3. Fenetika

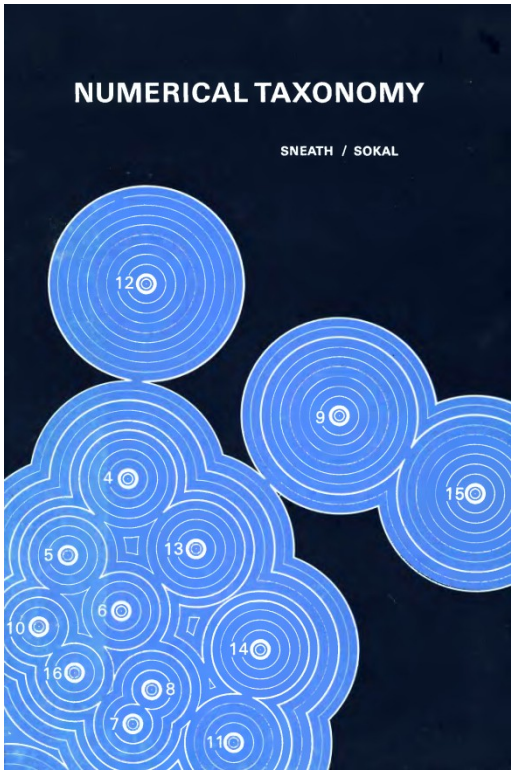
- numerická taxonomie
- použití fenetického přístupu v současné taxonomii
- taxonomický znak ze statistického hlediska
- tradiční a geometrická morfometrika
- shlukové analýzy
- ordinace (PCA)
- diskriminační analýza (CVA)
- ANN a automatické určování taxonů

Numerická taxonomie - Fenetika

- rozvoj výpočetní techniky
- Michener & Sokal (1957), Sneath (1957)
- Sokal R. & Sneath P. (1963): *Principles of Numerical Taxonomy*

Robert Sokal

• taxonomie jako praktická a empirická věda
• klasifikace založená na celkové podobnosti ve fenotypu
• čím více znaků, tím lépe
• každý znak má stejnou váhu
• jednotlivé taxony mohou být rozeznány díky korelaci různých znaků
• použití metod mnohorozměrné statistiky



Sneath & Sokal 1973

Postup fenetiků

- matice znaků x taxonů
- z ní výpočet matice podobnosti
- klasifikace taxonů do skupin

	holub	pštros	krokodýl	ještěrka	pes
povrch těla	peří	peří	šupiny	šupiny	srst
teplota těla	teplokrevný	teplokrevný	studenokrevný	studenokrevná	teplokrevný
počet nohou	2	2	4	4	4
typ lebky	diapsidní	diapsidní	diapsidní	diapsidní	synapsidní
péče o mláďata	ano	ano	ano	ne	ano

	holub	pštros	krokodýl	ještěrka	pes
holub	100	100	40	20	40
pštros	100	100	40	20	40
krokodýl	40	40	100	80	40
ještěrka	20	20	80	100	20
pes	40	40	40	20	100

Zrzavý et al. 2004

Postup fenetiků

1) výběr *operational taxonomic units* (OTU) – jedinci, populace, druhy, vyšší taxony
2) zaznamenání co největšího počtu znaků (ca. 30-100)
3) selekce znaků (korelace, závislost na prostředí apod.)
4) zakódování znaků, vytvoření matice znaků (*character matrix)*

druh / znak	A	B	C	D
1	0	0	1	1
2	1	0	0	1
3	1	1	1	1
4	0	1	1	0
5	1	0	0	1
6	0	0	0	1
7	1	0	0	0
8	1	1	1	0
9	0	1	1	0
10	0	0	0	1

Drozd 2004

Postup fenetiků

5)

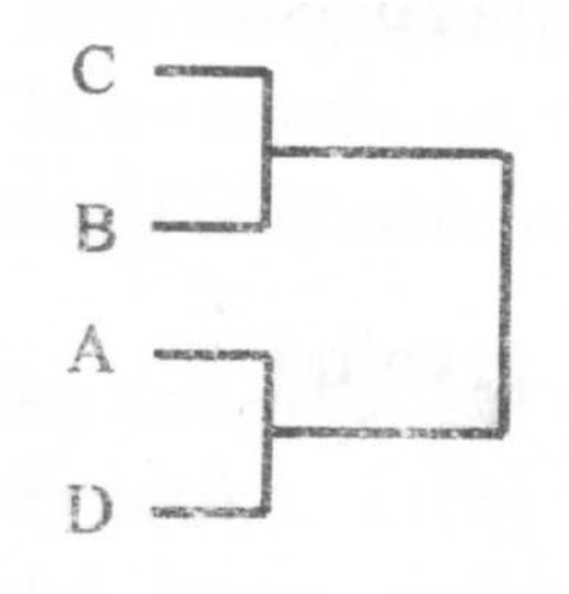
matice koeficientů vzdáleností (*distance matrix)*

Ja	B	C	D
A	0,2	0,2	0,3
B		0,3	0,1
C			0,2

6)

shluková analýza:

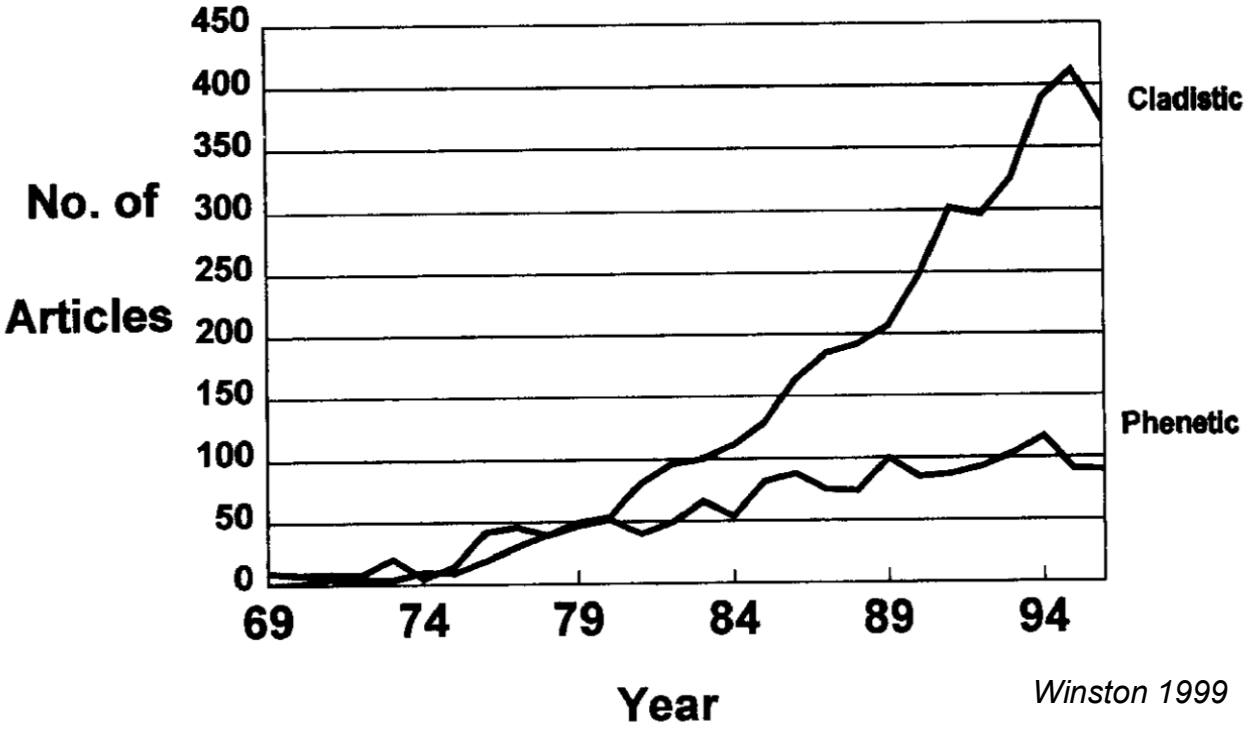
konstrukce fenogramu



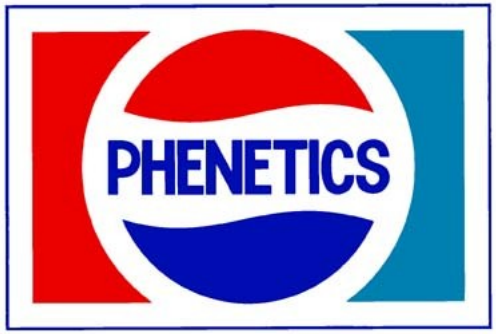
Úskalí fenetiky při tvorbě biologických klasifikací

- vychází z přístupu, že fylogeneze není poznatelná
- odlišné statistické metody = odlišné výsledky
- problém stejnocennosti znaků:
  - různý obsah informací vhodných pro klasifikaci, relativní dle hierarchické úrovně (nestejná rychlost evoluce)
  - nerozlišuje povrchní podobnost (např. konvergence) od podobnosti zděděné (homologie)
  - velké množství znaků = mnoho informačního balastu

Kladistika vs. fenetika: frekvence použití pro biologické klasifikace



Winston 1999



- fenetika je ve většině případů **nevhodná pro rekonstrukci fylogeneze**
- přínos: nutnost přesné definice metod, znaků, využití výpočetní techniky

Použití fenetického přístupu v současné taxonomii

- hodnocení vnitrodruhové a mezidruhové variability (vymezení taxonů, nalezení diagnostických znaků)
- „z nouze cnost“: pragmatická klasifikace jen na základě podobnosti bez nároku na fylogenetickou správnost
- hodnocení molekulárně-biologických dat, např. DNA-hybridizace, fingerprinting, imunologie, sekvence nukleotidů po korekci substitučními modely: tzv. distanční metody (*UPGMA*, *neighbor joining*)

Dělení znaků ze statistického hlediska

- kvalitativní (*qualitative*)
  - binární (*binary*): dva stavy: 0,1
  - vícestavové (*multistate*): 0,1, 2, 3, …
- semikvantitativní (*semiquantitative*)
- kvantitativní (*quantitative*):
  - nespojité, diskrétní (*discontinuous, discrete, meristic)*
  - spojité, kontinuální (*continuous*)

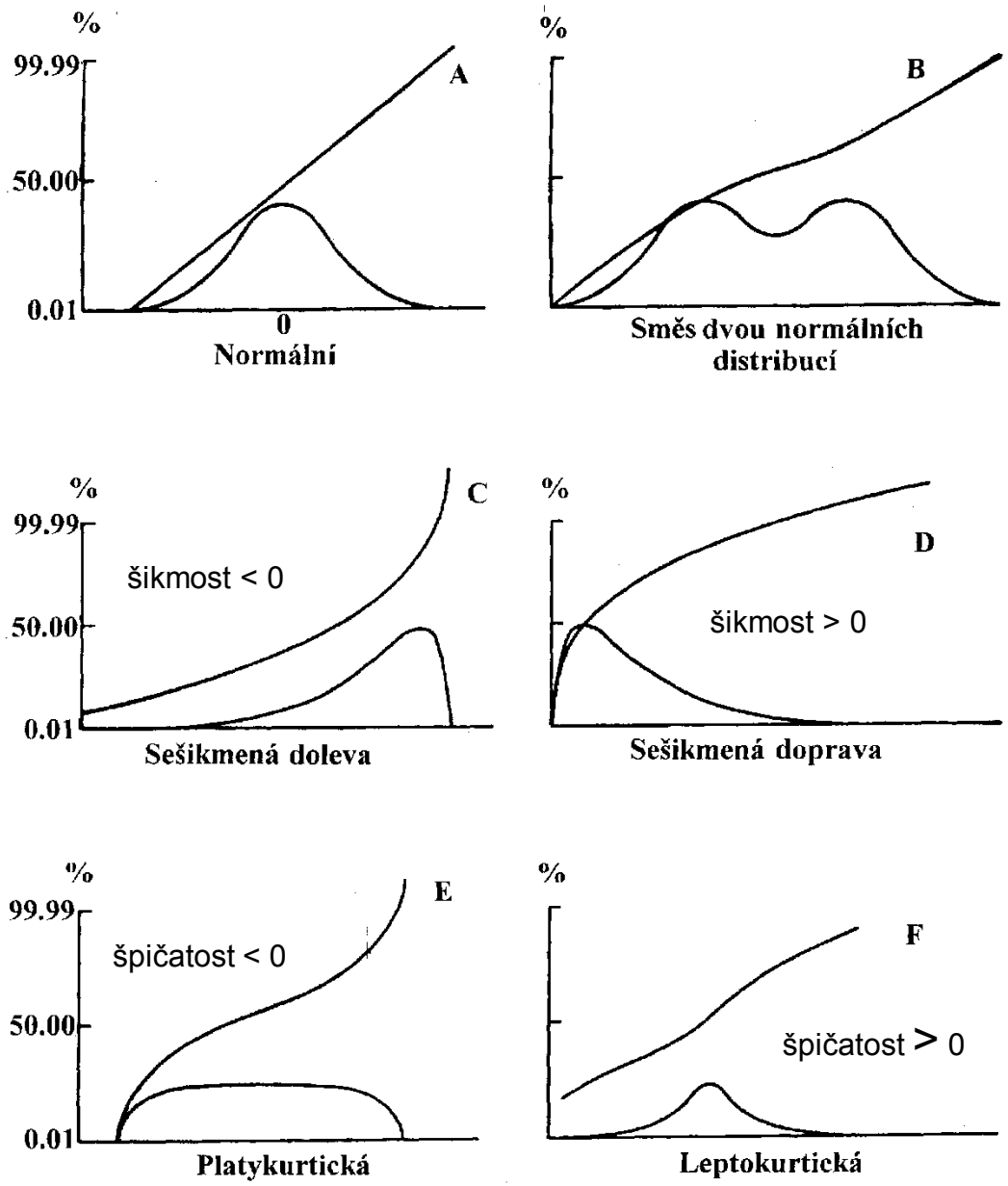
Převod vícestavového znaku na binární pomocí umělých proměnných (*dummy variables)*

stavy kvalitativního znaku	umělé binární proměnné			
a	1	0	0	0
b	0	1	0	0
c	0	0	1	0
d	0	0	0	1

