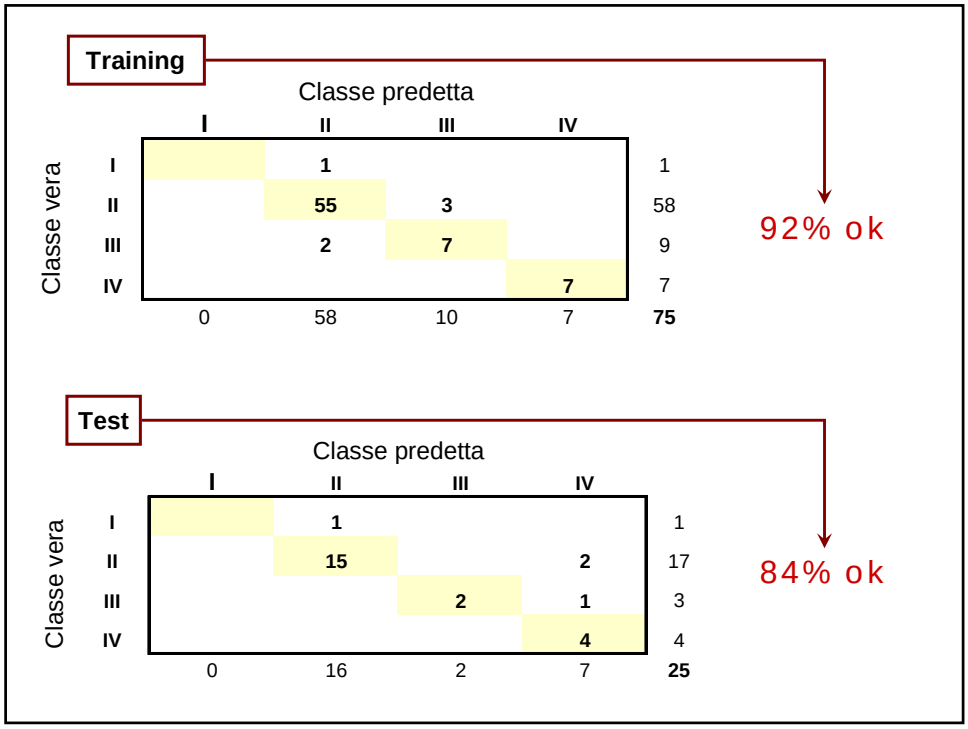
- Profondità
- Varianza della profondità (raggio=500 m)
- Distanza dal punto noto senza *Posidonia* più vicino (distanza dal margine della prateria stimata per eccesso)

Variabile da predire:

- Classe di densità della prateria (secondo Giraud)

Un modello per Rosignano:




Training
Classe predetta
I II III IV
Classe vera
I 1 1
II 55 3 58
III 2 7 9
IV 7 7
0 58 10 7 75
92% ok
Test
Classe predetta
I II III IV
Classe vera
I 1 1
II 15 2 17
III 2 1 3
IV 4 4
0 16 2 7 25
84% ok

In sintesi…

Alcuni metodi di classificazione

Tipo	Tecnica	Uso
Non gerarchico	*k*-means	Partizionare insiemi di dati
Gerarchico divisivo	*k*-means ripetuto	Buona tecnica su piccoli insiemi di dati
Gerarchico agglomerativo		
	Legame semplice	Tassonomia
	Legame completo	Classi omogenee
	Legame intermedio, centroide, di Ward	Scelta più frequente
Machine learning	SOM	Visualizzazione di insiemi di dati complessi
	Classification trees	Previsione e pattern recognition