Marhold & Suda 2002

Popisná statistika kvantitativního znaku

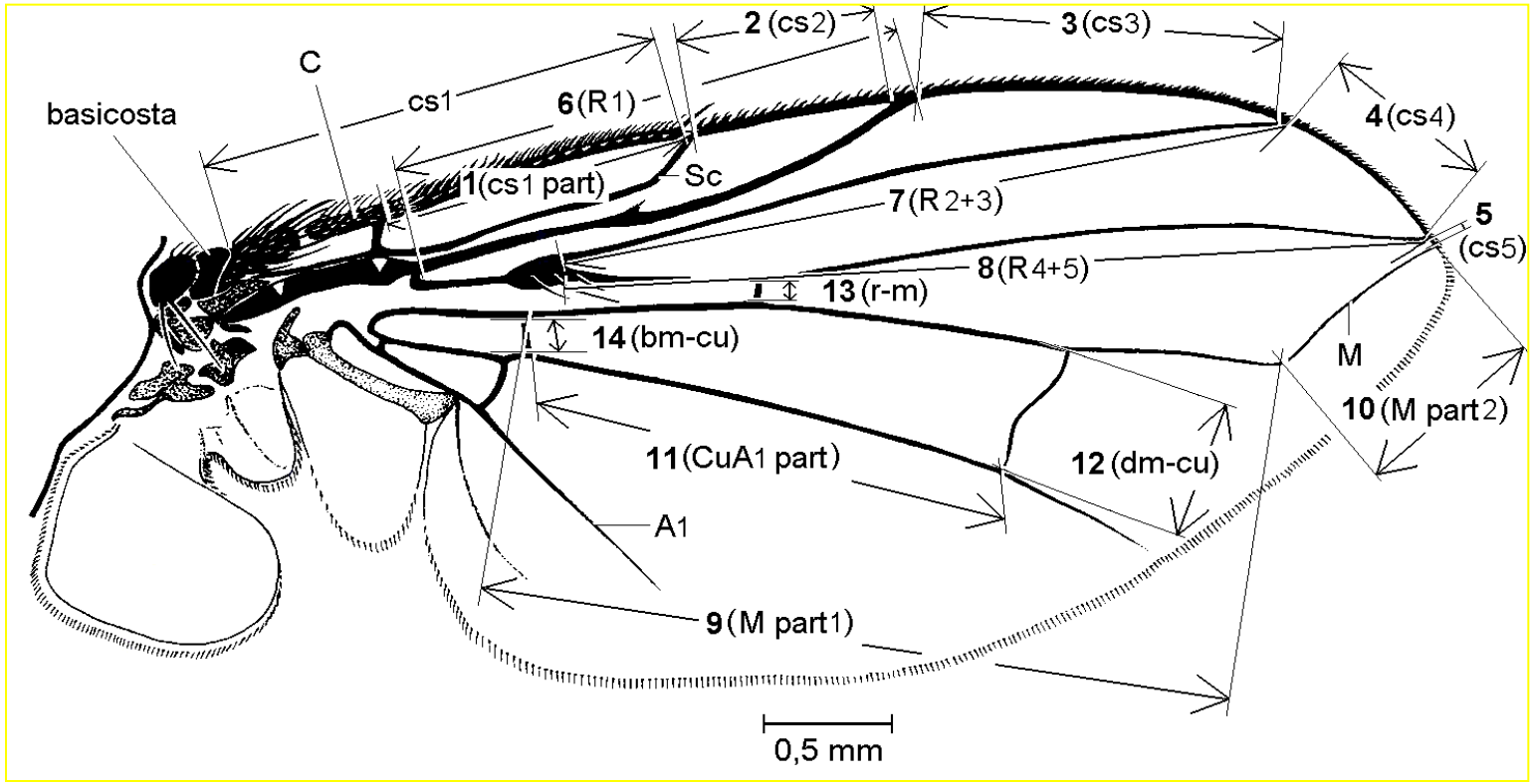
- ukazatel středu (průměr, medián, modus)
- ukazatel variability (rozpětí min-max, kvantily, rozptyl, směrodatná odchylka)
- rozložení (grafické srovnání, šikmost, špičatost, testy normality)
- korelace mezi znaky



Lepš 1996

Tradiční morfometrika

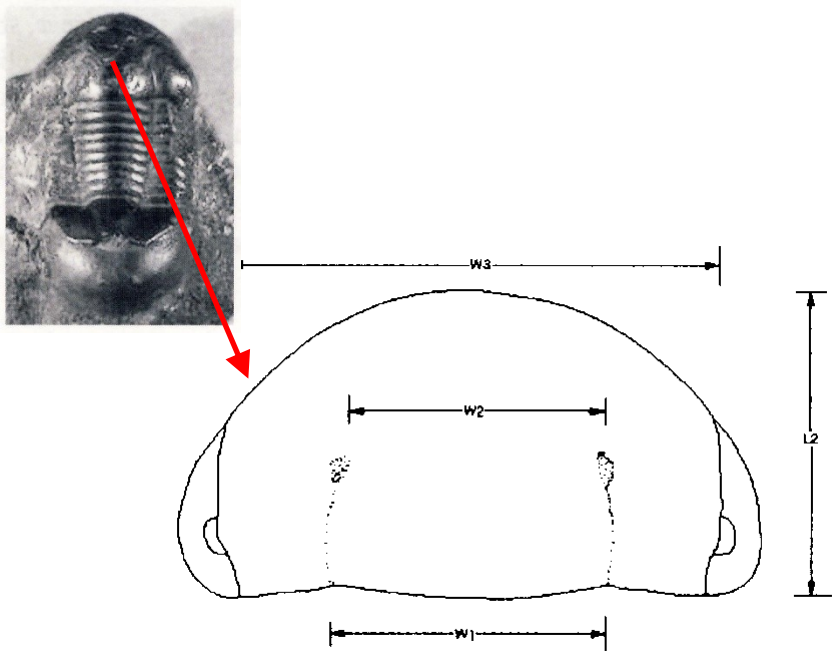
- měření délek, ploch, objemů, úhlů (lze nahradit cos α pro jednodušší hodnocení)

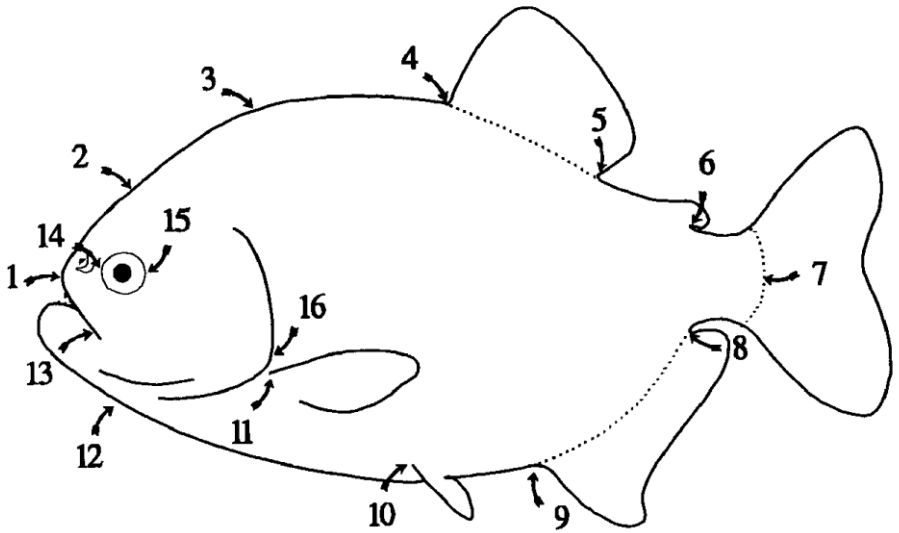


Vaňhara et al. 2007

Geometrická morfometrika: analýza tvaru

- tvar lze vyjádřit kvantitativními znaky
- nevychází ze vzdáleností, ale ze srovnávání:
  - obrysů (*outline analysis*)
  - polohy význačných bodů (*landmarks*)



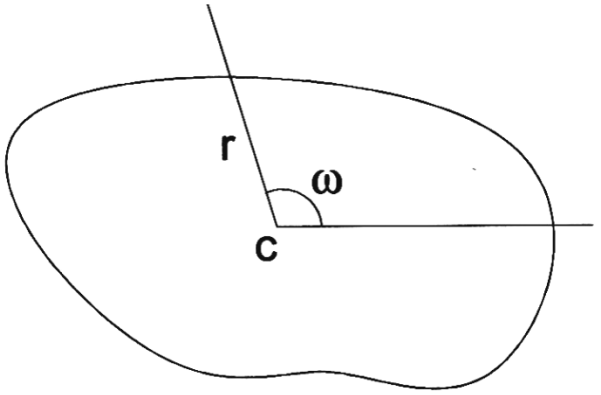


Hammer 2002

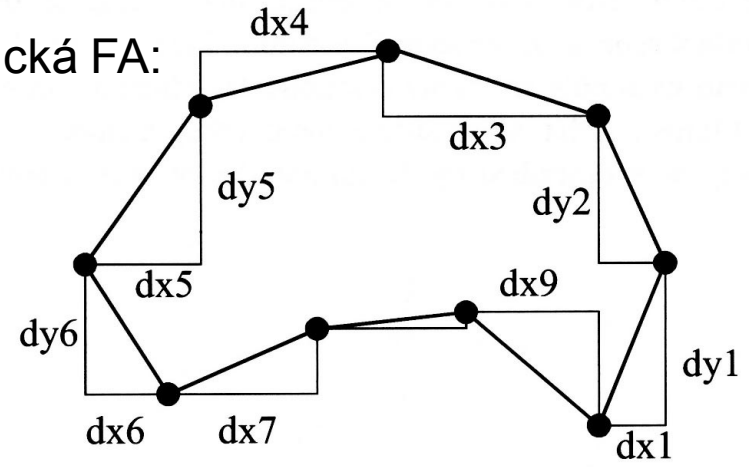
Zelditch et al. 2000

Analýza obrysů

- Fourierova analýza, eliptická Fourierova analýza
- uzavřený obrys jakožto periodická funkce
- každou periodickou funkci lze rozložit na sérii několika harmonických složek (sin, cos s příslušnými koeficienty - amplitudami), které jsou násobky původní funkce – matematické vyjádření tvaru



Eliptická FA:



$$r(\omega) = a_0 \cos\omega + b_0 \sin\omega + a_1 \cos\omega + b_1 \sin\omega + a_2 \cos\omega + b_2 \sin\omega + \ldots$$

$$r(\omega) = a_0 + \Sigma(a_i \cos\omega + b_i \sin\omega)$$

- řeší problém vícenásobného překryvu poloměru s obrysem při komplikovanějších tvarech
- odečtení x- a y-přírůstků od většího počtu pravidelně umístěných bodů na křivce
- 2 samostatné periodické funkce pro x a y
- dvojnásobný počet Fourierových koeficientů

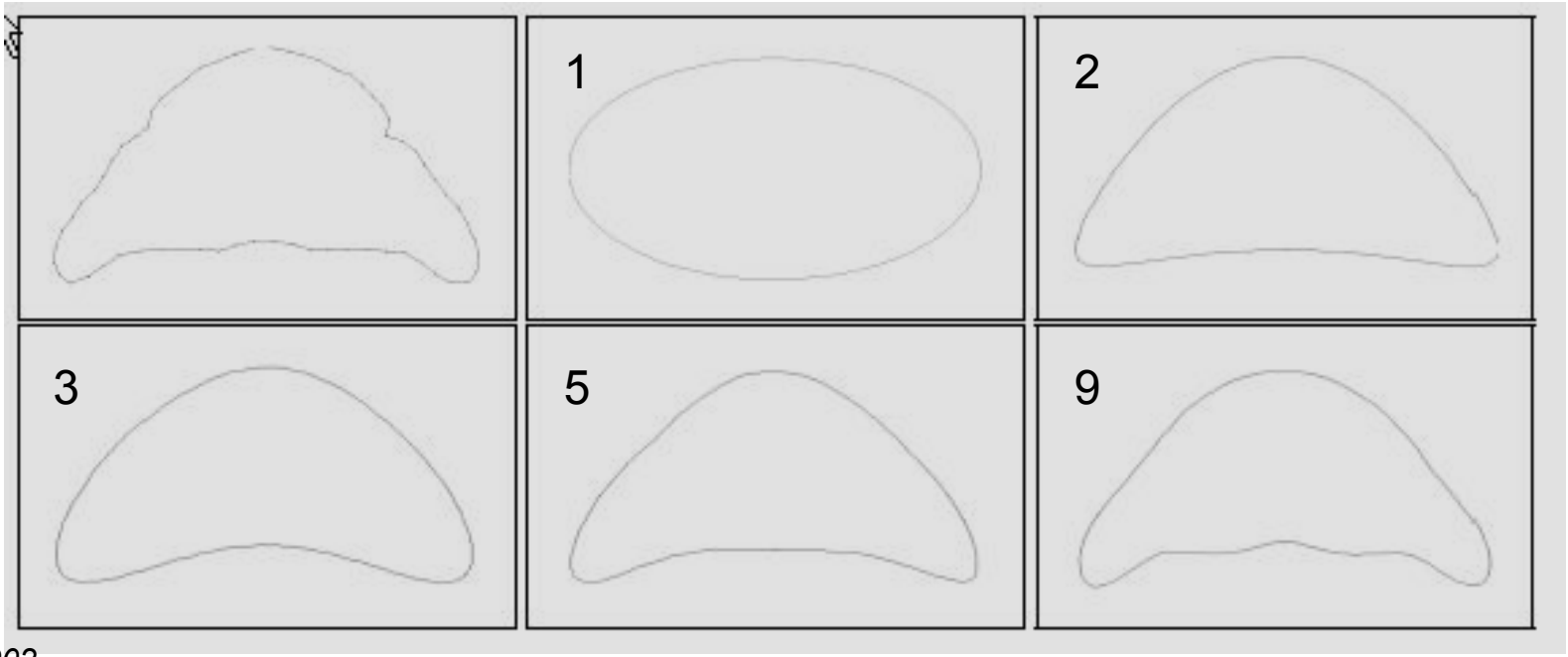
Hammer 2002

Analýza obrysů

- např. tvar hlavohrudi trilobita
- obrys digitalizován pomocí 64 bodů
- k adekvátnímu popisu tvaru pomocí EFA dostačuje 9 harmonických složek, tj. 36 koeficientů (=2*2*9)
- výhoda analýzy obrysů – tvar lze zpětně rekonstruovat

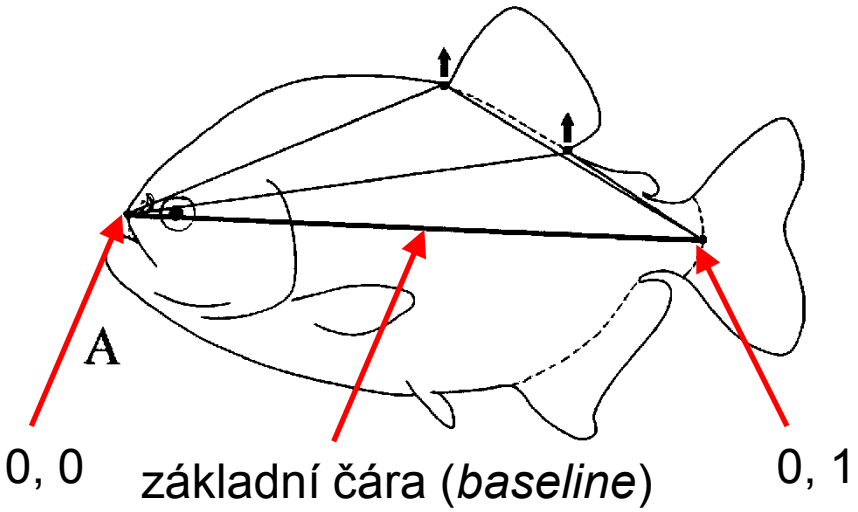
digitalizovaný obrys

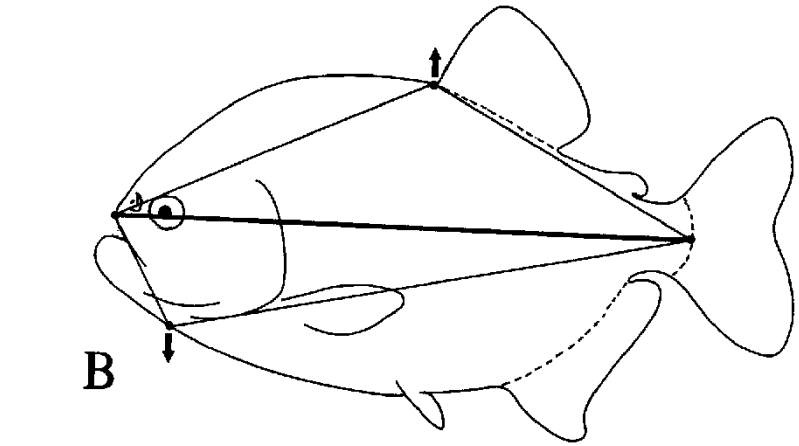
obrys rekonstruovaný pomocí různého počtu harmonických složek:



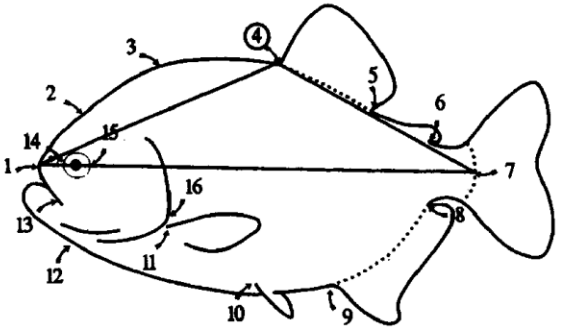
Hammer 2002

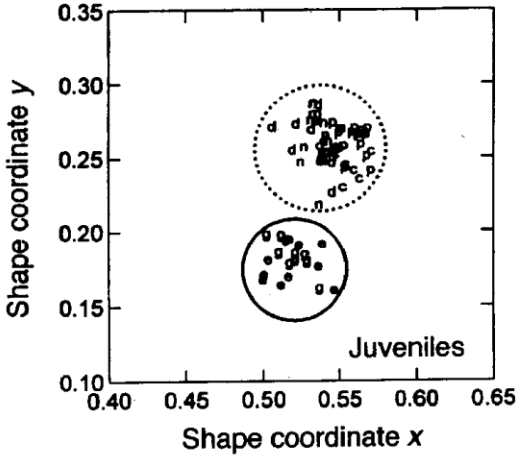
Význačné body: superpoziční metody (a)

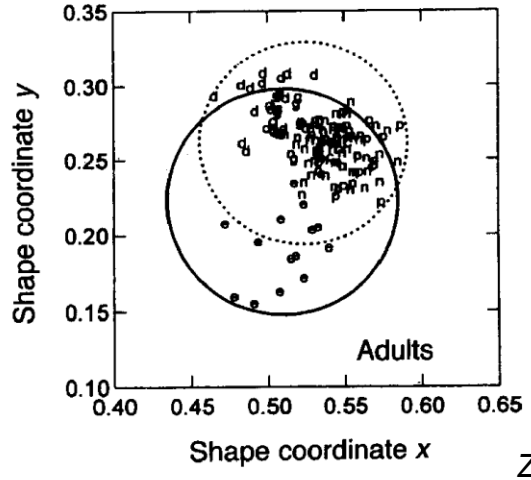




Booksteinovy souřadnice





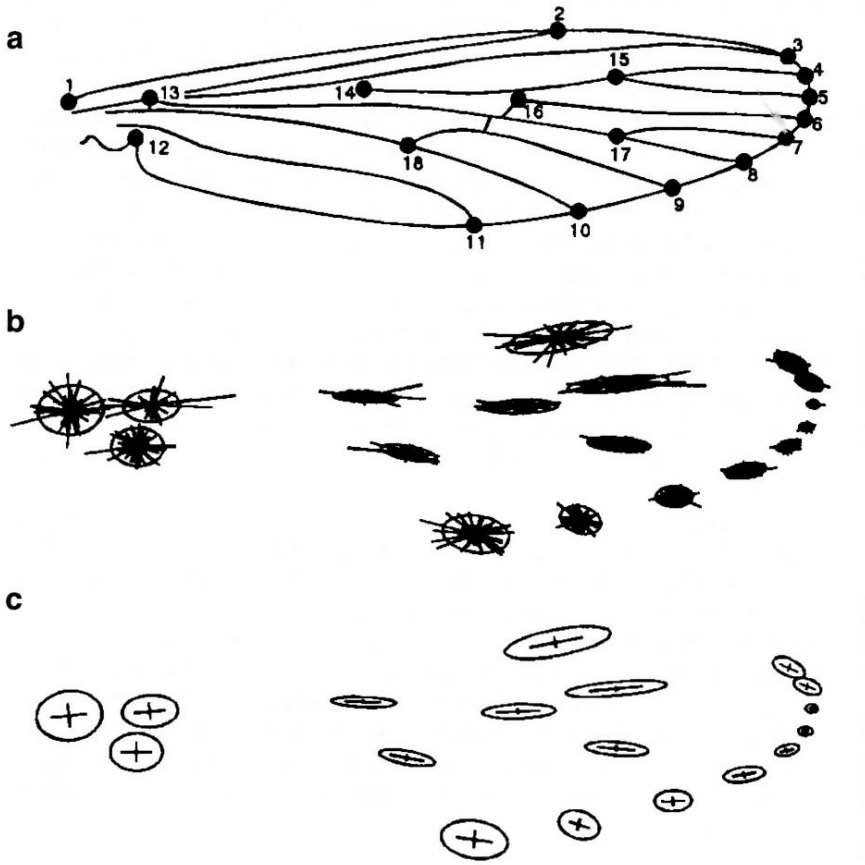


Zelditch et al. 2000

Význačné body: superpoziční metody (b)

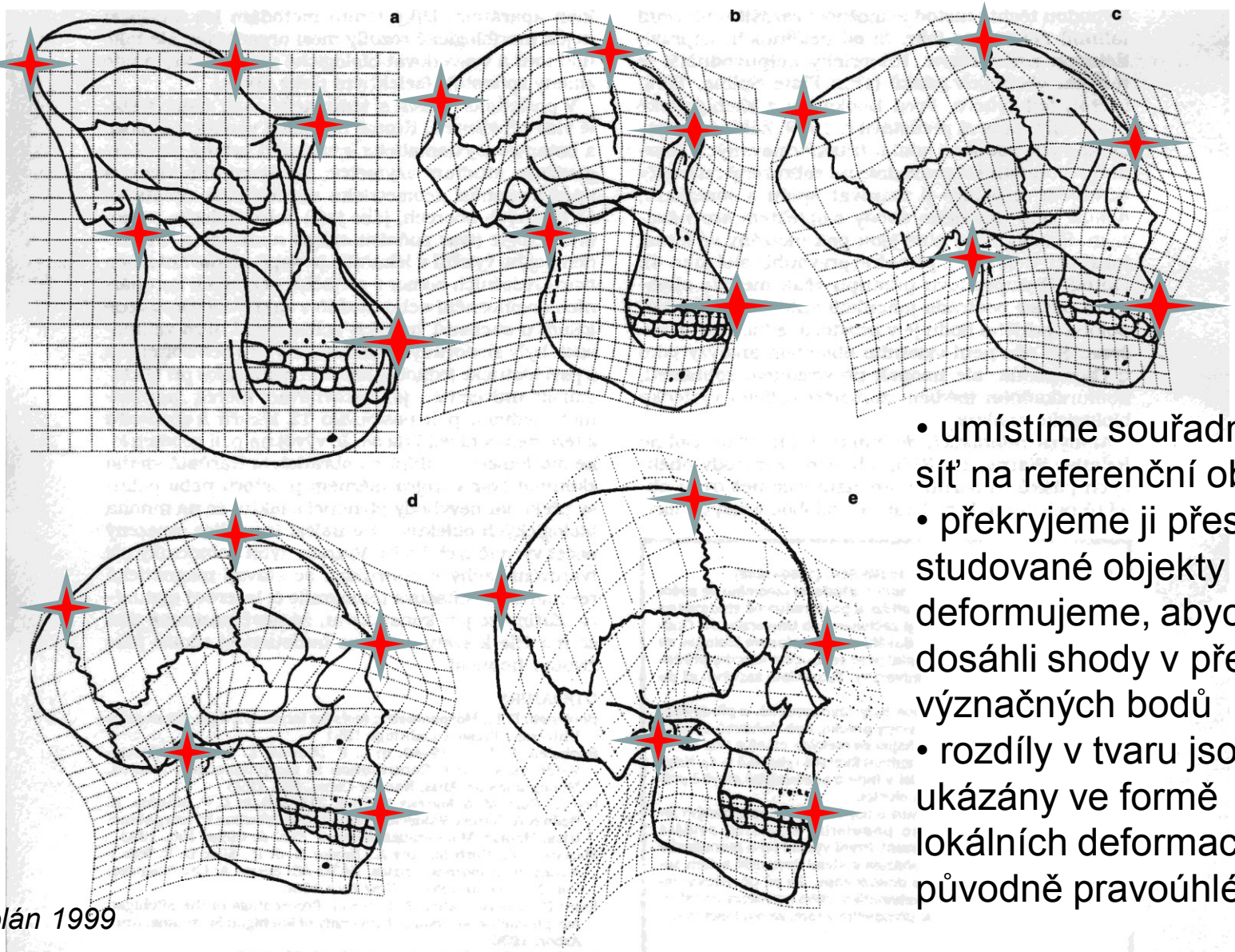
Prokrustovská superpozice (*Procrustes superimposition*)

- optimalizace míry shody v konfiguracích význačných bodů dvou a více objektů s využitím rotace, posunu a celkové (izometrické) změny velikosti tak, aby suma druhých mocnin rozdílů souřadnic mezi homologickými body byla minimální (podobné regresi, GLS)
- míra podobnosti mezi různými tvary: prokrustovská vzdálenost



Rohlf & Slice 1990

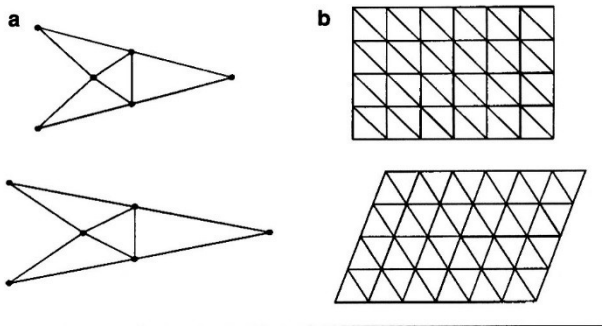
Deformační metody: Metoda ohebných pásků (*thin-plate spline)*



- umístíme souřadnicovou síť na referenční objekt
- překryjeme ji přes další studované objekty a deformujeme, abychom dosáhli shody v překrytí význačných bodů
- rozdíly v tvaru jsou ukázány ve formě lokálních deformací původně pravoúhlé sítě

Macholán 1999

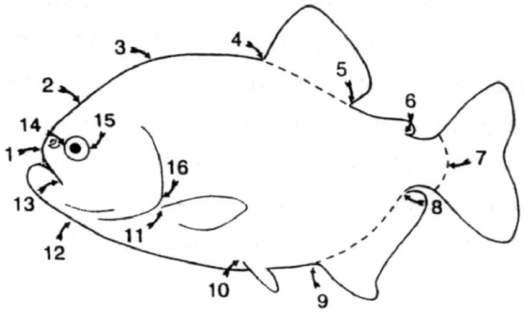
Metoda ohebných pásků (*thin-plate spline)*

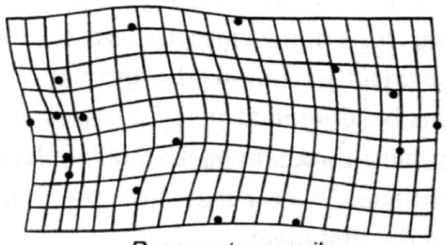


afinní složka

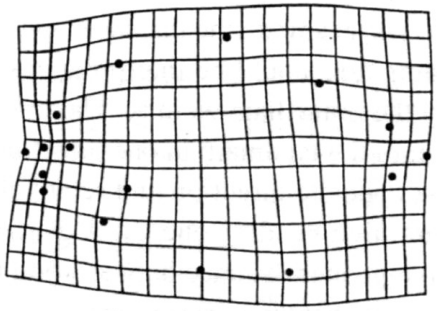
Macholán 1999

- umožňuje rozlišit uniformní (afinní) a nepravidelné (lokální) změny tvaru
- matice souřadnic a matice deformační energie
- vektory deformací podél každé osy vzhledem k referenční konfiguraci: parciální deformace (*partial warps*)
- analýza relativních deformací (*relative warps*) – obdoba PCA

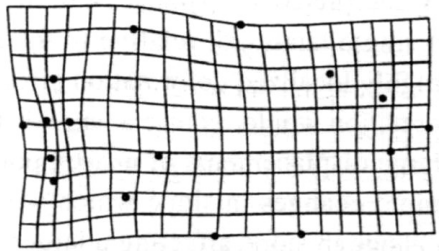




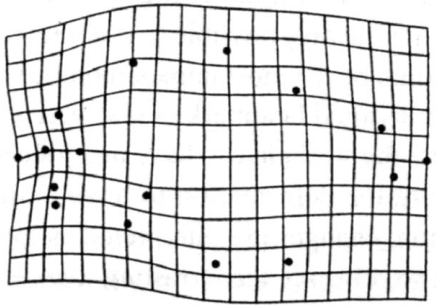
Pygocentrus cariba



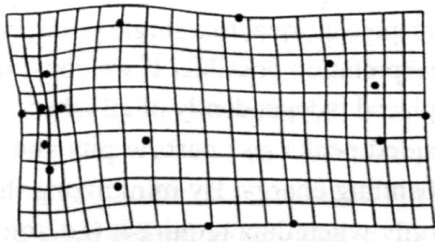
Pygopristis denticulata



Pygocentrus nattereri



Serrasalmus gouldingi



Pygocentrus piraya

Zelditch et al. 2000

Sběr morfologických dat pro statistickou analýzu

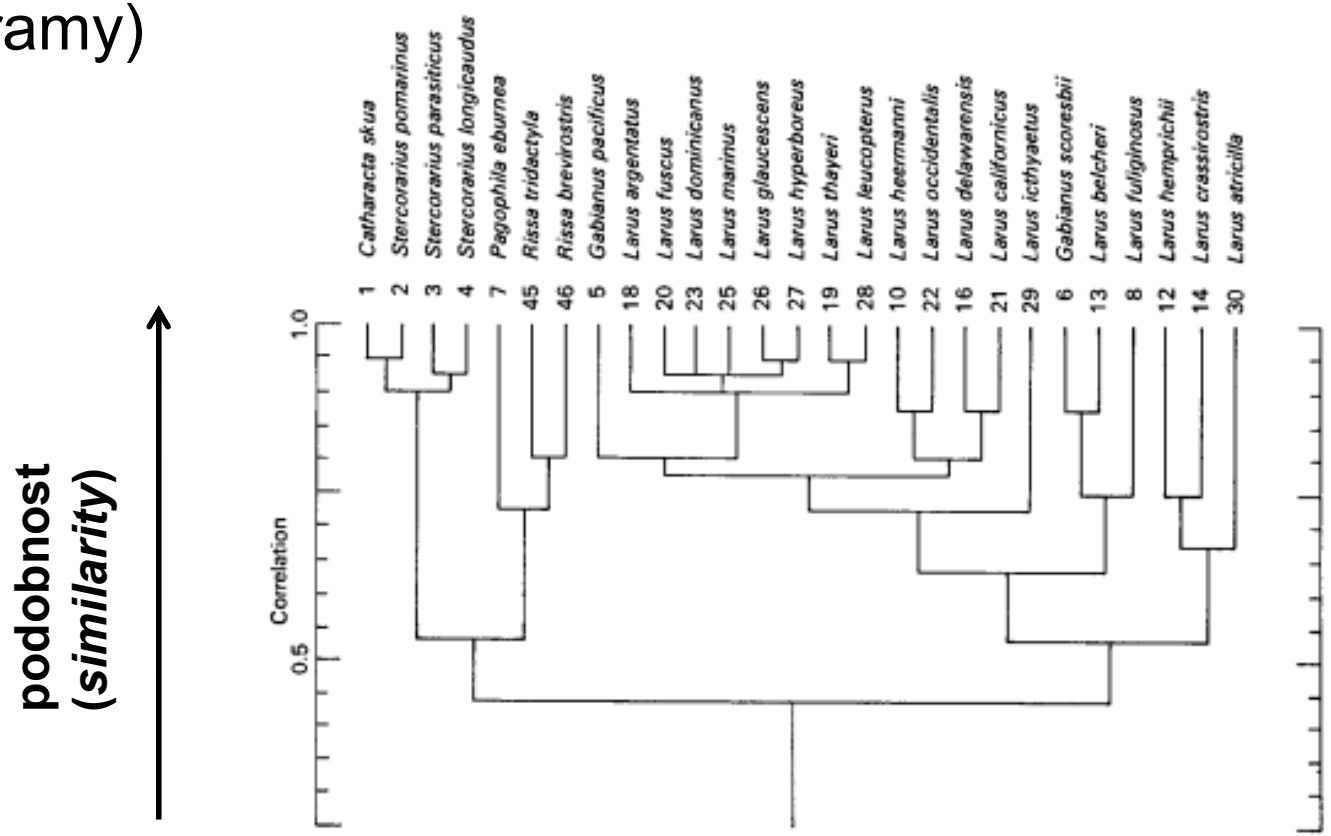
- jen kvantitativní a binární znaky
- vyloučení znaků závislých pouze na prostředí
- poměry mohou být někdy užitečné, ale mohou být problematické při statistickém vyhodnocení
- korelace mezi znaky – vyloučení silně korelovaných
- kolik znaků sledovat? – kompletnost vs. časová náročnost
- kolik jedinců prohlédnout? – podchycení variability
- počet jedinců vs. počet populací
- přesnost měření – pomůcka: počet jednotek mezi min a max by měl být mezi 30 a 300 (např. 5–10 mm, měřit s přesností na desetiny mm)
- chybějící data – vyřazení nebo nahrazení (např. průměrem)

Úprava matice dat

- matice znaků x OTU, *n*-rozměrů (*n*=počet znaků)
- <u>standardizace</u> (*standardization*) – převedení na stejné měřítko
  - centrováním: změní polohu nulového bodu $x'_{ij} = x_{ij} - \bar{x}_i$
  - rozpětím: když jsou znaky ve stejném měřítku, ale mezi jejich hodnotami jsou velké rozdíly $x'_{ij} = \frac{x_{ij} - \min_j \{x_{ij}\}}{\max_j \{x_{ij}\} - \min_j \{x_{ij}\}}$
  - směrodatnou odchylkou: kdvž isou znaky měřeny v odlišných škálách a jednotkách $x'_{ij} = \frac{x_{ij} - \bar{x}_i}{s_i}$
- <u>transformace</u> (*transformation*) - náprava odchylek od normality, odstranění heterogenity rozptylů
  - logaritmická, y=log(x+1)
  - odmocninová, $y=(x+1)^{-2}$
  - arkussinová (např. pro poměry a %)

Shlukové analýzy *(cluster analysis)*

• slouží k detekování přirozených skupin (shluků) v datech a často též k jejich uspořádání do hierarchických tříd (klasifikaci)
• výsledkem jsou obvykle stromové diagramy (dendrogramy)



Shlukové analýzy *(cluster analysis)*

Postup:

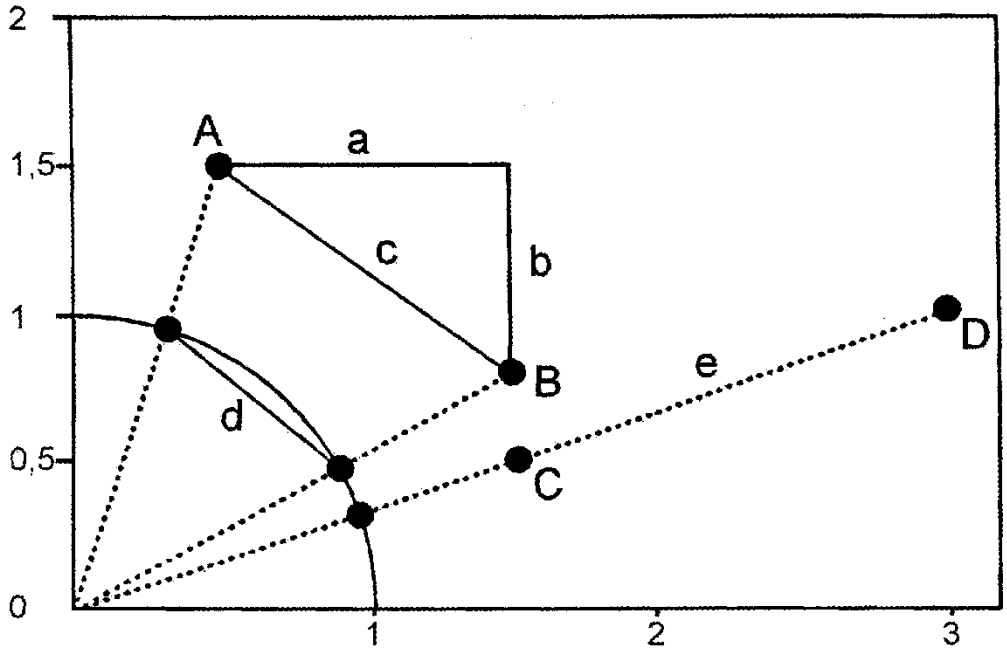
- 1. krok: výběr koeficientu podobnosti/vzdálenosti (metriky)

- 2. krok: výběr shlukovacího algoritmu

1a. Koeficienty podobnosti pro kvantitativní znaky:

- Eukleidovská vzdálenost (c)
- tětivová vzdálenost (*chord distance,* d)
- Manhattanská vzdálenost (a+b)
- Mahalanobisova vzdálenost (odstraňuje vliv korelace a závislosti na měřítku)

$$EU_{ij} = \sqrt{\sum_{k=1}^{p} \left(x_{ik} - x_{jk} \right)^2}$$



Marhold & Suda 2002

1b. Koeficienty podobnosti pro binární znaky a smíšená data

- Jaccardův index
  Ja=a/(a+b+c)
- jednoduchá shoda
  (*simple matching*)
  SM=(a+d)/(a+b+c+d)
- Sörensenův index
  Sor=2a/2a+b+c
- Gowerův index
  (smíšená data)

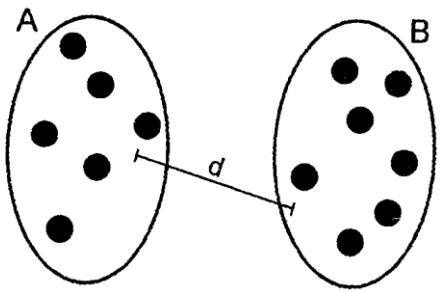
	objekt *i*		
objekt *j*	**kód znaku**	1	0
	1	**a**	**b**
	0	**c**	**d**

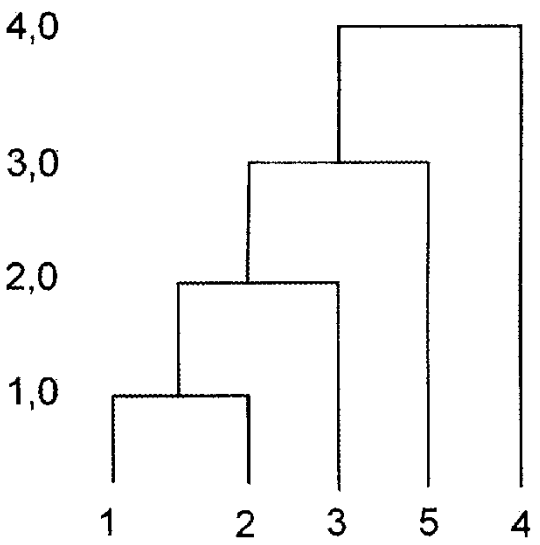
Marhold & Suda 2002

2. Shlukovací algoritmy

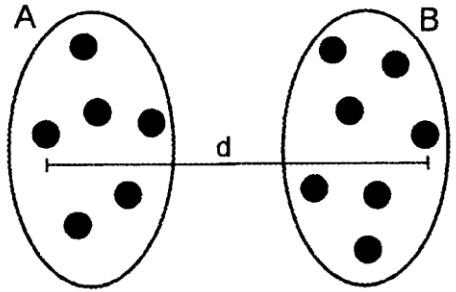
• rozdíly spočívají v tom, jak je definována vzdálenost mezi dvěma skupinami objektů

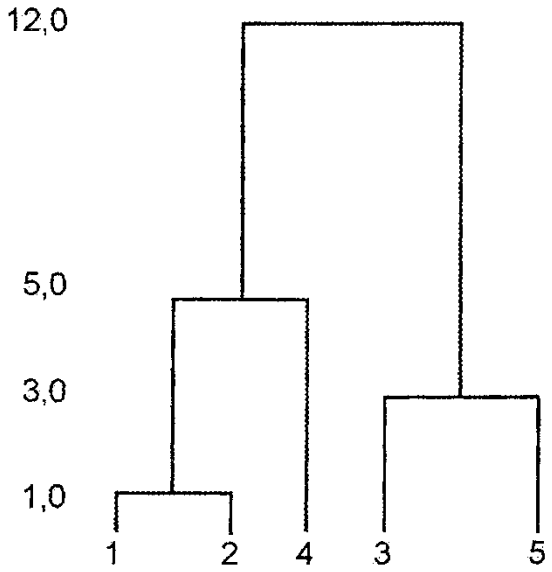
Metoda jednospojná
(*single linkage*)





Metoda všespojná
(*complete linkage*)

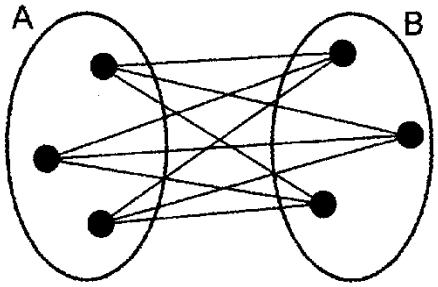


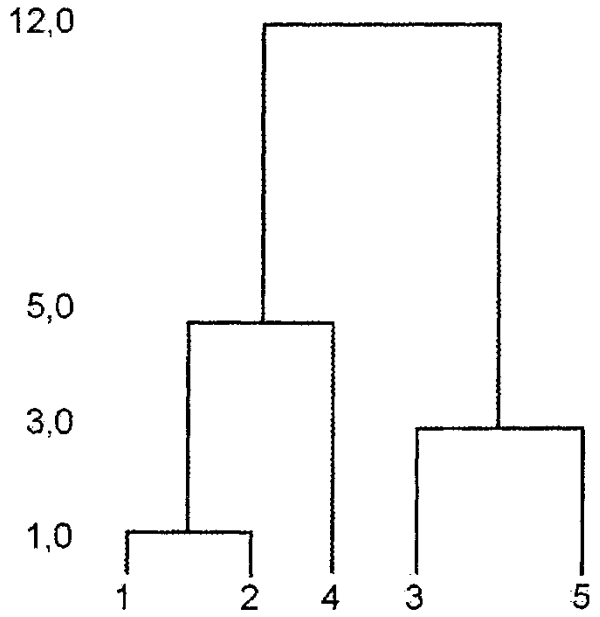


Marhold & Suda 2002

2. Shlukovací algoritmy

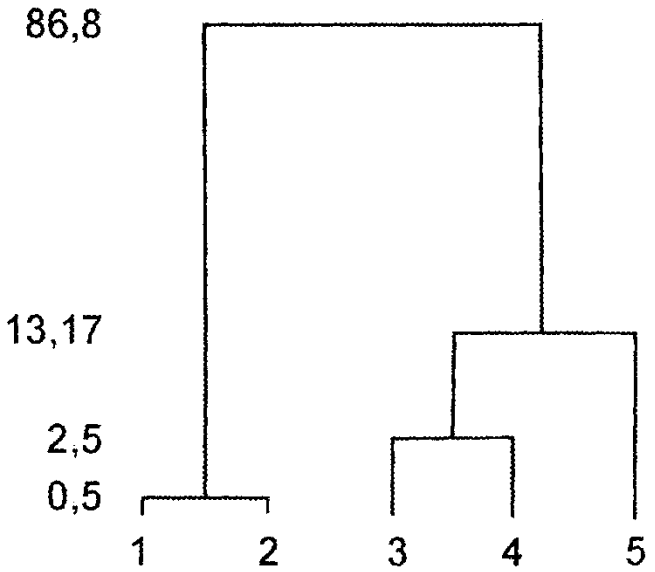
Metoda středospojná
(average linkage, UPGMA)





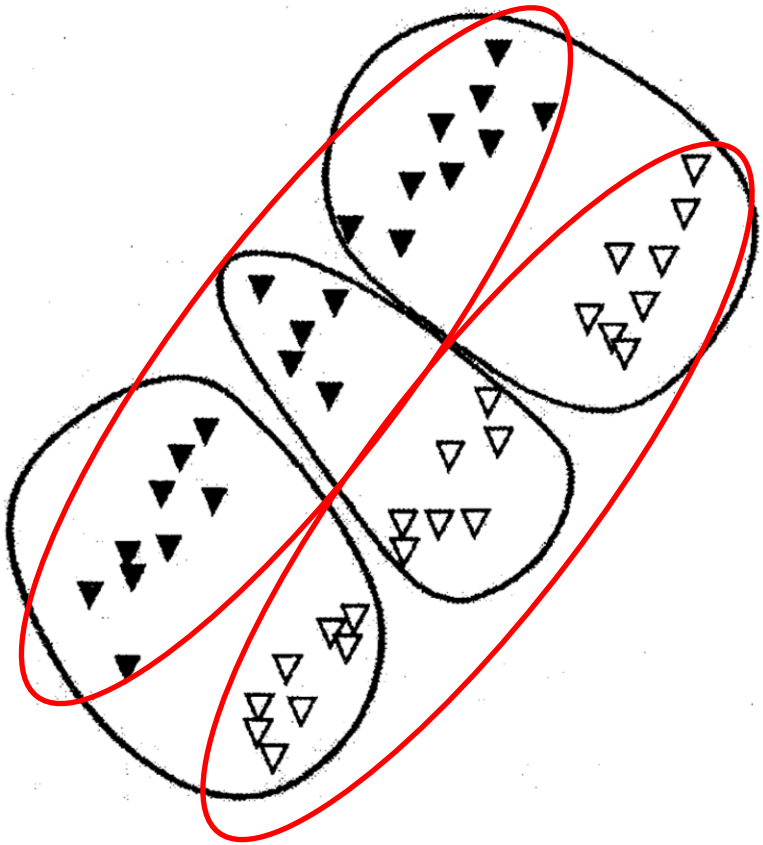
Wardova metoda

(minimalizace vnitroshlukového rozptylu)



Marhold & Suda 2002

Shlukové analýzy *(cluster analysis)* - shrnutí

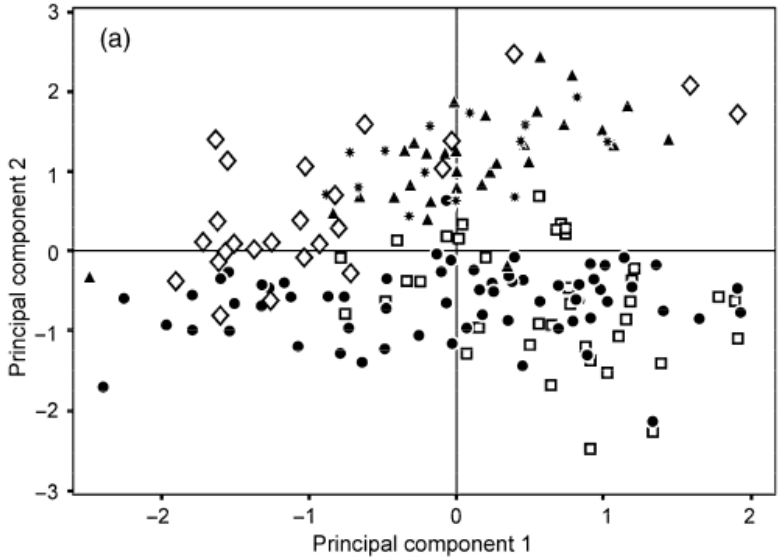


- nelze univerzálně doporučit optimální koeficient a metodu
- úspěšnost výsledku záleží na struktuře v datech
- zkusit více metod
- citlivost na odlehlé objekty
- nevhodné např.pro studium klinální variability

Marhold & Suda 2002

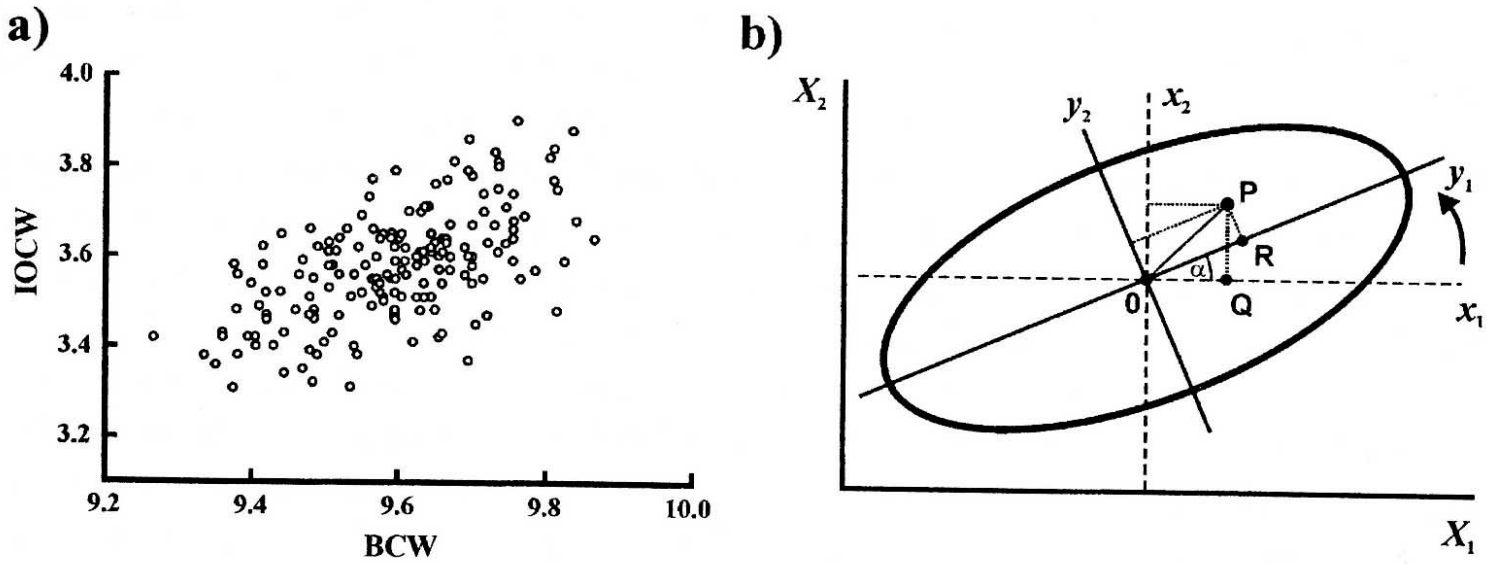
Ordinační metody

- cílem je nahradit velký počet znaků menším počtem hypotetických proměnných při minimální ztrátě informace (ideálně 2-3 osy)
- grafickým výstupem je ordinační diagram
- nepředpokládají a priorní seskupení objektů – explorační techniky k tvorbě hypotéz, k odhalení struktury v datech
- analýza hlavních komponent (PCA), analýza hlavních koordinát (PCoA), nemetrické mnohorozměrné škálování (NMDS), korespondenční analýza (CA)



Analýza hlavních komponent (PCA)

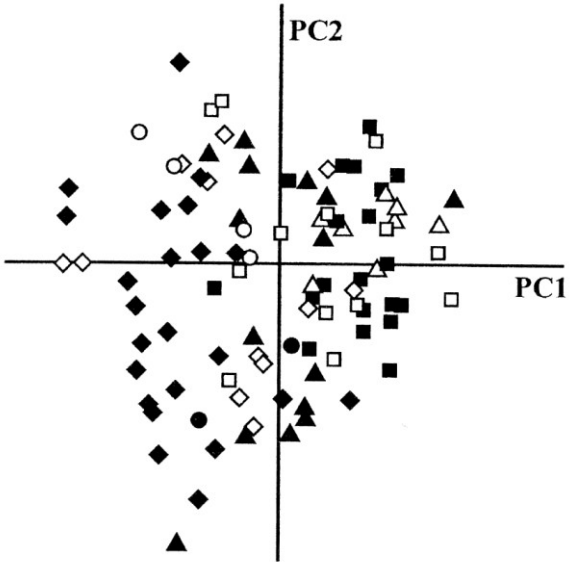
- osy (PC) vedeny ve směru největší variability vždy kolmo na sebe
- prvních několik PC na sebe váže nejvíce variability
- každá PC je lineární kombinací původních znaků
- hlavně pro kvantitativní znaky
- robustní k rozložení
- počet objektů by měl být větší než počet znaků
- kovariance vs. korelace

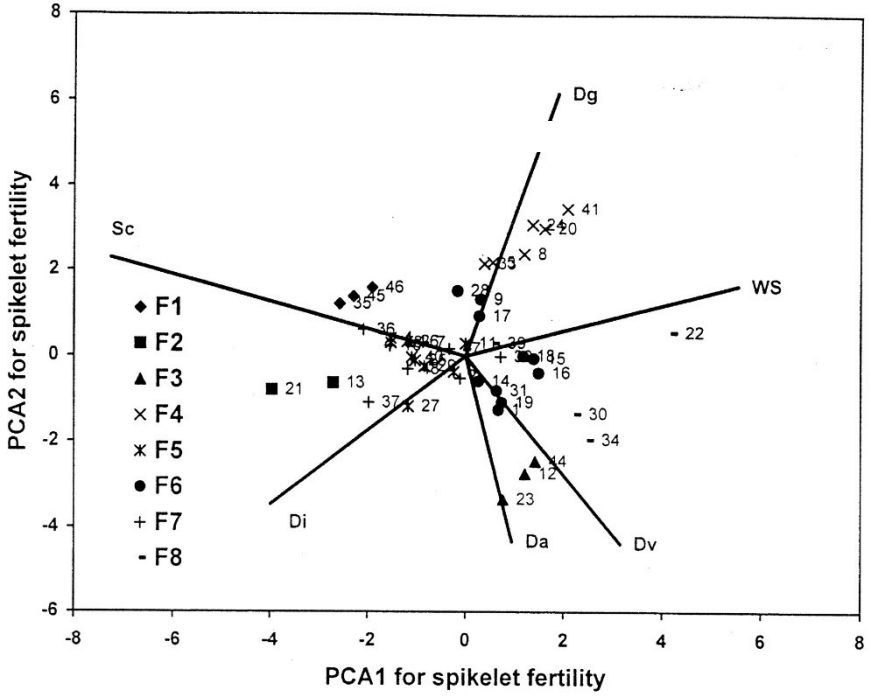


Zima & Macholán 2004

Interpretace výsledků PCA

- ordinace objektů a znaků, *biplot* (grafické znázornění) – podobné objekty leží blízko sebe, vektory korelovaných faktorů míří podobným směrem
- korelace znaků s jednotlivými PC: zátěže (*factor loadings*)
- vlastní čísla, latentní kořeny (*eigenvalues*) – míra variability v datech vyjádřená jednotlivými PC (absolutní hodnota, % podíl ze součtu EV)





Diskriminační analýza (DA)

- studujeme rozdíly mezi dvěma či více předem stanovenými skupinami (druhy, populacemi, pohlavími…)
- metoda testování hypotéz

	PCA, PCoA, NMDS	DA
Předem stanovené skupiny	Ne	Ano
Vysvětlení maximální variability	Celkové	meziskupinové
Vážení znaků	Ne	ano

Marhold & Suda 2002

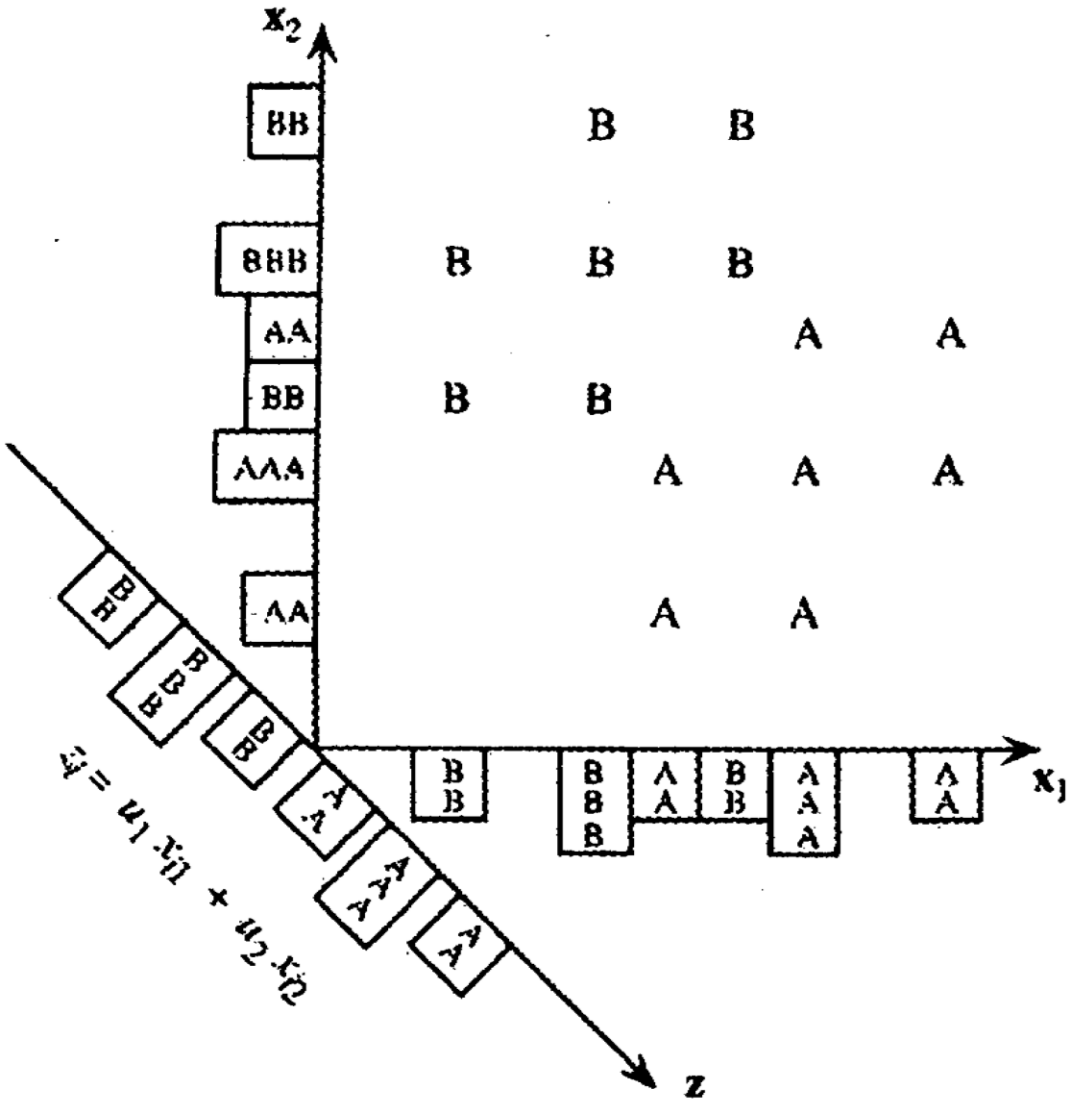
Kanonická diskriminační analýza, CDA (*canonical variates analysis, CVA*)

a) je možné odlišit předem stanovené skupiny objektů (druhy, populace,…) na základě znaků, které máme k dispozici, a do jaké míry?

b) které znaky jsou pro rozlišení skupin nejlepší?

- neumožňuje odhalit další možné přítomné skupiny (druhy, poddruhy apod.) v datech

- osy jsou vedeny ve směru největší variability mezi skupinami
- nová osa = kanonická diskriminační funkce je lineární kombinací původních znaků



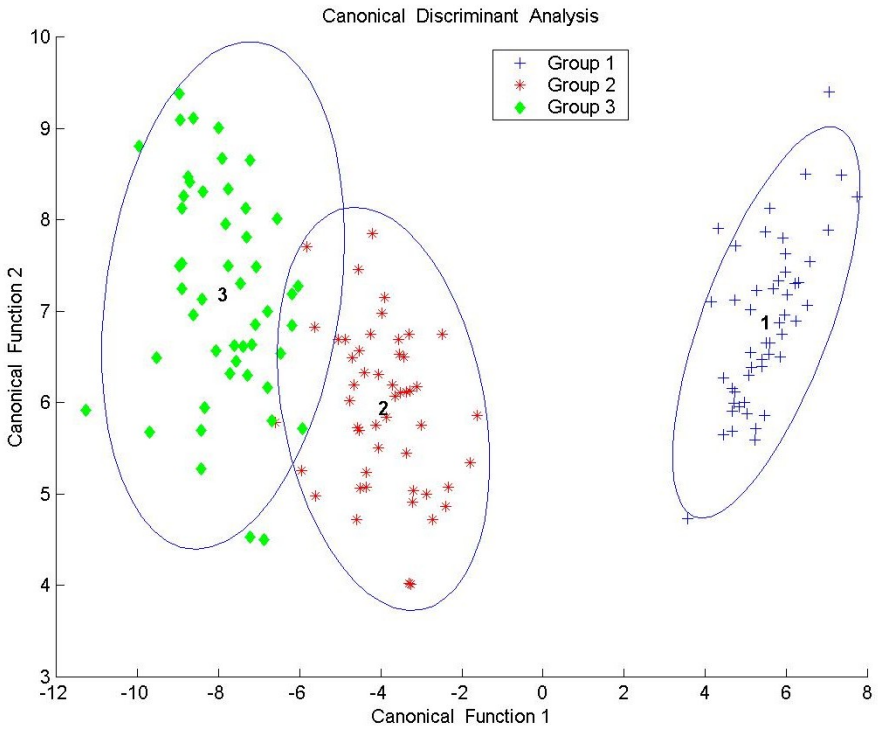
Marhold & Suda 2002

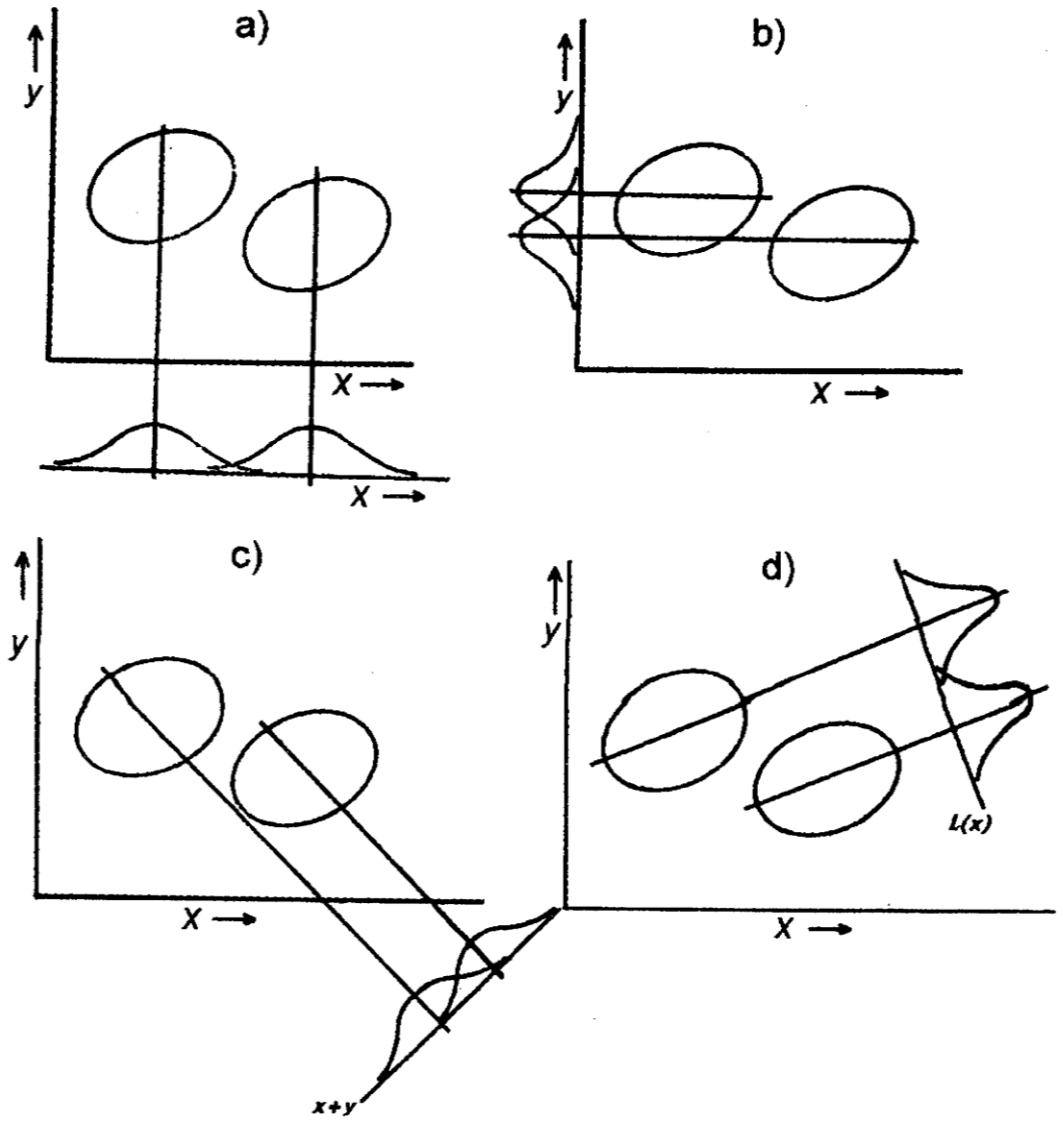
Požadavky CDA

- kvantitativní a binární znaky
- vyloučit znaky, které jsou navzájem lineární kombinací, silně korelované, a třídní znak
- mnohorozměrné normální rozložení
- alespoň 2 skupiny, v každé min. 2 objekty
- žádný znak by neměl být v nějaké skupině konstantní

Interpretace výsledků CDA

- relativní pozice objektů a skupinových centroidů (např. konfidenční intervaly)
- celková kanonická struktura – vztah mezi jednotlivými znaky a kanonickými osami (standardizované kanonické koeficienty, korelace mezi znaky a diskriminačními funkcemi)
- stačí interpretovat několik prvních os (významnost os: *eigenvalues, % eigenvalues,* kanonické korelační koeficienty, Wilksovo lambda)





Klasifikační diskriminační analýza

- slouží k identifikaci objektů
- cílem je odvodit rovnici, která kombinuje jednotlivé znaky pomocí vah

Marhold & Suda 2002

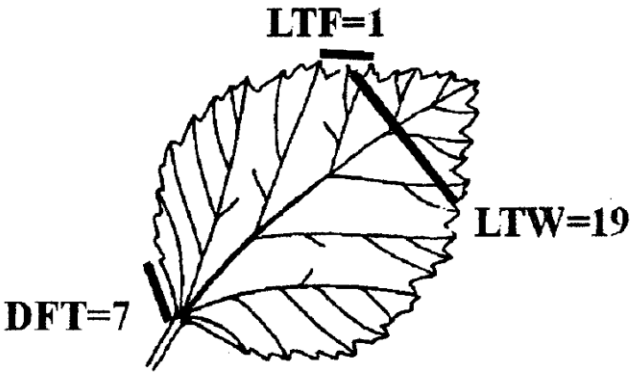
Klasifikační diskriminační analýza

- např. listy břízy
- klasifikační funkce:

  y = 12LTF + 2DFT – 2LTW – 23

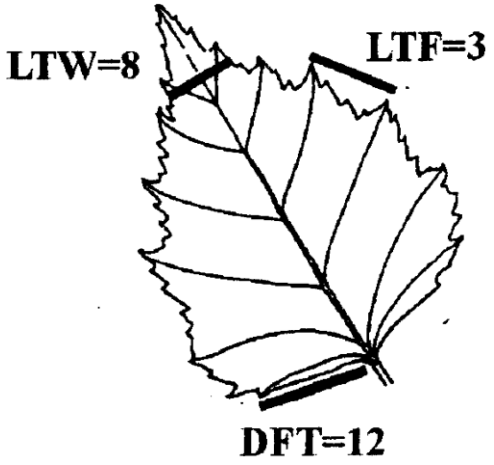
y < 0

***B. pubescens* = -35**



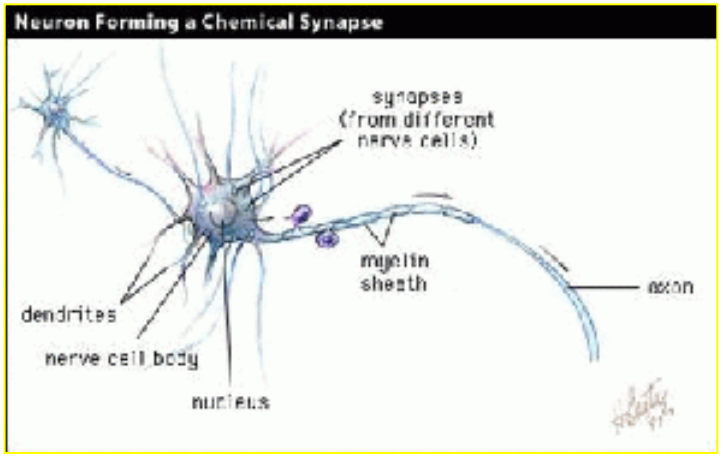
y > 0

***B. pendula* = +21**

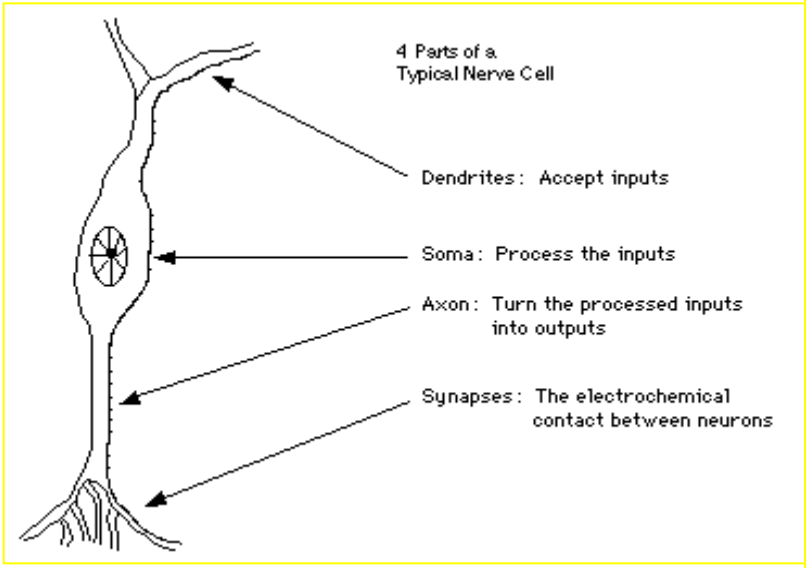


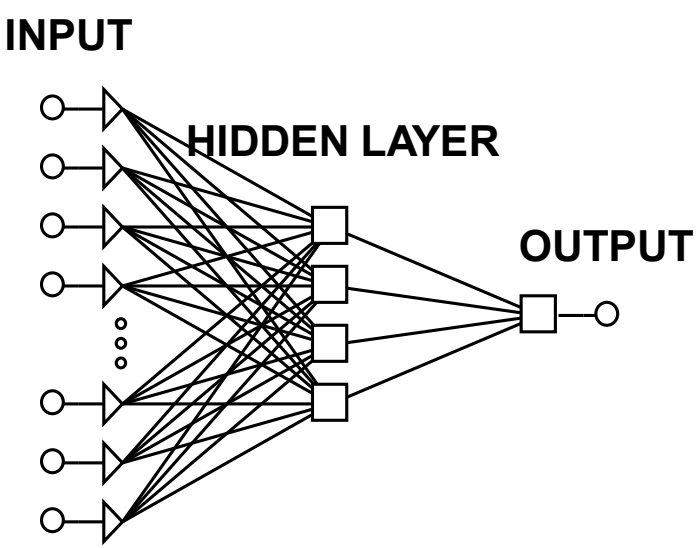
Marhold & Suda 2002

Umělé neuronové sítě (ANN)

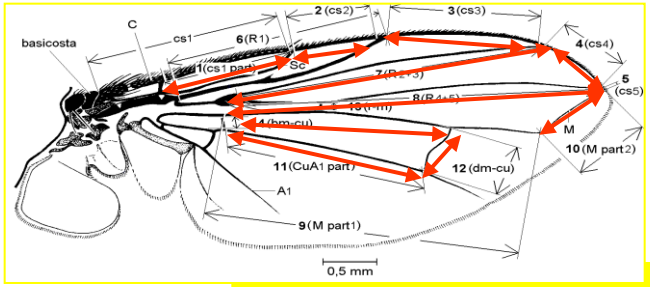


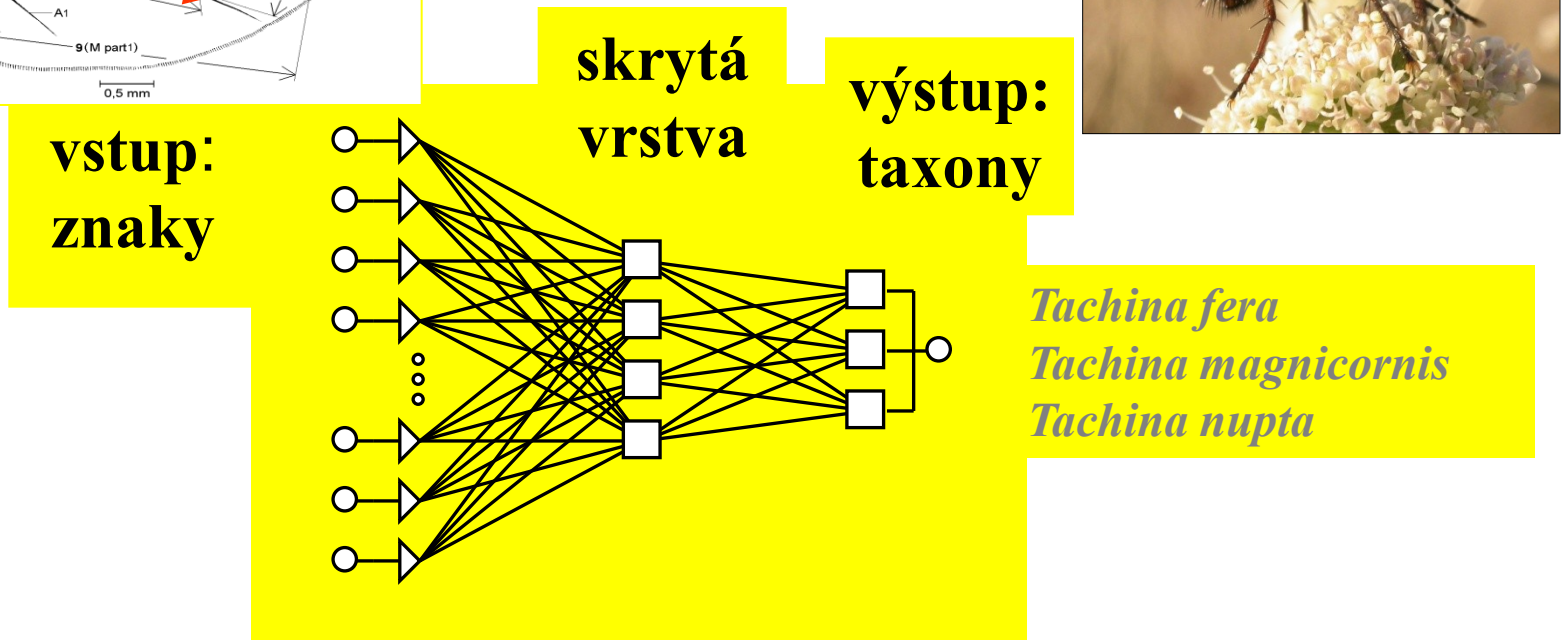
- matematické modely napodobující strukturu a funkci nervové soustavy
- složeny z mnoha dílčích funkčních jednotek - uzlů (umělých neuronů) hierarchicky uspořádaných a vzájemně provázaných ve vrstvách
- architektura sítě závisí na komplexitě problému





Umělé neuronové sítě (ANN) v taxonomii





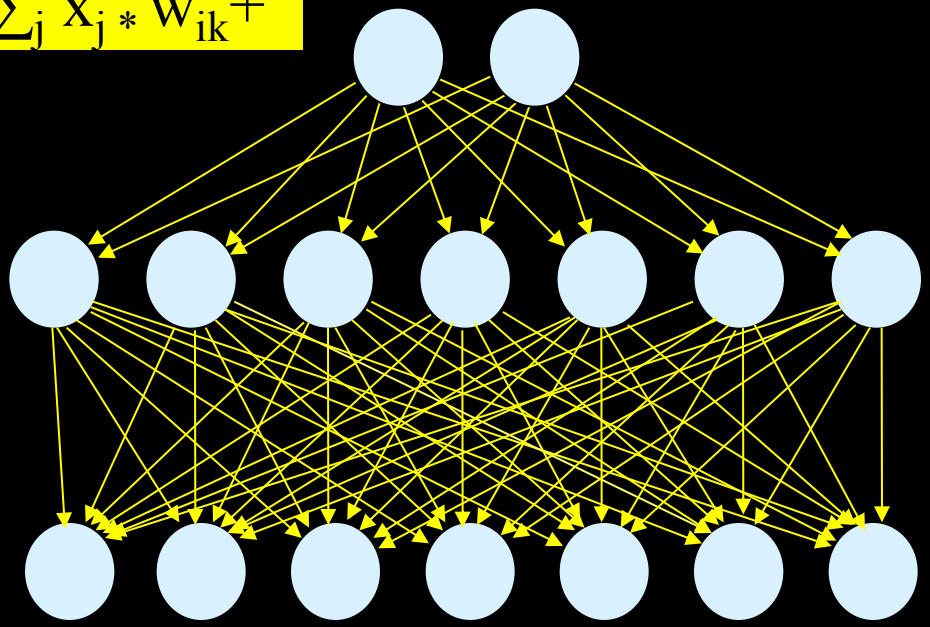
3 fáze: učení (*training*): iterativní tvorba modelu na základě trénovacího souboru námi určených jedinců – nastavení vah mezi jednotlivými neurony s cílem minimalizovat chybu v určování

verifikace (*verification*): ověření správnosti modelu

predikce (*prediction*): určování neznámých jedinců

$$t_{ss} = \sum_i (t_i - o_i)^2 \longrightarrow \text{MINIMUM}$$

$$sum_k = \sum_j x_j * w_{ik} +$$



input

hidden layer

output

$$output_j = f\left(\sum f\left(\sum w x_j + d_1\right) + d_2\right) + e$$

Automatické určování taxonů

- ANN jsou statisticky velmi robustní, nelineární metoda (nezávisí na rozložení a typu dat) se schopností učit se z příkladů
- ideální základ pro automatické systémy určování organizmů
- vstupní data: morfometrie, světelná spektra, bioakustika, koncentrace chemických látek v těle, transformované digitální fotografie,…

např. určování přílipek (*Patella* spp.) na základě koncentrací nasyc. uhlovodíků (Hernández-Borges et al. 2003)

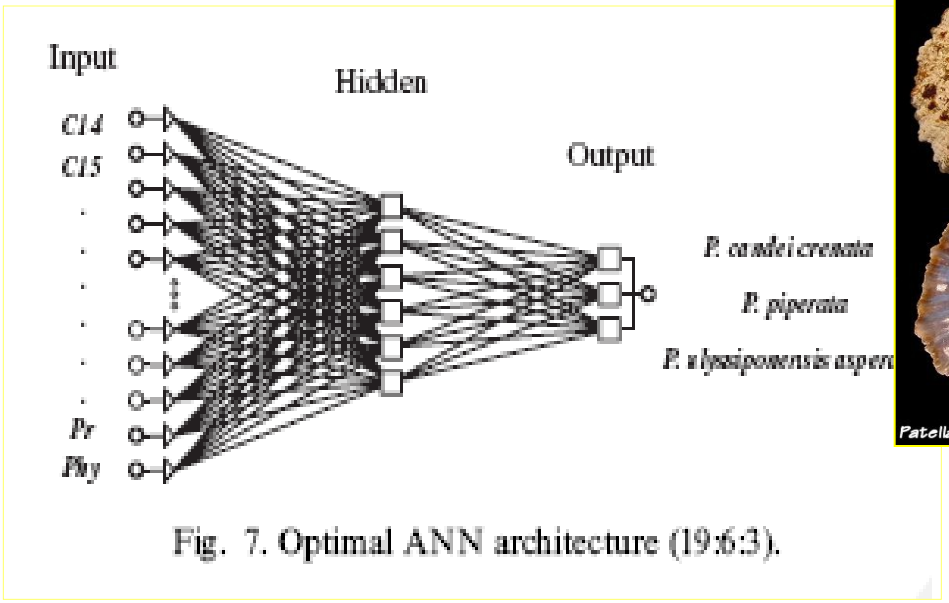
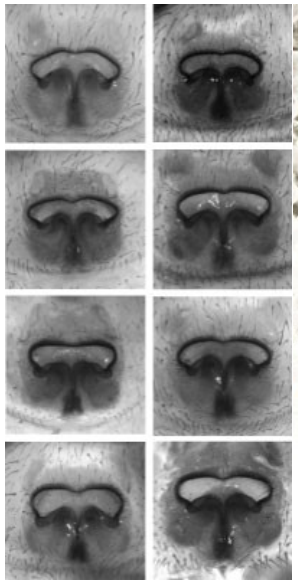
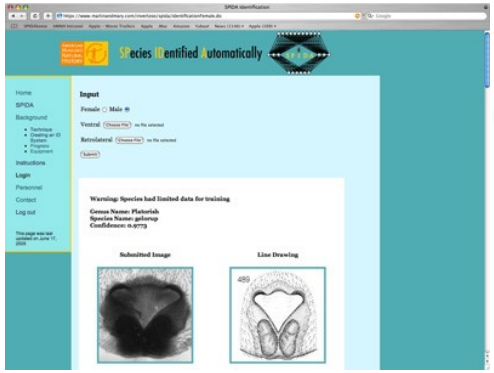


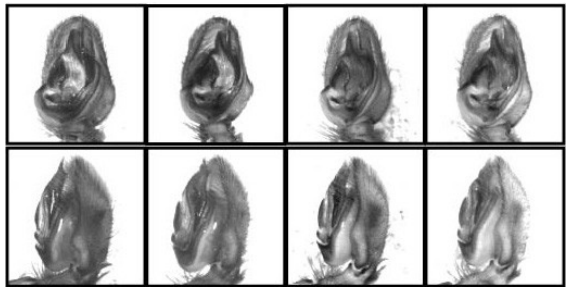
Fig. 7. Optimal ANN architecture (19:6:3).

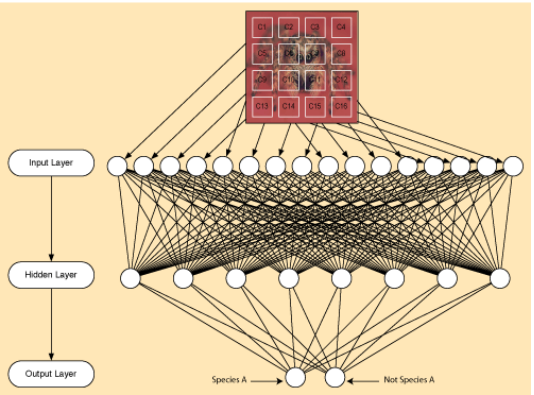
Automatické určování organizmů

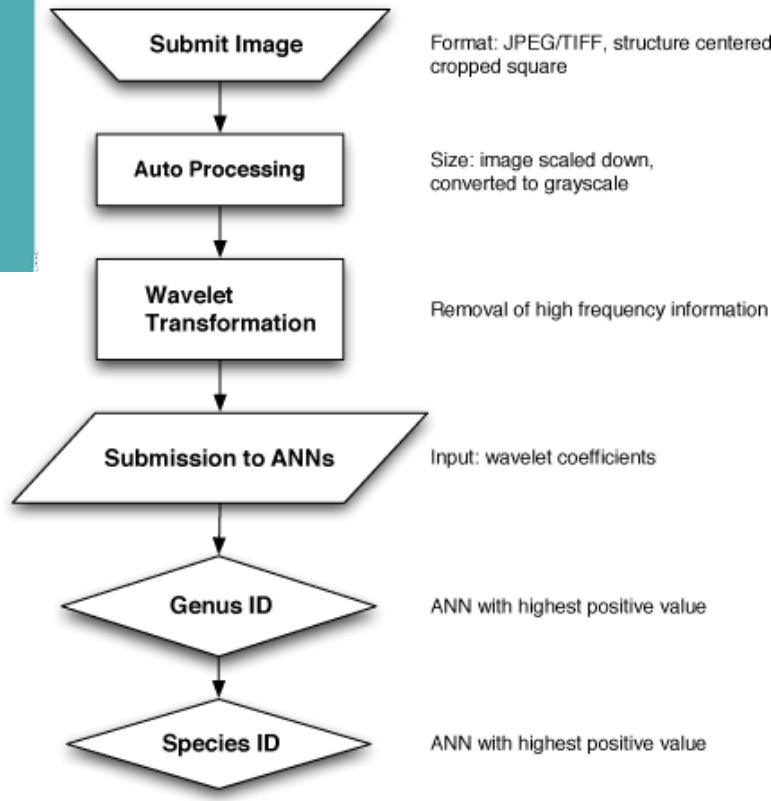
- např. SPIDA – web (Platnick et al. 2005) https://research.amnh.org/invertzoo/spida/common/index.htm
- automatický systém určování australských pavouků čel. Trochanteriidae (15 rodů, 121 druhů) přes internet na základě zaslaných fotografií





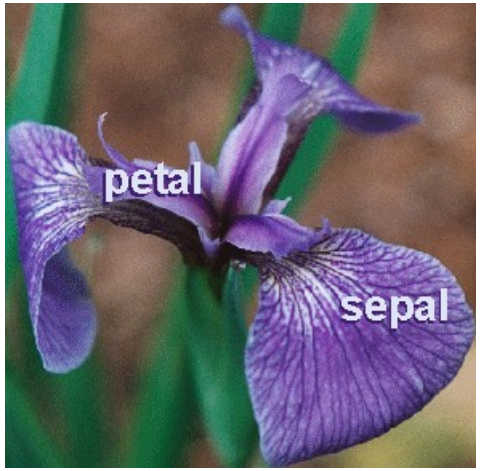






Příklad 1 – „Iris flower dataset“

- R. A. Fisher (1936): 3 druhy blízce příbuzných severoamerických kosatců
- od každého druhu 50 jedinců
- měřeny 4 znaky – délka a šířka okvětních lístků
- liší se jednotlivé druhy od sebe?
- analýza v programu PAST a STATISTICA



Iris virginica

Iris versicolor

Iris setosa

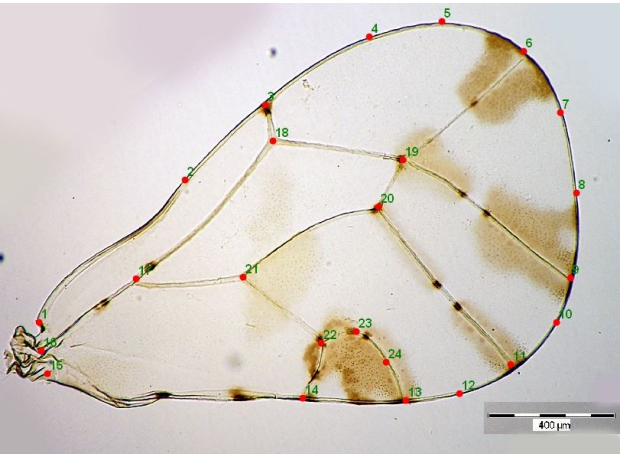
Příklad 2: Mouchy komplexu *Dinera carinifrons*

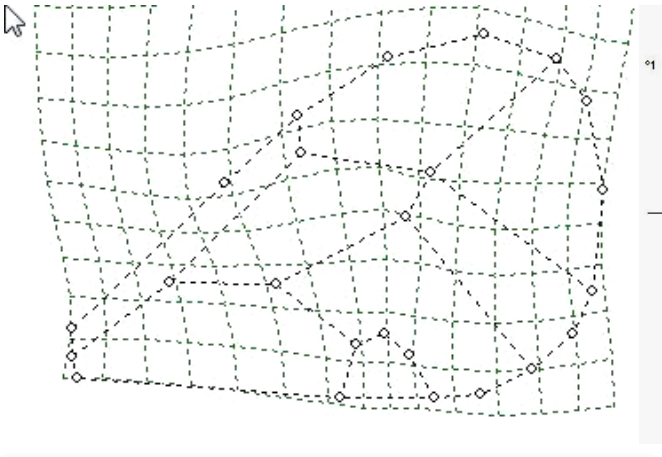
- Lutovinas *et al.* (2013), Diptera: Tachinidae – parazitoidi vrubounovitých brouků
- 2-3 druhy, 55 jedinců
- 19 znaků (délek, 8 na hlavě, 11 na křídlech)
- analýza v programu STATISTICA

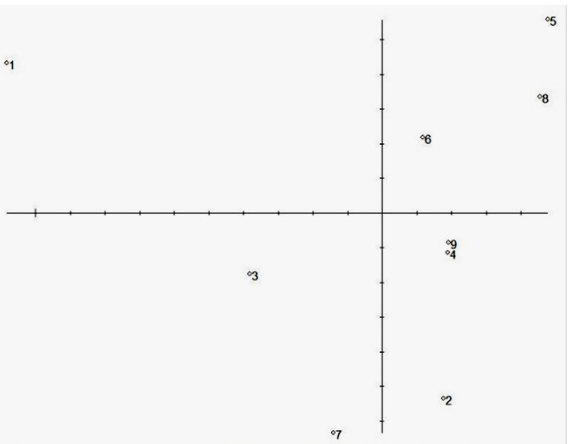
Dinera ferina

Příklad: Mery rodu *Pseudophacopteron*

- geometrická morfometrika předních křídel 9 afrotropických druhů na základě polohy význačných bodů
- příprava datového souboru, výběr landmarků a semilandmarků, metoda ohebných pásků, analýza relativních deformací v souboru programů TPS







Odkazy

- Marhold K. & Suda J. (2002): Statistické zpracování mnohorozměrných dat v taxonomii (Fenetické metody). Univerzita Karlova v Praze, Karolinum, Praha, 159 s.
- Zima J. & Macholán M. (2004) Analýza fenotypu. Pp. 9-49. In. Zima J., Macholán M., Muclinger P., Piálek J. (2004) Genetické metody v zoologii. Univerzita Karlova.
- http://folk.uio.no/ohammer/past/: freewarový statistický balík PAST se širokým použitím v taxonomii a ekologii (O. Hammer)
- Zelditch M.L., Swiderski D.L., Sheets H.D., Fink W.L. (2004): Geometric Morphometrics for Biologists: A Primer. Academic Press, New York, 443 s.
- http://life.bio.sunysb.edu/morph/: různé informace o geometrické morfometrice včetně softwaru (J. Rohlf)
- McLeod (ed.) (2007): Automated Taxon Identification in Systematics. Theory, Approaches and Applications. Systematics Association Special Volumes Series 74. CRC Press, London, 339 s.